CENGAGE Learning™

Introducción a la econometría
Un enfoque moderno

Jeffrey M. Wooldridge

Introducción a la **econometría**
Un enfoque moderno

4a. edición

Jeffrey M. Wooldridge
Michigan State University

Traducción
Ma. del Carmen Enriqueta Hano Roa
Érika M. Jasso Hernan D´Borneville
Traductoras profesionales

Revisión técnica
Roberto Palma Pacheco
Centro de Alta Dirección en Economía y Negocios
Universidad Anáhuac

Domingo Rodríguez Benavides
Facultad de Economía
Universidad Nacional Autónoma de México

CENGAGE Learning™

Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur

CENGAGE Learning™

Introducción a la econometría
Un enfoque moderno
4a. edición
Jeffrey M. Wooldridge

Presidente de Cengage Learning Latinoamérica:
Javier Arellano Gutiérrez

Director general México y Centroamérica:
Pedro Turbay Garrido

Director editorial Latinoamérica:
José Tomás Pérez Bonilla

Director de producción:
Raúl D. Zendejas Espejel

Coordinadora editorial:
María Rosas López

Editor senior:
Javier Reyes Martínez

Editora de producción:
Abril Vega Orozco

Diseño de portada:
DIHO Comunicación gráfica

Imagen de portada:
Stock.xchng

Composición tipográfica:
Heriberto Gachúz Chávez

Traducido del libro *Introductory Econometrics*, Fourth Edition.
Publicado en inglés por South-Western Cengage Learning ©2009
ISBN-13: 978-0-324-66054-8
ISBN-10: 0-324-66054-5

Datos para catalogación bibliográfica:
Wooldridge, Jeffrey M. *Introducción a la econometría. Un enfoque moderno,* 4a. edición.
ISBN-13: 978-607-481-312-8
ISBN-10: 607-481-312-4

Visite nuestro sitio en:
http://latinoamerica.cengage.com

Contenido breve

Contenido detallado

CAPÍTULO 4
Análisis de regresión múltiple: inferencia 117

CAPÍTULO 5
Análisis de regresión múltiple: MCO asintóticos 167

CAPÍTULO 6
Análisis de regresión múltiple: temas adicionales 184

CAPÍTULO 7
Análisis de regresión múltiple con información cualitativa: variables binarias (o *dummy*) 225

CAPÍTULO 8

Heterocedasticidad 264

CAPÍTULO 9

Más sobre especificación y temas de datos 300

PARTE 2

Análisis de regresión con datos de series de tiempo 339

CAPÍTULO 10

Análisis básico de regresión con datos de series de tiempo 340

CAPÍTULO 18

Temas avanzados de series de tiempo 623

CAPÍTULO 19

Realización de un proyecto empírico 668

APÉNDICE A

Herramientas matemáticas básicas 695

APÉNDICE B

Fundamentos de probabilidad 714

APÉNDICE C
Fundamentos de estadística matemática 747

APÉNDICE D
Resumen de álgebra matricial 788

APÉNDICE E

El modelo de regresión lineal en forma matricial 799

APÉNDICE F

Respuestas a las preguntas del capítulo 813

APÉNDICE G

Tablas estadísticas 823

Prefacio

Lo que me motivó a escribir la primera edición de *Introducción a la econometría: un enfoque moderno* fue la brecha tan amplia que existe entre la enseñanza de la materia en los cursos universitarios y la manera en la que los investigadores empíricos entienden y aplican los métodos econométricos. Quedé convencido de que una introducción a la econometría desde la perspectiva de los usuarios profesionales simplificaría la exposición y haría el tema mucho más interesante.

Con base en la reacción positiva a las ediciones anteriores, parece que esta fue una idea acertada. Muchos docentes, con formación e intereses diversos y cuyos estudiantes tienen niveles desiguales de preparación, han adoptado el enfoque moderno de la econometría expuesto en este libro. En esta edición sigo haciendo énfasis en la econometría aplicada a cuestiones reales. Todos los métodos econométricos están motivados por problemas particulares con los que se encuentran los investigadores al analizar datos no experimentales. El punto central en el libro es la comprensión e interpretación de los supuestos a la luz de aplicaciones empíricas reales: las matemáticas requeridas no van más allá del álgebra universitaria y la probabilidad y estadística básicas.

Creado para los profesores de econometría de hoy

En esta cuarta edición se conserva la estructura general de la tercera. La característica más sobresaliente que distingue a este libro de la mayoría es la división de los temas con base en el tipo de datos analizados. Este es un claro distanciamiento de la metodología tradicional, en la que se presenta un modelo lineal, se enumeran todos los supuestos que pueden necesitarse en algún punto posterior del análisis y después se prueban o presentan resultados sin relacionarlos claramente con los supuestos. Mi metodología consiste en tratar primero, en la parte 1, el análisis de regresión múltiple con datos de corte transversal bajo el supuesto de un muestreo aleatorio. Esto resulta conocido para el lector, porque ya está familiarizado con el muestreo aleatorio de poblaciones por los cursos de introducción a la estadística. En gran medida, esto permite distinguir los supuestos acerca del modelo de regresión poblacional subyacente –supuestos a los que se les puede dar un contenido económico o conductual– de los supuestos acerca de cómo se muestrearon los datos. El análisis de las consecuencias de un muestreo no aleatorio puede verse de forma intuitiva una vez que los estudiantes tengan una adecuada comprensión del modelo de regresión múltiple estimado usando muestras aleatorias.

Una característica importante de un enfoque moderno es que las variables explicativas –junto con la variable independiente– son tratadas como resultados (valores) de las variables aleatorias. En las ciencias sociales, considerar variables explicativas aleatorias es mucho más razonable que el supuesto tradicional de variables explicativas no aleatorias. Una ventaja no trivial, es que el modelo poblacional/enfoque del muestreo aleatorio reduce la cantidad de supuestos que el lector tiene que absorber y entender. Curiosamente, el enfoque clásico al análisis de regresión, que trata a las variables explicativas como fijas en muestras repetidas, y que sigue siendo dominante en los libros introductorios, se usa de manera literal con datos recolectados experimentalmente.

Además, los argumentos que se requieren para establecer y explicar los supuestos pueden ser poco claros para el lector.

Hago hincapié que en el modelo poblacional los principales supuestos que subyacen al modelo del análisis de regresión, como el supuesto de media cero para los factores no observables, son condicionales adecuadamente impuestos sobre las variables explicativas. Esto lleva a una clara comprensión de los tipos de problemas, como la heterocedasticidad (varianza no constante), que pueden invalidar los procedimientos estándar de inferencia. Además, logro disipar diversas concepciones erróneas que surgen en los libros de econometría de todos los niveles. Por ejemplo, explico por qué la R-cuadrada común es también válida como medida de bondad de ajuste en presencia de heterocedasticidad (capítulo 8) o de errores serialmente correlacionados (capítulo 12); demuestro que las pruebas para la forma funcional no deben ser vistas como pruebas generales de variables omitidas (capítulo 9), y explico por qué en un modelo de regresión siempre deben incluirse variables adicionales de control que no estén correlacionadas con la variable explicativa de interés, con frecuencia la variable clave de política (capítulo 6).

Como los supuestos para el análisis de corte transversal son relativamente sencillos y razonables, los estudiantes pronto pueden abordar aplicaciones serias de corte transversal sin tener que preocuparse por los complicados temas de tendencias, estacionalidad, correlación serial, alta persistencia y regresión espuria, omnipresentes en los modelos de regresión para series de tiempo. En un principio pensé que mi estudio de la regresión con datos de corte transversal seguido de regresión con datos de series de tiempo encontraría la aceptación de los investigadores cuyo interés fuera la microeconomía aplicada, y este parece haber sido el caso. Ha sido grato que a quienes han adoptado este libro y tienen inclinación al uso de series de tiempo también les haya entusiasmado la estructura del mismo. Posponer el análisis econométrico de las series de tiempo permite dar la atención adecuada a los potenciales escollos encontrados en el análisis de datos de series de tiempo, que no surgen con datos de corte transversal. En efecto, la econometría de las series de tiempo por fin recibe el tratamiento serio que se merece en un libro de introducción a la econometría.

Como en ediciones anteriores, he elegido intencionalmente temas que permitan la lectura de artículos de revistas y la realización de investigación empírica básica. En cada tema he omitido deliberadamente muchas pruebas y procedimientos de estimación que, aunque por tradición incluidos en los libros, no han resistido la prueba empírica del tiempo. De la misma manera, hago hincapié en temas más recientes que han demostrado con claridad su utilidad como, por ejemplo, la obtención de estadísticos de prueba que sean robustos a la heterocedasticidad (o a la correlación serial) de formas desconocidas, el uso de años múltiples de datos para el análisis de políticas o la solución del problema de la variable omitida mediante métodos de variables instrumentales. Me parece que ésta fue una buena elección, dado que lo único que he recibido son unas cuantas sugerencias de agregar o eliminar material.

A lo largo del libro sigo una metodología sistemática, lo que significa que los temas se presentan construyendo, de manera lógica, sobre el material previo, y los supuestos se introducen a medida que son necesarios para la obtención de alguna conclusión. Por ejemplo, los profesionales de la econometría comprenden que no se necesitan todos los supuestos de Gauss-Markov para mostrar que los estimadores por mínimos cuadrados ordinarios (MCO) son insesgados. Sin embargo, la gran mayoría de los libros de econometría introduce el conjunto completo de supuestos (muchos de los cuales son redundantes o incluso lógicamente contradictorios).

Mi enfoque sistemático se ilustra por el orden de los supuestos que utilizo para la regresión múltiple en la Parte 1. Este orden resulta en una progresión natural para resumir brevemente el rol de cada supuesto:

RLM.1: Presenta el modelo poblacional e interpreta los parámetros poblacionales (que se espera estimar).

RLM.2: Presenta el muestreo aleatorio de una población y describe los datos que se emplean para estimar los parámetros poblacionales.
RLM.3: Agrega el supuesto, sobre las variables explicativas, que permite calcular los estimados a partir de la muestra; este es el llamado supuesto de ausencia de colinealidad perfecta.
RLM.4: Supone que, en la población, la media de los errores no observables no depende de los valores de las variables explicativas; este es el supuesto de "independencia de la media" combinado con una media poblacional cero del error, y es el supuesto clave que produce el insesgamiento de los MCO.

Una vez introducidos los supuestos RLM.1 a RLM.3 se pueden analizar las propiedades algebraicas de los MCO, es decir, las propiedades algebraicas de los MCO en un determinado conjunto de datos. Agregando el supuesto RLM.4, se puede demostrar que los MCO son insesgados (y consistentes). El supuesto RLM.5 (homocedasticidad) se agrega para el teorema de Gauss-Markov (y para que sean válidas las fórmulas usuales de la varianza de los MCO), y RLM.6 (normalidad) se agrega para completar los supuestos clásicos del modelo lineal (para la inferencia estadística exacta).

Al pasar al estudio de las propiedades de las muestras grandes y al tratar la regresión de series de tiempo en la parte 2, empleo métodos paralelos. La cuidadosa presentación y el análisis de los supuestos hace relativamente fácil tratar temas más avanzados, como el uso de combinaciones de corte transversal, el aprovechamiento de la estructura de los datos de panel y la aplicación de los métodos de variables instrumentales. En general, me he esforzado por dar una visión unificada de la econometría en la que todos los estimadores y estadísticos se obtengan empleando sólo algunos principios intuitivamente razonables de estimación y prueba (los que, por supuesto, también tienen una rigurosa justificación). Por ejemplo, las pruebas de correlación lineal y heterocedasticidad basadas en la regresión resultan fáciles de entender para el lector, debido a que ya tienen una sólida comprensión de la regresión. Esto contrasta con tratamientos que dan un conjunto de recetas desarticuladas para obsoletos procedimientos econométricos de prueba.

En todo el libro hago énfasis en las relaciones *ceteris paribus*, a lo que se debe que después de un capítulo sobre el modelo de regresión simple, pase al análisis de regresión múltiple. La regresión múltiple motiva al lector a pensar tempranamente acerca de aplicaciones serias. También doy gran importancia al análisis de políticas con todo tipo de estructuras de datos. Los temas prácticos, como el uso de variables proxy para obtener efectos *ceteris paribus* y la interpretación de efectos parciales en modelos con términos de interacción, son vistos de manera sencilla.

Lo nuevo en esta edición

Entre los cambios en esta edición, en el capítulo 3 se encuentra un análisis sobre los factores de inflación de la varianza. Hasta ahora me he resistido a incluir un análisis formal de los diagnósticos existentes para detectar multicolinealidad. En esta edición proporciono, con algunas reservas, un breve análisis. Como en ediciones anteriores, mi opinión –de que la multicolinealidad es aún un tema poco comprendido y de que las afirmaciones que dicen que uno puede detectar y corregir multicolinealidad están equivocadas– no ha cambiado. Pero como me he encontrado teniendo que explicar repetidamente el uso y límites de estadísticos como los factores de inflación de la varianza, he decidido enfrentar el tema directamente.

En el capítulo 6 agrego un análisis del llamado estimado smearing para la transformación después de estimar un modelo lineal en el que la variable dependiente está en forma logarítmica. El método del smearing es muy útil y fácil de implementar; fue mi descuido no incluirlo en ediciones anteriores. También he añadido material sobre la obtención de un intervalo de predicción de 95% después de transformar un modelo que satisface los supuestos clásicos del modelo lineal.

En el capítulo 8 cambié el ejemplo 8.6 por uno en el que se emplea una base de datos, más moderna y mucho más grande, sobre riqueza financiera, ingreso y participación en los planes de pensión 401(k). Este ejemplo, junto con una nueva subsección sobre mínimos cuadrados ponderados con una función de varianza mal especificada, proporciona una bella ilustración de cómo los mínimos cuadrados ponderados pueden ser significativamente más eficientes que los mínimos cuadrados ordinarios, aun cuando se permita que la función de varianza esté mal especificada.

En otra nueva subsección del capítulo 8 se analiza el problema de la predicción después de la retransformación en un modelo con una variable dependiente logarítmica y heterocedasticidad en el modelo original.

En el capítulo 9 hay varios puntos nuevos. Primero proporciono un breve análisis de los modelos con pendientes aleatorias. Suministro este material como introducción a la idea de que los efectos marginales pueden depender de la heterogeneidad individual no observada. En el análisis de puntos atípicos y observaciones influyentes incluyo una descripción de los "residuales estudentizados" como medio para determinar observaciones influyentes. También hago notar cómo éstos pueden ser obtenidos fácilmente convirtiendo una observación en variable ficticia. Por último el cada vez más importante modelo de las desviaciones mínimas absolutas (DMA), ahora se describe de forma más completa en una nueva subsección. En los ejercicios para computadora se emplea una nueva base de datos sobre la compensación de los profesores de las escuelas elementales de Michigan para ilustrar la resistencia de las DMA a la inclusión de datos sospechosos.

En los capítulos sobre series de tiempo, capítulos 10, 11 y 12, se incluyen dos ejemplos nuevos (y bases de datos sobre la economía de Estados Unidos). El primer ejemplo es una ecuación sencilla conocida en la economía como Ley de Okun; el segundo es un análisis sectorial-específico de los efectos del salario mínimo. Estos ejemplos ilustran bellamente aplicaciones de la regresión a la economía empleando datos de series de tiempo.

En los capítulos avanzados ahora se incluye un análisis de la prueba de Chow para datos de panel (capítulo 13), un análisis más detallado de los métodos de MCO combinados y datos de panel para muestras de aglomerados (capítulo 14) y mejores análisis de los problemas de un instrumento débil y del carácter de las pruebas de sobreidentificación con variables instrumentales).

En el capítulo 17 amplié el análisis de la estimación de efectos parciales en modelos no lineales, haciendo hincapié en la diferencia entre efectos parciales evaluados en los promedios de los regresores frente a promediar los efectos parciales sobre todas las unidades.

En esta edición se han agregado más bases de datos. Ya antes mencioné la base de datos sobre la compensación de los profesores (ELEM94_95.RAW). Además, en algunos problemas nuevos se usa una base de datos sobre donaciones de caridad en Holanda (CHARITY.RAW). Hay dos nuevas bases de datos de series de tiempo OKUN.RAW y MINWAGE.RAW.

Hay también algunas otras bases de datos que no se usan en el libro, pero que pueden encontrarse en el sitio web de la obra: sobre los salarios y las publicaciones de los profesores de economía de las 10 principales universidades de Estados Unidos.

Dirigido a estudiantes de licenciatura, adaptable para estudiantes de posgrado

Este libro está pensado para estudiantes de licenciatura que hayan tomado los cursos de álgebra universitaria y uno de introducción a la probabilidad y estadística. (Los apéndices A, B y C contienen el material requerido.) En los cursos de econometría de un semestre o de un trimestre no puede esperarse que se vean todos, o incluso ninguno, de los temas más avanzados de la parte 3. Un curso introductorio típico comprende los capítulos 1 a 8 en los que se ven las bases de la regresión simple y de la regresión múltiple para datos de corte transversal. En tanto en este se conserve el acento en la intuición y la interpretación de ejemplos empíricos, el material de los primeros ocho capítulos deberá ser accesible para los estudiantes de licenciatura de casi todos los departamentos de economía. La mayoría de los docentes deseará ver también, con diverso grado de profundidad, al menos parte de los capítulos sobre análisis de regresión con datos de series de tiempo, capítulos 10, 11 y 12. En el curso de un semestre en la Universidad Estatal de Michigan donde imparto clases, veo el capítulo 10 con bastante detalle, proporciono una visión general de los temas del capítulo 11 y analizo el material sobre correlación serial del capítulo 12. Considero que este curso básico proporciona a los estudiantes bases sólidas para redactar trabajos empíricos, como trabajos semestrales, trabajos de seminarios de los estudiantes de último semestre o tesis. El capítulo 9 contiene temas más especializados que surgen en el análisis de datos de corte transversal y problemas como, por ejemplo, observaciones atípicas y muestreo no aleatorio; en un curso de un semestre, puede omitirse este capítulo sin pérdida de la continuidad.

La estructura del libro lo hace ideal para un curso centrado en datos de corte transversal y análisis de políticas: en lugar de los temas de los capítulos 9, 13, 14 y 15, pueden saltarse los capítulos sobre series de tiempo. El capítulo 13 es avanzado sólo en el sentido de que se tratan dos nuevas estructuras de datos: cortes transversales independientes combinados y análisis de datos de panel de dos periodos. Estas estructuras de datos son especialmente útiles en el análisis de políticas, y en este capítulo se proporcionan varios ejemplos. Quien haya comprendido bien los capítulos 1 a 8 tendrá pocas dificultades con el capítulo 13. En el capítulo 14 se analizan métodos más avanzados para datos de panel y probablemente sólo podrán ser vistos en un segundo curso. Una buena manera de concluir un curso sobre métodos para corte transversal es estudiando las nociones elementales de la estimación de variables instrumentales del capítulo 15.

El material selecto de la parte 3, que comprende los capítulos 13, 14, 15 y 17, lo he usado en un seminario para estudiantes de último año, orientado a la elaboración de un artículo de investigación. Los estudiantes que además de cursos básicos de un semestre hayan tenido contacto con el análisis básico de datos de panel, con la estimación de variables instrumentales y con los modelos de variables dependientes limitadas estarán en condiciones de leer amplios segmentos de la literatura aplicada a las ciencias sociales. En el capítulo 17 se proporciona una introducción a los modelos más comunes de variables dependientes limitadas.

Este libro es, también, adecuado para un curso introductorio a nivel maestría, en el que se haga hincapié en las aplicaciones y no en las deducciones empleando álgebra de matrices. De todos modos, para los docentes que deseen presentar el material de forma matricial, los apéndices D y E son exposiciones del álgebra de matrices y del modelo de regresión múltiple de forma matricial.

En la Universidad Estatal de Michigan, los estudiantes de doctorado de diversas áreas que requieren el análisis de datos –contaduría, economía agrícola, economía del desarrollo, finanzas, economía internacional, economía laboral, macroeconomía, ciencias políticas y finanzas públicas– han encontrado en este libro un puente entre el trabajo empírico que leen y la econometría más teórica que estudian a nivel doctorado.

Características didácticas

A lo largo de todo el libro se encuentran diseminadas preguntas, cuyas respuestas se dan en el apéndice F. Estas preguntas tienen como finalidad proporcionar al lector una retroalimentación inmediata. Todos los capítulos contienen numerosos ejemplos, muchos de los cuales son estudios de caso tomados de publicaciones recientes, ligeramente modificados para simplificar el análisis y en los que espero no haber sacrificado el punto principal.

Los problemas al final de los capítulos y los ejercicios para computadora están fuertemente orientados hacia el trabajo empírico y no a deducciones complicadas. Al estudiante se le pide que piense cuidadosamente con base en lo que ha aprendido. Con frecuencia, los ejercicios para computadora son una ampliación de los ejemplos presentados dentro del capítulo. En varios ejercicios se emplean bases de datos de trabajos publicados o bases de datos motivadas por publicaciones de investigaciones sobre economía u otros campos.

Una innovadora característica de este libro es el extenso glosario. Las breves descripciones y definiciones son un útil repaso para los estudiantes al preparar exámenes o leer investigaciones empíricas que emplean métodos econométricos. En esta cuarta edición se agregan varios términos y otros han sido actualizados.

Complementos para el estudiante

El *Student Solutions Manual* (ISBN 0-324-58658-2) contiene sugerencias sobre cómo leer cada capítulo, así como las respuestas a algunos problemas y ejercicios para computadora.

Complementos para el profesor

El *Instructor's Manual with Solutions* (ISBN 0-324-58657-4) contiene las respuestas a todos los ejercicios, así como sugerencias sobre cómo presentar el material de cada capítulo. El manual del instructor contiene también fuentes para cada uno de los archivos de datos, y muchas sugerencias de cómo usarlas en bases de datos, exámenes y trabajos del trimestre.

Bases de datos – disponibles en cuatro formatos

Hay aproximadamente 100 bases de datos disponibles en ASCII, EViews, Excel y Stata. Debido a que la mayoría de las bases de datos proviene de investigaciones reales, algunas son muy extensas. Salvo algunas listas incompletas de bases de datos que se emplean como ilustración de las diversas estructuras de datos, estas bases de datos no aparecen en la obra. El libro está pensado para un curso en que el uso de computadoras forme parte integral del mismo. En línea puede obtenerse un amplio manual con la descripción de los datos. Éste contiene una lista de fuentes de datos, así como sugerencias de cómo usar las bases de datos que no están descritas en el libro.

Recursos para el profesor

Este libro cuenta con una serie de recursos para el profesor, los cuales están disponibles en inglés y sólo se proporcionan a los docentes que lo adopten como libro de texto en sus cursos. Para mayor información, comuníquese a las oficinas de nuestros representantes o a las siguientes direcciones de correo electrónico:

Cengage Learning México y Centroamérica	clientes.mexicoca@cengage.com
Cengage Learning Caribe	clientes.caribe@cengage.com
Cengage Learning Cono Sur	clientes.conosur@cengage.com
Cengage Learning Pacto Andino	clientes.pactoandino@cengage.com

Las direcciones de los sitios web referidas en el libro no son administradas por Cengage Learning Latinoamérica, por lo que ésta no es responsable de los cambios o actualizaciones de las mismas.

Sugerencias para organizar un curso

Ya se ha hablado sobre el contenido de la mayoría de los capítulos, así como sobre las posibles formas de organizar los cursos. Aquí se proporcionan comentarios más precisos acerca de qué material puede verse u omitirse en algunos de los capítulos.

El capítulo 9 tiene algunos ejemplos interesantes (como una regresión del salario en la que intervienen el CI y una variable explicativa). Las variables proxy no es necesario que se presenten formalmente para tratar este tipo de ejemplos y así suelo hacerlo al terminar el análisis de corte transversal. En el capítulo 12, tratándose de un curso de un semestre, omito el material sobre inferencia robusta a la correlación serial para mínimos cuadrados ordinarios, así como los modelos dinámicos de heterocedasticidad.

Incluso en un segundo curso, acostumbro dedicar poco tiempo al capítulo 16 sobre análisis de ecuaciones simultáneas. Si hay algún tema acerca del cual hay diversas opiniones, es la importancia de las ecuaciones simultáneas. Algunos piensan que este tema es fundamental; otros piensan que es difícilmente aplicable. Mi opinión es que los modelos de ecuaciones simultáneas se emplean de forma exagerada (vea el capítulo 16). Si uno lee las aplicaciones cuidadosamente, las variables omitidas y los errores de medición parecen ser las razones más probables por las que se adopte la estimación de variables instrumentales y a esto se debe que, en el capítulo 15, usé las variables omitidas para motivar la estimación de variables instrumentales. Sin embargo, los modelos de ecuaciones simultáneas son indispensables para estimar funciones de oferta y demanda y son también usados en otros casos importantes.

El capítulo 17 es el único en el que se consideran modelos en los que los parámetros son inherentemente no lineales, lo que es una dificultad más para los estudiantes. Lo primero que se debe ver en este capítulo son los modelos probit y logit para respuesta binaria. Mi presentación de los modelos Tobit y de la regresión censurada parece seguir siendo novedosa: reconozco explícitamente que el modelo Tobit se usa para soluciones de esquina en muestras aleatorias, mientras que la regresión censurada se usa cuando el proceso de recolección de los datos censura la variable dependiente.

En el capítulo 18 se ven importantes temas recientes de la econometría de las series de tiempo, que comprenden detección de raíces unitarias y cointegración. Este material únicamente lo veo en un segundo curso, ya sea a nivel licenciatura o posgrado. En este capítulo se presenta también una introducción bastante detallada al pronóstico.

El capítulo 19 es mucho más amplio que capítulos similares en otros libros. En él se resumen algunos de los métodos adecuados para distintos tipos de problemas y estructuras de datos, se señalan escollos potenciales, se explica con cierto detalle cómo escribir un trabajo semestral sobre economía empírica y se dan sugerencias para proyectos.

Agradecimientos

Quisiera agradecer a aquellos que revisaron el proyecto para la cuarta edición o que aportaron comentarios útiles sobre la tercera:

Swarnjit S. Arora
University of Wisconsin—Milwaukee

Jushan Bai
New York University

Edward Coulson
Penn State University

Lisa M. Dickson
University of Maryland—Baltimore County

Angela K. Dills
Clemson University

Michael Jansson
University of California—Berkeley

Subal C. Kumbhakar
State University of New York—Binghamton

Angelo Melino
University of Toronto

Daniel Monchuk
University of Southern Mississippi

Kevin J. Murphy
Oakland University

Leslie Papke
Michigan State University

Subhash Ray
University of Connecticut

Edwin A. Sexton
Brigham Young University—Idaho

Lara Shore-Sheppard
Williams College

Jeffrey Smith
University of Michigan

Stephen Stageberg
University of Mary Washington

Timothy Vogelsang
Michigan State University

Anne E. Winkler
University of Missouri—St. Louis

Dec Mullarkey
Boston College

Varios de los cambios que señalé antes se deben a los comentarios recibidos de personas que se encuentran en esta lista y aún sigo considerando algunas de las sugerencias hechas por uno o más de los revisores.

Muchos estudiantes y profesores asistentes, demasiados para ser mencionados, han encontrado algún error en las ediciones pasadas o han sugerido una redacción diferente para algunos párrafos. Agradezco sus sugerencias.

Gracias al personal de Cengage Learning, el proceso de revisión se realizó sin tropiezos. Mike Worls, mi editor de adquisiciones, desde hace tiempo, me dio, como siempre, todo su apoyo y Laura Bofinger llegó en plena actividad como mi nueva editora de desarrollo. Me sirvió mucho su entusiasmo en este proyecto.

Martha Conway fue muy eficiente como gerente de proyecto y Charu Khanna de Macmillan Publishing Solutions supervisó eficiente y profesionalmente la composición tipográfica del manuscrito.

Dedico este libro a mi esposa, Leslie –quien sometió a sus estudiantes de seminario a la tercera edición– y a nuestros hijos, Edmund y Gwenyth –quienes ya entienden lo suficiente de economía como para saber que más les vale ser científicos “reales.”

Jeffrey M. Wooldridge

Acerca del autor

Jeffrey M. Wooldridge es Profesor Universitario Distinguido de Economía en la State University en donde ha impartido clases desde 1991. A partir de 1986, y hasta 1991, fue profesor asistente de economía en el Instituto Tecnológico de Massachussets (MIT). Obtuvo su título de Licenciado en letras, con especializaciones en ciencias computacionales y economía, en la Universidad de California, Berkeley, en 1982 y su doctorado en economía en la Universidad de California, San Diego en 1986. Ha publicado cerca de 40 artículos en periódicos reconocidos a nivel internacional, así como varios capítulos de libros. Es también autor de *Econometric Analysis of Cross Section and Panel Data*. Sus premios incluyen la Beca Alfred P. Sloan Research, el premio Plura Script de *Econometric Theory*, el premio Sir Richard Stone del *Journal of Applied Econometrics* y tres premios de profesor del año de graduados del MIT. Es miembro de *Econometric Society* y del *Journal of Econometrics*. Ha sido editor del *Journal of Business and Economic Statistics*, coeditor de econometría de *Economic Letters* y colaborador de los consejos editoriales de *Econometric Theory*, el *Journal of Economic Literature*, el *Journal of Econometrics*, de *Review of Economics and Statistics* y del *Stata Journal*. También ha fungido como consultor en econometría para Arthur Andersen, Charles River Associates y el Washington State Institute for Public Policy.

CAPÍTULO 1

La naturaleza de la econometría y los datos económicos

En el capítulo 1 se presenta el campo de acción de la econometría y los problemas generales que surgen al aplicar los métodos econométricos. En la sección 1.3 se examinan los tipos de bases de datos que se emplean en los negocios, la economía y en otras ciencias sociales. En la sección 1.4 se presenta un análisis intuitivo sobre las dificultades relacionadas con la inferencia de la causalidad en las ciencias sociales.

1.1 ¿Qué es la econometría?

Imagine que el gobierno lo contrata para evaluar la efectividad de un programa de capacitación para el trabajo financiado con fondos públicos. Suponga que se trata de un programa para instruir a los trabajadores sobre diversas maneras de utilizar las computadoras en los procesos de fabricación. Este programa, de veinte semanas, ofrece cursos en horarios fuera de la jornada laboral. Cualquier trabajador de la industria puede participar e inscribirse de manera voluntaria a todo el programa o a una parte de él. Usted tiene que determinar si este programa de capacitación laboral tiene algún efecto sobre los posteriores salarios por hora de los trabajadores.

Ahora, suponga que trabaja para la banca de inversión. Tiene que estudiar el rendimiento de varias estrategias de inversión a corto plazo en certificados o letras del tesoro de Estados Unidos para probar si se cumplen las teorías económicas implicadas.

A primera vista, la tarea de responder a estas preguntas puede parecer desalentadora. Por ahora, puede que de lo único que tenga una vaga idea sea del tipo de datos que debe recolectar. Al finalizar el curso de econometría usted sabrá emplear los métodos econométricos para evaluar, de manera formal, un programa de capacitación laboral o para probar una teoría económica sencilla.

La econometría se basa en el desarrollo de métodos estadísticos que se utilizan para estimar relaciones económicas, probar teorías económicas y evaluar e implementar políticas públicas y de negocios. La aplicación más común de la econometría es en el pronóstico de variables macroeconómicas tan importantes como las tasas de interés, de inflación y el producto interno bruto. Si bien el pronóstico de indicadores económicos es un tema muy conocido y al que se le suele dar mucha publicidad, los métodos econométricos también se emplean en áreas de la economía que no están relacionadas con la elaboración de pronósticos macroeconómicos. Por ejemplo, se estudiarán los efectos de los gastos de campaña política sobre los resultados de las votaciones. En el campo de la educación, se considerará el efecto que tiene el gasto público en escuelas sobre el desempeño de los estudiantes. Además, se aprenderá a emplear los métodos econométricos para pronosticar series de tiempo económicas.

La econometría se ha convertido en una disciplina independiente de la estadística matemática por ocuparse de la recolección y análisis de datos económicos no experimentales. **Datos no experimentales** son datos sobre individuos, empresas o segmentos de la economía que no son obtenidos por medio de experimentos controlados. (A los datos no experimentales en ocasiones también se les llama **datos retrospectivos** o **datos observacionales**, para subrayar el hecho de que el investigador es recolector pasivo de los datos.) En las ciencias naturales los **datos experimentales** suelen ser obtenidos en el laboratorio, pero en las ciencias sociales son mucho más difíciles de obtener. Aunque es posible idear experimentos sociales, suele ser imposible, prohibitivamente caro o moralmente indeseable realizar la clase de experimentos controlados que serían necesarios para abordar problemas económicos. En la sección 1.4 se dan ejemplos concretos de la diferencia entre datos experimentales y datos no experimentales.

Como es natural, los econometristas han tomado prestado de la estadística matemática todo lo que les ha sido posible. El método del análisis de regresión múltiple es un pilar fundamental en ambos campos, pero su enfoque e interpretación pueden ser notablemente diferentes. Además, los economistas han ideado nuevas técnicas para lidiar con la complejidad de los datos económicos y para probar las predicciones de las teorías económicas.

1.2 Pasos en un análisis económico empírico

Los métodos econométricos tienen importancia en casi todas las ramas de la economía aplicada. Se emplean cuando se desea probar una teoría económica o cuando se piensa en una relación que tiene alguna importancia para decisiones de negocios o para el análisis de políticas. En un **análisis empírico** se utilizan datos para probar teorías o estimar relaciones.

¿Cómo se procede para estructurar un análisis económico empírico? Aunque parezca obvio, vale la pena subrayar que el primer paso en cualquier análisis empírico es la cuidadosa formulación de la pregunta de interés, la cual puede estar relacionada con la prueba de un aspecto determinado de una teoría económica o puede ser adecuada para probar los efectos de una política pública. En principio, los métodos econométricos se pueden emplear para responder a una gama muy amplia de interrogantes.

En algunos casos, en especial en aquellos relacionados con la prueba de teorías económicas, se construye un **modelo económico** formal, el cual consiste en ecuaciones matemáticas que describen diversas relaciones. Los economistas son conocidos por construir modelos para la descripción de una gran variedad de comportamientos. Por ejemplo, en microeconomía intermedia, las decisiones de consumo de un individuo, sujetas a una restricción de presupuesto, se describen mediante modelos matemáticos. La premisa básica que subyace a estos modelos es la *maximización de la utilidad*. El supuesto de que al hacer una elección los individuos, sujetos a restricciones de recursos, eligen aquello que maximice su bienestar, proporciona un sólido marco para la elaboración de modelos económicos manejables y de predicciones claras. En el contexto de las decisiones de consumo, la maximización de la utilidad conduce a un conjunto de *ecuaciones de demanda*. En una ecuación de demanda, la cantidad demandada de cada artículo depende de su precio, del precio de los bienes sustitutos y complementarios, del ingreso del consumidor y de características individuales que influyen en las preferencias.

Los economistas también han empleado herramientas económicas básicas, como la maximización de la utilidad, para explicar comportamientos que a primera vista pueden no parecer de carácter económico. Un ejemplo clásico es el modelo económico del comportamiento delictivo de Becker (1968).

Ejemplo 1.1

[Modelo económico del comportamiento delictivo]

En un artículo extraordinario el premio Nobel, Gary Becker, postuló un marco de maximización de la utilidad para describir la participación de una persona en un acto delictivo. Ciertos delitos tienen claras recompensas económicas, pero la mayoría de las conductas delictivas tienen costos. El costo de oportunidad del delito evita que el delincuente desarrolle otras actividades, por ejemplo, desempeñar un empleo legal. Además, hay costos como la posibilidad de ser atrapado y, en caso de ser declarado culpable, el costo del encarcelamiento. Desde la perspectiva de Becker, la decisión de emprender una actividad ilegal es una decisión de asignación de recursos tomando en cuenta los beneficios y los costos de actividades en competencia.

Bajo supuestos generales, es posible deducir una ecuación que describa el tiempo invertido en una actividad delictiva en función de diversos factores. Esta función se puede representar como

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \tag{1.1}$$

donde

y = horas invertidas en actividades delictivas,
x_1 = "salario" por cada hora invertida en la actividad delictiva,
x_2 = salario por hora en un empleo legal,
x_3 = otro ingreso que no provenga ni de la delincuencia ni del empleo,
x_4 = probabilidad de ser atrapado,
x_5 = probabilidad de ser declarado culpable una vez que haya sido atrapado,
x_6 = pena prevista si es declarado culpable y
x_7 = edad.

En general, también influyen otros factores en la decisión de una persona para participar en una actividad delictiva, pero la lista anterior es representativa de los que pueden resultar de un análisis económico formal. Como es usual en la teoría económica, no se ha sido específico acerca de la función $f(\cdot)$ de (1.1). Ésta depende de una función subyacente de utilidad que rara vez se conoce. No obstante, la teoría económica —o introspección— puede emplearse para predecir el efecto que tendrá cada variable en la actividad delictiva. Esta es la base para un análisis econométrico de la actividad delictiva de un individuo.

Algunas veces, el modelado económico formal es el punto de partida del análisis empírico, pero es más común el empleo de teorías económicas menos formales o incluso apoyarse por completo en la intuición. El lector estará de acuerdo en que los determinantes del comportamiento delictivo que se muestran en la ecuación (1.1) están basados en el sentido común; podría haberse llegado de manera directa a esta ecuación, sin necesidad de partir de la maximización de la utilidad. Esta perspectiva también tiene su valor, aunque hay casos en los que una deducción formal hace más claro lo que para la intuición puede pasar inadvertido.

El siguiente es un ejemplo de una ecuación que puede deducirse mediante un razonamiento menos formal.

Ejemplo 1.2

[Capacitación laboral y productividad de los trabajadores]

Considérese el problema planteado al comienzo de la sección 1.1. Un economista laboral desea examinar los efectos de la capacitación sobre la productividad de los trabajadores. En este caso no se necesita una teoría económica formal. Una comprensión básica de la economía es suficiente para advertir que factores tales como la educación, la experiencia y la capacitación laboral afectan la productividad de los trabajadores. Además, los economistas saben que a los trabajadores se les paga en razón de su productividad. Este sencillo razonamiento lleva a un modelo como el siguiente:

$$salario = f(educ, exper, capacitación), \tag{1.2}$$

donde

$$\begin{aligned} salario &= \text{salario por hora,} \\ educ &= \text{años de escolaridad formal,} \\ exper &= \text{años de experiencia laboral y} \\ capacitación &= \text{semanas de capacitación laboral.} \end{aligned}$$

De nuevo hay otros factores que influyen sobre la tasa salarial, pero la ecuación (1.2) encierra la esencia del problema.

Una vez precisado el modelo económico, es necesario transformarlo en lo que se llama un **modelo econométrico**. Dado que a lo largo de este libro trataremos con modelos econométricos, es importante saber cómo se relaciona un modelo econométrico con un modelo económico. Como ejemplo se tomará la ecuación (1.1). Antes de poder emprender un análisis econométrico debe especificarse la forma de la función $f(\cdot)$. Otro problema relacionado con (1.1) es qué hacer con las variables que no pueden ser observadas de manera razonable. Por ejemplo, considere el salario que puede ganar una persona mediante una actividad delictiva. En principio, esa es una cantidad bien definida, pero sería muy difícil, si no imposible, precisar cuál es esta cantidad para un determinado individuo. Incluso variables como la probabilidad de ser detenido no pueden ser evaluadas de manera realista para un individuo determinado, pero se puede tomar nota de las estadísticas de detenciones y deducir una variable que aproxime la probabilidad de ser detenido. Hay, además, muchos otros factores que influyen en el comportamiento delictivo y que no es posible enumerar, y mucho menos precisar, pero que de alguna manera deben ser tomados en cuenta.

Las ambigüedades inherentes al modelo económico de la actividad delictiva se resuelven especificando un modelo econométrico:

$$\begin{aligned} actdelic = \beta_0 + \beta_1 salario_m + \beta_2 otringr + \beta_3 frecdet + \beta_4 frecculp \\ + \beta_5 promsent + \beta_6 edad + u, \end{aligned} \tag{1.3}$$

donde

$$\begin{aligned} actdelic &= \text{una medida de la frecuencia de la actividad delictiva,} \\ salario_m &= \text{salario que puede ganar en el empleo legal,} \\ otringr &= \text{ingresos provenientes de otras fuentes (activos, herencias, etcétera),} \\ frecdet &= \text{frecuencia de las detenciones por delitos anteriores (para aproximar} \\ & \text{ la probabilidad de ser detenido),} \\ frecculp &= \text{frecuencia de ser declarado culpable y} \\ promsent &= \text{duración promedio de la sentencia.} \end{aligned}$$

La elección de estas variables es determinada tanto por la teoría económica como por consideraciones acerca de los datos. El término u comprende factores no precisados, como el salario obtenido por la actividad delictiva, costumbres morales, antecedentes familiares y errores en las mediciones de factores como la actividad delictiva y la probabilidad de ser detenido. Aunque pueden agregarse al modelo variables de antecedentes familiares, tales como cantidad de hermanos, educación de los padres, etc., u no puede eliminarse por completo. En efecto, cómo tratar este *término de error* o *de perturbación* es quizás el componente más importante de todo análisis econométrico.

Las constantes $\beta_0, \beta_1, \ldots, \beta_6$ son los *parámetros* del modelo econométrico y describen dirección y fuerza de la relación entre la *actividad delictiva* y cada uno de los factores empleados para determinar la *actividad delictiva* en el modelo.

Un modelo econométrico completo para el ejemplo 1.2, puede ser

$$salario = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 capacitación + u, \tag{1.4}$$

donde el término u comprende factores como "habilidades innatas", calidad de la educación, antecedentes familiares y otros innumerables factores que influyen en el salario de una persona. Si lo que interesa en concreto es la capacitación laboral, entonces el parámetro de interés es β_3.

La mayoría de las veces, un análisis econométrico inicia con la especificación de un modelo econométrico sin atender a los detalles de la creación del modelo. Aquí, en general, se seguirá este método debido, en gran parte, a que una deducción cuidadosa de un modelo, como el modelo económico de la actividad delictiva, toma mucho tiempo y puede llevar a áreas especializadas, y a menudo complicadas, de la teoría económica. En los ejemplos presentados en este libro, el razonamiento económico será importante, y toda teoría económica subyacente se incorporará en las especificaciones del modelo econométrico. En el ejemplo del modelo económico para la actividad delictiva se empezará con un modelo econométrico como el (1.3) y se usarán el razonamiento económico y el sentido común como guías para la elección de las variables. A pesar de que con este método se pierde algo de la riqueza del análisis económico, suele ser un modelo empleado de manera frecuente y efectiva por investigadores meticulosos.

Una vez que se ha especificado un modelo econométrico como el (1.3) o el (1.4), pueden plantearse diversas *hipótesis* en relación con los parámetros desconocidos. Por ejemplo, en la ecuación (1.3), se puede plantear la hipótesis de que $salario_m$, el salario que puede obtenerse en un empleo legal, no tenga efecto alguno sobre el comportamiento delictivo. En el contexto de este modelo econométrico particular, esta hipótesis es equivalente a $\beta_1 = 0$.

Por definición, en un análisis empírico se necesitan datos. Una vez recolectados los datos sobre las variables relevantes, se emplean los métodos econométricos para estimar los parámetros del modelo econométrico y para probar, formalmente, las hipótesis de interés. En algunos casos, el modelo econométrico se emplea para hacer predicciones, ya sea al probar una teoría o al estudiar el impacto de alguna política.

Dada la gran importancia de la recolección de datos en el trabajo empírico, en la sección 1.3 se describe el tipo de datos que suelen encontrarse.

1.3 Estructura de los datos económicos

Las bases de datos económicos pueden ser de diversos tipos. Aunque algunos métodos econométricos pueden ser empleados, con alguna o ninguna pequeña modificación, para distintos tipos de bases de datos, las características especiales de algunas bases de datos deben ser tomadas en cuenta y aprovecharse. A continuación se describen las estructuras de datos que suelen encontrarse.

Datos de corte transversal

Una **base de datos de corte transversal** consiste en una muestra de individuos, hogares, empresas, ciudades, estados, países u otras unidades, tomada en algún punto dado en el tiempo. Algunas veces no todos los datos de estas unidades corresponden exactamente a un mismo momento. Por ejemplo, puede ser que, un conjunto de familias sea entrevistado durante diferentes semanas de un año. En un análisis de corte transversal puro, diferencias menores de tiempo en la recolección de los datos son ignoradas. Aun cuando un conjunto de familias haya sido entrevistado en semanas distintas de un mismo año, se considerará como una base de datos de corte transversal.

Una característica importante de los datos de corte transversal es que a menudo puede suponerse que han sido obtenidos de la población subyacente mediante un **muestreo aleatorio**. Por ejemplo, si se obtiene información sobre salarios, educación, experiencia y otras características tomando de manera aleatoria 500 personas de la población trabajadora, entonces se tiene una muestra aleatoria de la población de todos los trabajadores. El muestreo aleatorio es el esquema de muestreo que se estudia en los cursos introductorios de estadística y simplifica el análisis de datos de corte transversal. En el apéndice C se encuentra un repaso del muestreo aleatorio.

Algunas veces, el muestro aleatorio no es una premisa apropiada para analizar datos de corte transversal. Por ejemplo, suponga que se desea analizar los factores que intervienen en la acumulación del patrimonio familiar. Se puede entrevistar a un conjunto de familias de una muestra aleatoria, pero algunas de ellas se rehusarán a informar sobre su riqueza. Si, por ejemplo, las familias más acaudaladas están menos dispuestas a revelar su nivel de riqueza, entonces la muestra obtenida sobre el patrimonio no es una muestra aleatoria de la población de todas las familias. Esto ilustra un problema de selección de la muestra, tema avanzado que se verá en el capítulo 17.

Otra violación al muestreo aleatorio ocurre cuando se muestrea de unidades que son grandes con relación a la población, en especial de unidades geográficas. El problema potencial en estos casos es que la población no es tan grande para suponer que las observaciones muestreadas sean independientes. Por ejemplo, si se desea explicar el surgimiento de nueva actividad comercial en los estados, en función de las tasas salariales, los precios de los energéticos, las tasas de impuestos empresariales y prediales, los servicios disponibles, la calidad de la fuerza de trabajo y otras características del estado, es poco probable que las actividades comerciales de estados contiguos sean independientes entre sí. Sin embargo, los métodos econométricos que se estudiarán aquí sí funcionan en estas situaciones, aunque algunas veces deben ser afinados. En general, la complejidad que resulta al analizar tales situaciones será ignorada y estos problemas se tratarán en el marco de un muestreo aleatorio, aun cuando esto no sea técnicamente correcto.

Los datos de corte transversal son muy empleados en economía y en otras ciencias sociales. En economía, el análisis de datos de corte transversal está relacionado de manera estrecha con los campos de la microeconomía aplicada, por ejemplo, economía laboral, finanzas públicas locales y estatales, organización industrial, economía urbana, demografía y economía de la salud. Datos sobre individuos, hogares, empresas y ciudades en un punto dado del tiempo son importantes para probar hipótesis microeconómicas y para evaluar políticas económicas.

Los datos de corte transversal que se emplean en el análisis econométrico pueden representarse y almacenarse en una computadora. La tabla 1.1 contiene, en forma resumida, una base de datos de corte transversal de 526 trabajadores correspondientes a 1976. (Este es un subconjunto de los datos en el archivo WAGE1.RAW.) Las variables son *wage* (salario en dólares por hora), *educ* (años de escolaridad), *exper* (años de experiencia laboral), *female* (mujer, un indicador del género), y *married* (casado, estado marital). Estas dos últimas variables son de carácter binario (cero-uno) y se usan para indicar características cualitativas de la persona (es mujer o no; está casada o no). En el capítulo 7 y posteriores se verá más acerca de las variables binarias.

En la tabla 1.1, la variable *obsno* es el número de observación asignado a cada persona de la muestra. A diferencia de las otras variables, ésta no es una característica de la persona. Todos los paquetes de software para econometría o para estadística asignan un número de observación a cada dato. La intuición indica que en datos como los de la tabla 1.1 no importa qué persona sea catalogada como la observación 1, qué persona sea la observación 2, etc. El hecho de que el orden de los datos no importe es una característica clave del análisis econométrico de bases de datos de corte transversal.

Algunas veces, en datos de corte transversal las distintas variables corresponden a periodos diferentes. Por ejemplo, para determinar los efectos de las políticas públicas sobre el crecimiento económico de largo plazo, los economistas han estudiado la relación entre el crecimiento del producto interno bruto (PIB) real *per cápita* a lo largo de cierto periodo (por ejemplo, 1960 a 1985) y variables determinadas, en parte, por las políticas públicas de 1960 (consumo del gobierno como porcentaje del PIB y tasa de educación secundaria en adultos). Esta base de datos puede representarse como en la tabla 1.2, que es una parte de la base de datos empleada en un estudio sobre tasa de crecimiento realizado por De Long y Summers (1991).

TABLA 1.3

Base de datos de corte transversal sobre salarios y otras características de los individuos

obsno	salario (wage)	educ	exper	mujer (female)	casado (married)
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

TABLA 1.2

Base de datos sobre tasas de crecimiento económico y características de un país

obsno	país	cpibrpc	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Bélgica	2.56	13	69
4	Bolivia	1.24	18	12
⋮	⋮	⋮	⋮	⋮
61	Zimbabwe	2.30	17	6

La variable *cpibrpc* representa el crecimiento promedio del PIB real *per cápita* a lo largo del periodo 1960 a 1985. El hecho de que *govcons60* (consumo del gobierno como porcentaje del PIB) y *second60* (porcentaje de población adulta con educación secundaria) correspondan al año 1960, mientras que *cpibrpc* sea el promedio del crecimiento a lo largo del periodo de 1960 a 1985, no crea ningún problema para considerar esta información como una base de datos de corte transversal. Las observaciones se presentan por país en orden alfabético, el cual no afecta ningún análisis posterior.

Datos de series de tiempo

Una base de **datos de series de tiempo** consiste de las observaciones de una o varias variables a lo largo del tiempo. Ejemplos de datos de series de tiempo son los precios de acciones, la cantidad de dinero en circulación, el índice de precios al consumidor, el producto interno bruto, la tasa anual de homicidios y las cifras de venta de automóviles. Debido a que los eventos pasados pueden influir sobre los eventos futuros y los comportamientos rezagados son frecuentes en las ciencias sociales, el tiempo es una dimensión importante en las bases de datos de series de tiempo. A diferencia de los datos de corte transversal, en una serie de tiempo el orden cronológico de las observaciones proporciona información potencialmente importante.

Una característica fundamental de los datos de series de tiempo, que las hace más difíciles de analizar que los datos de corte transversal, es que rara vez, si acaso, puede suponerse que las observaciones económicas sean independientes en el tiempo. La mayor parte de las series de tiempo económicas y otras series de tiempo están relacionadas, a menudo fuertemente, con sus historias recientes. Por ejemplo, saber algo sobre el producto interno bruto del último trimestre dice mucho acerca del rango probable del PIB durante este trimestre, debido a que el PIB tiende a permanecer bastante estable de un trimestre a otro. Aunque la mayor parte de los procedimientos econométricos pueden usarse tanto con datos de corte transversal como con datos de series de tiempo, la especificación de modelos econométricos para datos de series de tiempo requiere un poco más de trabajo para que se justifique el uso de los métodos econométricos estándar. Además, se han desarrollado modificaciones y embellecimientos de las técnicas econométricas estándar para tomar en cuenta y aprovechar el carácter dependiente de las series de tiempo económicas y para considerar otras cuestiones, como el hecho de que algunas variables económicas tiendan a mostrar una clara tendencia en el tiempo.

Otra característica de las series de tiempo que puede requerir atención especial es la **periodicidad de los datos**, la frecuencia con que éstos se recolectan. En economía, las frecuencias más comunes son diaria, semanal, mensual, trimestral y anual. Así, por ejemplo, los precios de las acciones se publican cada día (con excepción de sábados y domingos); la cantidad de dinero en circulación se publica de manera semanal; muchas series macroeconómicas se registran cada mes, incluidas las tasas de inflación y de desempleo. Otras series macroeconómicas se registran con menos frecuencia, por ejemplo, cada tres meses (o cada trimestre). El producto interno bruto es un ejemplo importante de una serie trimestral. Otras series de tiempo, como la tasa de mortalidad infantil en Estados Unidos, sólo están disponibles anualmente.

Muchas series de tiempo económicas semanales, mensuales o trimestrales muestran un fuerte patrón estacional, que puede ser un factor importante en el análisis de una serie de tiempo. Por ejemplo, los datos mensuales sobre la construcción de vivienda varían con los meses, debido simplemente a la variación de las condiciones climatológicas. En el capítulo 10 se verá cómo trabajar con series de tiempo estacionales.

La tabla 1.3 contiene una base de datos de series de tiempo tomada de un artículo de Castillo-Freeman y Freeman (1992) sobre los efectos del salario mínimo en Puerto Rico. En esta base de datos el año más antiguo, corresponde a la primera observación y el año más disponible corresponde a la última. Al emplear métodos econométricos para analizar datos de series de tiempo, los datos deben almacenarse en orden cronológico.

TABLA 1.3

Salario mínimo, desempleo y datos relacionados con Puerto Rico

obsno	año	prommin	coverprom	desempl	PNB
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
⋮	⋮	⋮	⋮	⋮	⋮
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

La variable *prommin* se refiere al salario mínimo promedio de ese año, *coverprom* es la tasa promedio de cobertura (el porcentaje de los trabajadores protegidos por la ley del salario mínimo), *desempl* es la tasa de desempleo y *PNB* es el producto nacional bruto. Estos datos se emplearán más adelante al analizar el efecto del salario mínimo sobre el desempleo.

Combinación de cortes transversales

Algunas bases de datos tienen características tanto de corte transversal como de series de tiempo. Por ejemplo, suponga que en Estados Unidos se realizan dos encuestas de corte transversal a los hogares, una en 1985 y otra en 1990. En 1985 se encuesta a los hogares de una muestra aleatoria acerca de variables como ingreso, ahorro, tamaño de la familia, etc. En 1990 se toma *otra* muestra aleatoria de hogares usando las preguntas de la encuesta anterior. Para tener un tamaño mayor de la muestra se pueden **combinar los cortes transversales** juntando los dos años.

Combinar (o juntar) los cortes transversales de años distintos suele ser una buena manera de analizar los efectos de las nuevas políticas públicas. La idea es recolectar datos de años anteriores y posteriores al cambio de la política. Considere, por ejemplo, la siguiente base de datos sobre los precios de la vivienda tomados en 1993 y 1995, antes y después de una reducción en el impuesto sobre la propiedad inmobiliaria en 1994. Suponga que de 1993 se tienen datos de 250 casas y de 1995 de 270 casas. Una manera de almacenar esta base de datos es como en la tabla 1.4.

Las observaciones 1 a 250 corresponden a casas vendidas en 1993 y de la 251 a 520 a las 270 casas vendidas en 1995. Aunque el orden en que se almacenen los datos no resulta crucial, conservar el registro del año de cada observación suele ser muy importante. A esto se debe que en esta tabla se ingrese *año* como una variable aparte.

Una combinación de corte transversal se analiza de manera muy parecida a como se analizan los datos de corte transversal, salvo que suelen tomarse en cuenta las diferencias que presentan las variables con el tiempo. En efecto, además de que se incrementa el tamaño de la muestra, lo importante en el análisis de combinaciones de cortes transversales es observar cómo ha cambiado con el tiempo una relación clave.

TABLA 1.4

Combinación de cortes transversales: precios de la vivienda en dos años diferentes

obsno	año	preciov	impprov	piescuadr	recs	baños
1	1993	85 500	42	1600	3	2.0
2	1993	67 300	36	1440	3	2.5
3	1993	134 000	38	2000	4	2.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	1993	243 600	41	2600	4	3.0
251	1995	65 000	16	1250	2	1.0
252	1995	182 400	20	2200	4	2.0
253	1995	97 500	15	1540	3	2.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
520	1995	57 200	16	1100	2	1.5

Datos de panel o longitudinales

Un conjunto de **datos de panel** (o *datos longitudinales*) consiste en una serie de tiempo por *cada* unidad de una base de datos de corte transversal. Por ejemplo, suponga que a lo largo de diez años se registran los datos de salario, educación y empleo de un conjunto de personas; o que durante cinco años se recolecte información sobre inversiones y datos financieros de un mismo conjunto de empresas. Datos de panel también pueden recolectarse sobre unidades geográficas. Por ejemplo, de un conjunto de condados de Estados Unidos pueden recolectarse los datos sobre el flujo de migrantes, tasas de impuestos, tasas salariales, gasto del gobierno, etc., de los años 1980, 1985 y 1990.

La característica fundamental de los datos de panel, que los distingue de las combinaciones de cortes transversales, es que durante un intervalo de tiempo se vigilan las *mismas* unidades (personas, empresas o condados, en los ejemplos precedentes) de un corte transversal. Los datos de la tabla 1.4 no se pueden considerar como una base de datos de panel debido a que es muy probable que las casas vendidas en 1993 sean diferentes de las vendidas en 1995, y aunque haya algunas que se encuentren en los dos años, esta cantidad puede ser demasiado pequeña para tener importancia. En cambio, la tabla 1.5 contiene una base de datos de panel sobre actividad delictiva y estadísticos relacionados de 150 ciudades de Estados Unidos en dos años diferentes.

En dicha tabla hay varios aspectos interesantes. Primero, a cada ciudad se le ha dado un número del 1 al 150. Es irrelevante a qué ciudad se le considera la número 1, la 2, etc. Como ocurre con los datos de corte transversal puro, el orden en el corte transversal de un conjunto de datos de panel no tiene importancia. También podrían haberse usado los nombres de las ciudades, en lugar de los números, pero suele ser útil tener los dos.

TABLA 1.5

Datos de panel, de dos años, sobre estadísticas de delincuencia urbana

obsno	ciudad	año	homicidios	población	desempl	policía
1	1	1986	5	350 000	8.7	440
2	1	1990	8	359 200	7.2	471
3	2	1986	2	64 300	5.4	75
4	2	1990	1	65 100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
297	149	1986	10	260 700	9.6	286
298	149	1990	6	245 000	9.8	334
299	150	1986	25	543 000	4.3	520
300	150	1990	32	546 200	5.2	493

Un segundo punto es que los datos de los dos años de la ciudad 1 son los dos primeros renglones u observaciones. Las observaciones 3 y 4 corresponden a la ciudad 2 y así sucesivamente. Como para cada una de las 150 ciudades se tienen dos filas de datos, cualquier software considerará que se tienen 300 observaciones. Esta base de datos puede tratarse como una combinación de cortes transversales, en la que da la casualidad de que las mismas ciudades aparecen en cada uno de los años. Pero, como se verá en los capítulos 13 y 14, la estructura de panel también puede usarse para analizar cuestiones a las que no se puede responder si este conjunto de datos se ve como una simple combinación de cortes transversales.

En la tabla 1.5, los datos de los dos años se colocan en filas contiguas, ubicando en todos los casos el primer año antes del segundo. Para casi cualquier fin práctico, esta es la manera que se prefiere para ordenar bases de datos de panel. Compárese la organización de estos datos con la manera en la que se han dispuesto los datos en la tabla 1.4. En pocas palabas, la razón para ordenar los datos como en la tabla 1.5 es que se necesitará realizar transformaciones para cada ciudad a lo largo de los dos años.

Dado que en los datos de panel se requiere que las unidades sean las mismas a lo largo del tiempo, las bases de datos de panel, en especial si son de personas, hogares o empresas, son más difíciles de obtener que las combinaciones de cortes transversales. Es comprensible que observar las mismas unidades a lo largo del tiempo tenga diversas ventajas sobre los datos de corte transversal o aun sobre combinaciones de cortes transversales. La ventaja de importancia en este libro es que tener varias observaciones de las mismas unidades permite controlar determinadas características no observadas de las personas, empresas, etc. Como se verá, el uso de más de una observación suele facilitar la inferencia causal en situaciones en las que inferir causalidad sería muy difícil si se contara sólo con un corte transversal. La segunda ventaja de los datos de panel es que permiten estudiar la importancia de desfases de conducta o los resultados de la toma de decisiones. Esta información puede ser significativa debido a que el impacto de muchas políticas económicas sólo puede esperarse después de pasado algún tiempo.

Muchos de los libros que no son para niveles de maestría o doctorado no contienen métodos econométricos para datos de panel. Sin embargo, en la actualidad muchos economistas reconocen que hay algunas cuestiones a las que es difícil, o imposible, responder de manera satisfactoria sin el uso de datos de panel. Como se verá, con el análisis de datos de panel, un método que no es mucho más complicado que el empleo de datos de corte transversal estándar, es posible lograr progresos considerables.

Comentario sobre las estructuras de datos

La primera parte de este libro se ocupa del análisis de datos de corte transversal, debido a que esto presenta menos dificultades técnicas y conceptuales. Al mismo tiempo permite ilustrar la mayor parte de los temas fundamentales del análisis econométrico. Los métodos e ideas del análisis de cortes transversales serán empleados en el resto del libro.

Aunque en el análisis econométrico de las series de tiempo se usan muchas de las herramientas propias del análisis de cortes transversales, el análisis de series de tiempo es más complicado debido a las trayectorias, de carácter altamente persistente, que presentan muchas series económicas. En la actualidad es muy aceptado que los ejemplos tradicionalmente empleados para ilustrar la aplicación de los métodos econométricos a las series de tiempo son inadecuados o erróneos. De manera que en un principio no tiene mucho sentido emplearlos ya que eso sólo reforzará una mala práctica econométrica. Por tanto, el estudio de las series de tiempo econométricas se pospondrá hasta la parte 2 del libro, en la que se introducen los temas que se refieren a tendencia, persistencia, dinámica y estacionalidad.

En la parte 3 se verán explícitamente combinaciones de cortes transversales y datos de panel. El análisis de cortes transversales combinados de manera independiente y el análisis de datos de panel sencillos son extensiones bastante sencillas del análisis de cortes transversales puro. Sin embargo, estos temas se tratarán hasta el capítulo 13.

1.4 Causalidad y la noción de *ceteris paribus* en el análisis econométrico

En la mayoría de las pruebas de teorías económicas, así como en la evaluación de políticas públicas, el objetivo de los economistas es inferir que una variable (por ejemplo, la educación) tiene un **efecto causal** sobre otra variable (por ejemplo, la productividad de los trabajadores). Encontrar simplemente una relación entre dos o más variables puede ser sugestivo, pero no concluyente, a menos que pueda establecerse causalidad.

El concepto ***ceteris paribus*** —"si todos los demás factores relevantes permanecen constantes"— tiene un papel importante en el análisis causal. Esta idea ha estado implícita en parte de lo hasta ahora dicho, en particular en los ejemplos 1.1 y 1.2, pero no se había mencionado de manera explícita.

Es probable que el lector recuerde de los cursos de introducción a la economía, que la mayor parte de las cuestiones económicas tienen un carácter *ceteris paribus*. Por ejemplo, cuando se analiza la demanda del consumidor interesa saber el efecto que tiene una modificación en el precio de un determinado bien sobre la cantidad demandada, mientras todos los demás factores —tales como ingreso, precios de los demás bienes y preferencias individuales— se mantienen constantes. Si no permanecen constantes los demás factores, entonces no se puede saber cuál es el efecto de una modificación en el precio sobre la cantidad demandada.

Mantener los demás factores constantes también es crítico en el análisis de las políticas. En el ejemplo de la capacitación laboral (ejemplo 1.2), interesa conocer, por ejemplo, el efecto de una semana más de capacitación sobre el salario, cuando todos los demás componentes (en particular, educación y experiencia) permanecen sin cambio. Si se logran mantener constantes todos los demás factores relevantes y se encuentra una relación entre capacitación laboral y salarios, puede concluirse que tal capacitación tiene un efecto causal sobre la productividad de los trabajadores. A pesar de que esto puede parecer bastante sencillo, aun ya en este nivel inicial debe ser claro que, salvo en casos muy especiales, no será posible mantener, literalmente, todo lo demás sin cambio. La pregunta fundamental en la mayor parte de los estudios empíricos es: ¿se han mantenido constantes suficientes factores para que se justifique la causalidad? Raras veces un estudio econométrico es evaluado sin que surja esta pregunta.

En la mayoría de las aplicaciones serias, la cantidad de factores que pueden tener influencia sobre una variable —como la actividad delictiva o los salarios— es inmensa y aislar cualquier variable particular puede parecer una tarea imposible. Sin embargo, se verá que empleando cuidadosamente los métodos econométricos pueden simular experimentos *ceteris paribus*.

Por ahora, aún no es posible explicar cómo usar los métodos econométricos para estimar efectos *ceteris paribus*, de manera que se considerarán algunos de los problemas que suelen surgir en la econometría al tratar de inferir causalidad, pero no se empleará ninguna ecuación. En cada ejemplo, el problema de inferir causalidad desaparece si es posible realizar un experimento apropiado. Por tanto, es útil describir cómo se puede estructurar un problema de este tipo, y observar que, en la mayoría de los casos, no es práctico obtener datos experimentales. También es útil reflexionar sobre la razón por la cual los datos con los que se cuenta no tienen las importantes características de las bases de datos experimentales.

Por el momento se confiará en la comprensión intuitiva de términos como *aleatorio*, *independencia* y *correlación*, los cuales ya le deben ser familiares al lector por cursos introductorios de probabilidad y estadística. (En el apéndice B se presenta un repaso de estos conceptos.) Se empezará por dar un ejemplo que ilustra algunos de estos importantes tópicos.

Ejemplo 1.3

[Efecto de un fertilizante en el rendimiento de un cultivo]

En algunos de los primeros estudios econométricos [por ejemplo, Griliches (1957)] se consideró el efecto de nuevos fertilizantes sobre el rendimiento de los cultivos. Suponga que el cultivo en consideración sea de frijol de soya. Como la cantidad de fertilizante es sólo uno de los factores que influyen en el rendimiento —algunos de los factores restantes son precipitación pluvial, calidad de la tierra y presencia de parásitos— este problema debe ser planteado como una cuestión *ceteris paribus*. Una manera de determinar el efecto causal de la cantidad de fertilizante sobre el rendimiento del frijol de soya es realizar un experimento que comprenda los pasos siguientes. Elegir varias parcelas de un acre. Aplicar a cada parcela una cantidad diferente de fertilizante y después medir el rendimiento. Esto proporciona una base de datos de corte transversal. Después, se emplean los métodos estadísticos (a introducirse en el capítulo 2) para medir la relación entre rendimiento y cantidad de fertilizante.

Como se ha descrito, éste puede no ser un muy buen experimento, ya que no se ha dicho nada acerca de la elección de las parcelas de tierra que son idénticas en todos los aspectos excepto en la cantidad de fertilizante. En efecto, no es posible elegir parcelas de terreno con esta característica: hay algunos factores, como la calidad de la tierra, que no pueden ser bien observados. ¿Cómo se sabe que los resultados de este experimento puedan emplearse para medir el efecto *ceteris paribus* del fertilizante? La respuesta depende de los detalles específicos para elegir las cantidades de fertilizantes. Si éstas son asignadas a las parcelas, independientemente de otros factores que influyan en el rendimiento, es decir, al elegir las cantidades de fertilizante se ignoran por completo otras características de las parcelas, entonces vamos por buen camino. Esta aseveración se justificará en el capítulo 2.

El ejemplo siguiente es más representativo de las dificultades que surgen para inferir causalidad en la economía aplicada.

Ejemplo 1.4

[Medición del rendimiento de la educación]

Desde hace tiempo, economistas laborales y hacedores de política pública han estado interesados en el "rendimiento de la educación". De manera un poco informal, la cuestión se puede plantear como sigue: si se toma una persona de la población y se le da un año más de educación, ¿en cuánto aumentará su salario? Como en el ejemplo anterior, esta es una cuestión *ceteris paribus*, lo que significa que todos los demás factores se mantienen constantes mientras se da a la persona un año más de educación.

Uno podría imaginar que un encargado de la planeación social podría diseñar un experimento para resolver este problema, así como investigadores agrícolas diseñan un experimento para estimar los efectos de los fertilizantes. Suponga, por el momento, que el encargado de la planeación social tiene la capacidad de asignarle cualquier nivel de educación a cualquier persona. ¿Cómo puede emular este planificador el experimento del fertilizante del ejemplo 1.3? El planificador puede elegir un grupo de personas y asignar a cada una de manera aleatoria una cantidad de educación. A algunas se les asignará una educación hasta el sexto grado, a otras hasta secundaria, a otras hasta dos años de universidad y así sucesivamente. A continuación, el encargado de la planeación puede medir los sueldos de las personas de este grupo (suponiendo que todas cuenten con un empleo). Las personas corresponden a las parcelas del ejemplo de los fertilizantes, la educación juega el papel de los fertilizantes y el salario corresponde a los rendimientos de frijol de soya. Como en el ejemplo 1.3, si los niveles de educación se asignan de manera independiente respecto a las demás características que afectan la productividad (por ejemplo, experiencia y capacidades innatas), los resultados de un análisis en el que se ignoren estos otros factores serán útiles. Una vez más, en el capítulo 2 se justificará esta aseveración; por ahora se enuncia sin mayor soporte.

A diferencia del ejemplo de los fertilizantes y el rendimiento, el experimento que se describe en el ejemplo 1.4 no es factible. Los aspectos éticos, por no mencionar los costos económicos, que implica determinar aleatoriamente los niveles de educación de un grupo de personas son obvios. Además, como asunto de logística, a una persona no se le podría asignar como nivel de educación el sexto grado si ya cuenta con un grado universitario.

Aunque no se puedan obtener datos experimentales que permitan medir el rendimiento de la educación, se pueden recolectar, de un grupo grande, datos no experimentales sobre niveles de educación y salarios, tomando una muestra aleatoria de la población formada por las personas trabajadoras. Datos de este tipo se encuentran disponibles a partir de las diversas encuestas empleadas en la economía laboral, sin embargo, estas bases de datos tienen una característica que dificulta estimar el rendimiento *ceteris paribus* de la educación. Son las propias personas las que *eligen* su nivel de educación; por tanto, es probable que los niveles de educación no sean independientes de todos los demás factores que afectan el salario. Este problema es una característica común a la mayor parte de las bases de datos no experimentales.

Uno de los factores que influyen en el salario es la experiencia laboral. Dado que, en general, tener más educación suele requerir que se posponga el ingreso a la fuerza de trabajo, las personas con más educación por lo general tienen menos experiencia. Por tanto, en las bases de datos no experimentales sobre salarios y educación, es muy probable que la educación esté relacionada de manera negativa con una importante variable que afecta el salario. También se cree que las personas con mayores habilidades innatas suelen elegir niveles superiores de educación. Como mejores habilidades llevan a salarios más altos, una vez más se observa una correlación entre educación y factores críticos que afectan el salario.

Los factores omitidos, experiencia y habilidad, en el ejemplo del salario tienen sus análogos en el ejemplo del fertilizante. La experiencia en general es fácil de medir y por tanto es similar a una variable tal como la precipitación pluvial. Por otro lado, las habilidades son algo nebuloso y difícil de cuantificar; en el ejemplo del fertilizante puede compararse con la calidad de la tierra. Como se verá a lo largo de este libro, tomar en cuenta otros factores observados, como la experiencia, al estimar el efecto *ceteris paribus* de otra variable, como la educación, es relativamente sencillo. También se verá que tomar en cuenta factores inherentes no observables, como las habilidades, es mucho más problemático. Es justo señalar que muchos de los avances en los métodos econométricos han tratado de ocuparse de los factores no observados en los modelos econométricos.

Hay un último paralelo que se puede observar entre los ejemplos 1.3 y 1.4. Suponga que en el ejemplo del fertilizante las cantidades de este producto no fueron determinadas de manera totalmente aleatoria; sino que la persona encargada de elegirlas pensó que sería mejor aplicar más fertilizante en las parcelas con tierra de mejor calidad. (Los investigadores agrícolas tienen una buena idea de qué parcelas de terreno son de mejor calidad, aun cuando no puedan cuantificar con exactitud estas diferencias.) Esta situación es análoga a relacionar el nivel de escolaridad con la habilidad no observada en el ejemplo 1.4. Dado que una mejor tierra da rendimientos mayores y que en las mejores parcelas se utilizó más fertilizante, la relación entre rendimiento y fertilizante puede ser espuria.

Ejemplo 1.5

[Efecto de la presencia de la policía sobre los niveles de delincuencia urbana]

El tema de cuál será la mejor manera de prevenir la delincuencia ha sido, y tal vez seguirá siendo de actualidad. Una de las cuestiones especialmente importantes a este respecto es si la presencia de más agentes de policía en las calles servirá para detener la delincuencia

La pregunta *ceteris paribus* es sencilla: si se elige aleatoriamente una ciudad y dados, 10 policías adicionales, por ejemplo, ¿en cuánto disminuirá la tasa de delincuencia? Otra manera de plantear esta pregunta es: si dos ciudades son iguales en todos los aspectos, salvo en que en la ciudad A haya 10 policías más que en la ciudad B, ¿de cuánto será la diferencia entre las tasas de delincuencia de las dos ciudades?

Es casi imposible hallar un par de comunidades idénticas en todos los aspectos, excepto en el tamaño de sus fuerzas policiacas. Por fortuna, esto no es necesario en el análisis econométrico. Lo que se necesita saber es si los datos que pueden recolectarse acerca de niveles de delincuencia y tamaño de las fuerzas policiacas en una comunidad pueden ser considerados experimentales. Podría uno imaginar un experimento en el que se tenga una cantidad grande de ciudades y a cada ciudad se le asignara una cantidad determinada de policías para el siguiente año.

Aunque hay políticas que se pueden emplear para influir en el tamaño de las fuerzas policiacas, no es posible indicar a cada ciudad cuántos policías debe de emplear. Si, como es probable, la elección de cuántos policías emplear está relacionada con otros factores, propios de la ciudad, que influyen sobre la delincuencia, entonces los datos deben verse como datos no experimentales. De hecho, una manera de enfocar este problema es considerar que la elección del tamaño de la fuerza policiaca de una ciudad y el volumen de la criminalidad son *determinados de forma simultánea*. Estos problemas se abordarán de manera explícita en el capítulo 16.

En los tres primeros ejemplos vistos, los datos han sido de corte transversal a diferentes niveles de agregación (por ejemplo, a niveles individuales o de una ciudad). Los mismos obstáculos surgen para inferir causalidad en problemas con series de tiempo.

Ejemplo 1.6

[Efectos del salario mínimo sobre el desempleo]

Un importante, y quizá controversial, asunto político es el que se refiere al efecto del salario mínimo sobre las tasas de desempleo en diversos grupos de trabajadores. Aunque este problema puede ser estudiado con diversos tipos de datos (corte transversal, series de tiempo o datos de panel), suelen usarse las series de tiempo para observar los efectos agregados. En la tabla 1.3 se presenta un ejemplo de una base de datos de series de tiempo sobre tasas de desempleo y salarios mínimos.

El análisis estándar de la oferta y la demanda implica que a medida que el salario mínimo aumenta más allá del salario del equilibrio del mercado, nos movemos sobre la curva de demanda de trabajo y el empleo total disminuye. (La oferta de trabajo supera la demanda.) Para cuantificar este efecto, se estudia la relación entre empleo y salario mínimo a lo largo del tiempo. Además de algunas dificultades especiales que pueden surgir al usar datos de series de tiempo, se tienen posibles problemas para inferir causalidad. El salario mínimo en Estados Unidos no se determina en el vacío. Existen varias fuerzas económicas y políticas que afectan el salario mínimo de un determinado año. (El salario mínimo, una vez determinado, suele mantenerse durante varios años, a menos que sea indexado —ajustado— a la inflación.) Por tanto, es probable que el monto del salario mínimo esté relacionado con otros factores que tienen efecto sobre los niveles de desempleo.

Imagine un experimento hipotético que podría realizar un gobierno interesado en determinar los efectos del salario mínimo sobre el desempleo (en lugar de preocuparse por la protección de los trabajadores de bajos ingresos). Cada año el gobierno podría determinar de forma aleatoria el salario mínimo y registrar después los resultados sobre el empleo. Después, empleando métodos econométricos bastante sencillos, podrían examinarse los datos experimentales de series de tiempo obtenidos. Sin embargo, esto está muy lejos de describir la manera en que se fijan los salarios mínimos.

Si se puede tener suficiente control sobre otros factores relacionados con el empleo, entonces es posible estimar el efectos *ceteris paribus* del salario mínimo sobre el empleo. En este sentido, el problema es muy parecido a los ejemplos anteriores con datos de corte transversal.

Aun cuando las teorías económicas no son descritas de la manera más natural en términos de causalidad, suelen tener predicciones que pueden probarse empleando métodos econométricos. Esto se demuestra en el siguiente ejemplo.

Ejemplo 1.7

[Hipótesis de las expectativas]

La *hipótesis de las expectativas* de la economía financiera establece que si a los inversionistas se les da toda la información disponible en el momento de invertir, el rendimiento *esperado* en cualquiera de dos inversiones es el mismo. Considérense, por ejemplo, dos inversiones con un horizonte de inversión a tres meses, adquiridas al mismo tiempo: 1) Se compran certificados del tesoro (Estados Unidos) a tres meses con valor nominal de 10 000 dólares a un precio inferior a 10 000 dólares, y en tres meses se obtienen 10 000 dólares. 2) Se compran certificados del tesoro (Estados Unidos) a seis meses (a un precio inferior a 10 000 dólares) y en tres meses se venden como certificados del tesoro a tres meses. Para las dos inversiones se necesita aproximadamente el mismo capital inicial, pero hay una diferencia importante. En la primera inversión, en el momento de la compra se sabe con exactitud cuál será el rendimiento, debido a que se conoce el precio de los certificados del tesoro a tres meses, así como su valor nominal. Con la segunda inversión no ocurre lo mismo: aunque en el momento de la compra se sabe cuál es el precio de los certificados del tesoro a seis meses, no se conoce el precio al que se pueden vender en tres meses. Por tanto, esta inversión presenta incertidumbre para alguien que tenga un horizonte de inversión de tres meses.

Por lo general, el rendimiento de estas dos inversiones será diferente. Con base en la hipótesis de las expectativas, el rendimiento esperado en la segunda inversión, contando con toda la información al momento de la inversión, deberá ser igual al rendimiento de la compra de certificados del tesoro a tres meses. Como se verá en el capítulo 11, esta teoría resulta bastante fácil de probar.

RESUMEN

En este capítulo introductorio se han presentado los objetivos y el campo de acción del análisis econométrico. La econometría se emplea en todas las áreas de la economía aplicada para probar teorías, proporcionar información a los encargados de elaborar las políticas públicas y privadas, y predecir series de tiempo económicas. Los modelos econométricos algunas veces se obtienen de modelos económicos formales, pero otras veces están basados en razonamientos económicos informales y en la intuición. Los objetivos de cualquier análisis econométrico son estimar los parámetros de un modelo y probar hipótesis acerca de ellos; los valores y signos de los parámetros determinan la validez de una teoría económica y los efectos de determinadas políticas.

Los tipos de estructuras de datos más comúnmente empleados en la econometría aplicada son los datos de corte transversal, las series de tiempo, las combinaciones de cortes transversales y los datos de panel. Los conjuntos de datos en los que interviene una dimensión de tiempo, como las series de tiempo y los datos de panel, requieren un tratamiento especial debido a la correlación en el tiempo de la mayoría de las series económicas. Otras cuestiones, como las tendencias y la estacionalidad, surgen en las series de tiempo, pero no en los datos de corte transversal.

En la sección 1.4 se trataron las nociones de *ceteris paribus* y de inferencia causal. En la mayoría de los casos, las hipótesis en las ciencias sociales son de carácter *ceteris paribus*, es decir, para estudiar una relación entre dos variables todos los demás factores relevantes deben mantenerse constantes. En las ciencias sociales, dado el carácter no experimental de la mayor parte de los datos que suelen recolectarse, hallar relaciones causales no es una tarea fácil.

TÉRMINOS CLAVE

Análisis empírico
Bases de datos de corte transversal
Ceteris paribus
Combinación de cortes transversales
Datos de panel
Datos de series de tiempo
Datos experimentales
Datos no experimentales
Datos observacionales
Datos retrospectivos
Efecto causal
Modelo econométrico
Modelo económico
Muestreo aleatorio
Periodicidad de los datos

PROBLEMAS

1.1 Suponga que se le pide que realice un estudio para determinar si grupos de clase pequeños contribuyen a un mejor desempeño de los estudiantes de cuarto grado.

i) Si pudiera realizar cualquier experimento que deseara, ¿qué haría? Explique con claridad.

ii) Siendo más realistas, suponga que puede obtener datos observacionales de varios miles de estudiantes de cuarto grado de un determinado estado. Puede conocer el tamaño de sus grupos y las calificaciones estandarizadas obtenidas en el examen final. ¿Por qué puede esperarse una correlación negativa entre el tamaño de los grupos y las puntuaciones en el examen final?

iii) Una correlación negativa, ¿indicaría necesariamente que tamaños de grupo menores causan un mejor desempeño?

1.2 Para justificar los programas de capacitación laboral se ha dicho que éstos mejoran la productividad de los trabajadores. Suponga que se le pide que evalúe si una mayor capacitación para el trabajo hace que los trabajadores sean más productivos. Pero, en lugar de que se le proporcionen datos sobre trabajadores individuales, se le facilitan datos de fábricas en Ohio. De cada firma se le proporcionan horas de capacitación laboral por trabajador (*capacitación*) y la cantidad de artículos no defectuosos producidos por hora por cada trabajador (*producción*).

i) Establezca cuidadosamente el experimento *ceteris paribus* subyacente a esta pregunta.

ii) ¿Parece razonable que la decisión de una empresa de capacitar a sus trabajadores sea independiente de las características de los mismos? ¿Cuáles son algunas de esas características medibles y no medibles de los trabajadores?

iii) Nombre un factor, que no sea una característica de los trabajadores, que influya en la productividad de los trabajadores.

iv) Si encontrara una correlación positiva entre *producción* y *capacitación*, ¿habría establecido de manera convincente que la capacitación para el trabajo hace que los trabajadores sean más productivos? Explique.

1.3 Suponga que en su universidad se le pide que encuentre una relación entre horas semanales de estudio (*estudio*) y horas semanales de trabajo (*trabajo*). ¿Tendría sentido considerar que en este problema se trata de inferir si *estudio* “causa” *trabajo* o *trabajo* “causa” *estudio*? Explique.

EJERCICIOS EN COMPUTADORA

C1.1 Para este ejercicio emplee la base de datos WAGE1.RAW.

i) Determine el nivel educativo promedio de la muestra. ¿Cuáles son los niveles de educación menor y mayor?

ii) Determine el salario promedio por hora (*wage*) en la muestra. ¿Parece ser alto o bajo?

iii) Los datos de los salarios están dados en dólares de 1976. Usando el *Economic Report of the President* (de 2004 o posterior o el Informe de Gobierno en países de habla hispana) obtenga y dé los índices de precios al consumidor (IPC) correspondientes a 1976 y 2003.

iv) Use los valores de los IPC del inciso iii) para determinar el salario promedio por hora en dólares de 2003. ¿Parece ahora más razonable el salario promedio por hora?

v) ¿Cuántas mujeres (*females*) hay en la muestra? ¿Cuántos hombres?

C1.2 Para responder estas preguntas emplee la base de datos BWGHT.RAW.

i) ¿Cuántas mujeres hay en la muestra (male = 0) y cuántas de las informantes fumaron durante un embarazo?

ii) ¿Cuál es la cantidad promedio de cigarros consumidos por día (cigs)? ¿Es el promedio, en este caso, una medida representativa de la mujer “típica”? Explique.

iii) Entre las mujeres que fumaron durante el embarazo, ¿cuál es la cantidad promedio de cigarros consumidos por día? ¿Cuál es la relación de esto con su respuesta al inciso ii) y por qué?

iv) Determine el promedio de *fatheduc* (años de educación del padre) en la muestra. ¿Por qué se emplean sólo 1192 observaciones para calcular este promedio?

v) Dé el ingreso familiar promedio (*famine*) y su desviación estándar en dólares.

C1.3 Los datos de MEAP01.RAW pertenecen al estado de Michigan en el año 2001. Emplee estos datos para contestar las preguntas siguientes.

i) Determine los valores mayor y menor de *math4*. ¿Es lógico este intervalo? Explique.

ii) ¿Cuántas escuelas tienen una tasa perfecta de aprobados en el examen de matemáticas? ¿A qué porcentaje del total de la muestra corresponde esta cantidad?

iii) ¿En cuántas escuelas la tasa de aprobados en matemáticas es exactamente 50%?

iv) Compare el promedio de las tasas de aprobados en matemáticas y en lectura. ¿Cuál de estas pruebas es más difícil de aprobar?

v) Encuentre la correlación entre *math4* y *read4*. ¿Qué concluye?

vi) La variable *exppp* es gasto por alumno. Determine el promedio y la desviación estándar de *exppp*. ¿Parece haber una gran variación en el gasto por alumno?

vii) Suponga que la escuela A gasta 6 000 dólares por alumno y la escuela B gasta 5 000 dólares por alumno. Dé el porcentaje en el que el gasto de la escuela A supera al gasto de la escuela B. Compare este porcentaje con $100 \cdot [\log(6\,000) - \log(5\,500)]$, que es la diferencia porcentual aproximada basada en la diferencia de los logaritmos naturales. (Veáse la sección A.4 del apéndice A.)

C1.4 La base de datos de JTRAIN2.RAW proviene de un experimento de capacitación para el trabajo realizado para hombres con bajos ingresos durante 1976-1977; véase Lalonde (1986).

i) Emplee la variable indicadora *train* para determinar la proporción de hombres a los que se les dio capacitación para el trabajo.

ii) La variable *re78* es ingresos desde 1978, dados en dólares de 1982. Determine el promedio de *re78* para la muestra de hombres a los que se les dio capacitación laboral y para la muestra de hombres a los que no se les dio. ¿Es esta diferencia económicamente grande?

iii) La variable *unem78* indica si un hombre estuvo desempleado o no en 1978. ¿Qué proporción de los hombres a los que se les dio capacitación para el trabajo están desempleados? ¿Y de aquellos a los que no se les dio capacitación laboral? Comente la diferencia.

iv) Con base en los incisos ii) y iii), ¿parece haber sido efectivo el programa de capacitación laboral? ¿Qué haría que nuestra conclusión fuera más convincente?

PARTE 1

Análisis de regresión con datos de corte transversal

En la parte 1 de este libro se estudiará el análisis de regresión empleando datos de corte transversal. Esta parte se apoya en el álgebra universitaria y en los conceptos básicos de probabilidad y estadística. En los apéndices A, B y C se presenta un repaso completo de estos temas.

El capítulo 2 empieza con el modelo de regresión lineal simple que explica una variable en términos de otra. Aunque la regresión simple no es muy empleada en la econometría aplicada, se usa de manera ocasional y sirve como punto de partida debido a que tanto el álgebra como las interpretaciones son relativamente sencillas.

En los capítulos 3 y 4 se ven los fundamentos del análisis de regresión múltiple, en la que se considera más de una variable que afecta a aquella que se trata de explicar. La regresión múltiple aún es el método más empleado en la investigación empírica y, por tanto, estos capítulos merecen una atención especial. En el capítulo 3 la atención se centra en el álgebra del método de mínimos cuadrados ordinarios (MCO) y también se establecen las condiciones bajo las cuales un estimador MCO es insesgado y el mejor estimador lineal insesgado. En el capítulo 4 se estudia el importante tema de la inferencia estadística.

En el capítulo 5 se analizan las propiedades de los estimadores MCO en muestras grandes, o asintóticas. Esto proporciona una justificación para los procedimientos de inferencia del capítulo 4, cuando los errores en un modelo de regresión no están distribuidos normalmente. En el capítulo 6 se tratan algunos otros temas sobre el análisis de regresión, incluyendo cuestiones como temas avanzados sobre formas funcionales, escalamiento de datos, predicción y bondad de ajuste. En el capítulo 7 se explica cómo se puede incorporar la información cualitativa a los modelos de regresión múltiple.

En el capítulo 8 se ilustra cómo probar y corregir el problema de la heterocedasticidad, o varianza no constante, en términos del error. Se muestra cómo pueden ajustarse los estadísticos MCO comunes y también se presenta una extensión de MCO, conocida como *mínimos cuadrados ponderados*, que considera explícitamente varianzas diferentes en los errores. En el capítulo 9 se profundiza un poco más en el importante problema de la correlación entre el término del error y una o más de las variables explicativas; se demuestra cómo la disponibilidad de una variable proxy puede resolver el problema de variables omitidas. Además, se establece el sesgo y la inconsistencia en los estimadores MCO en presencia de cierto tipo de errores de medición en las variables. También se discuten varios problemas, en los datos, como el de las observaciones atípicas.

CAPÍTULO 2

El modelo de regresión simple

El modelo de regresión simple puede utilizarse para estudiar la relación entre dos variables. Por razones que se verán más adelante, este modelo presenta limitaciones como herramienta general para el análisis empírico. No obstante, algunas veces es adecuado como herramienta empírica. Aprender a interpretar el modelo de regresión simple es una buena práctica para estudiar la regresión múltiple, lo cual se hará en capítulos subsiguientes.

2.1 Definición del modelo de regresión simple

Gran parte de los análisis en econometría aplicada parten de la premisa siguiente: y y x son dos variables que representan alguna población y se desea "explicar y en términos de x" o "estudiar cómo varía y cuando varía x". En el capítulo 1 se vieron algunos ejemplos: y es rendimiento de cultivos de frijol de soya y x la cantidad de fertilizante; y es salario por hora y x los años de educación escolar y y es la tasa de delincuencia en una comunidad y x la cantidad de policías.

Para establecer un modelo que "explique y en términos de x" hay que tomar en consideración tres aspectos. Primero, dado que entre las variables nunca existe una relación exacta, ¿cómo pueden tenerse en cuenta otros factores que afecten a y? Segundo, ¿cuál es la relación funcional entre y y x? Y, tercero, ¿cómo se puede estar seguro de que la relación entre y y x sea una relación *ceteris paribus* entre y y x (si es ese el objetivo buscado)?

Estas ambigüedades pueden resolverse estableciendo una ecuación que relacione y con x. Una ecuación sencilla es

$$y = \beta_0 + \beta_1 x + u. \tag{2.1}$$

La ecuación (2.1), que se supone válida en la población de interés, define el **modelo de regresión lineal simple**. A esta ecuación también se le llama *modelo de regresión lineal de dos variables* o *modelo de regresión lineal bivariada* debido a que en este modelo se relacionan las dos variables x y y. A continuación se analizará el significado de cada una de las cantidades que aparecen en la ecuación (2.1). [Dicho sea de paso, el origen del término "regresión" no tiene una importancia especial para la mayoría de las aplicaciones econométricas modernas, por lo que no se explicará aquí. Vea la historia del análisis de regresión en Stigler (1986).]

Cuando las variables y y x se relacionan mediante la ecuación (2.1) se les da diversos nombres que se usan indistintamente: a y se le conoce como la **variable dependiente**, la **variable explicada**, la **variable de respuesta**, la **variable predicha** o el **regresando**; a x se le conoce como la **variable independiente**, la **variable explicativa**, la **variable de control**, la **variable predictora** o el **regresor**. (Para x también se usa el término **covariada**.) En econometría con frecuencia se usan los términos "variable dependiente" y "variable independiente". Pero hay que hacer notar que aquí el término "independiente" no se refiere al concepto estadístico de independencia entre variables aleatorias (vea el apéndice B).

Los términos variables "explicada" y "explicativa" son probablemente los más descriptivos. "Respuesta" y "control" se usan más en las ciencias experimentales, en donde la variable x está bajo el control del experimentador. Aquí no se usarán los términos "variable predicha" y "predictor", aunque éstos a veces se encuentran en aplicaciones relacionadas sólo con la predicción y no con la causalidad. La terminología que se empleará aquí para la regresión simple se resume en la tabla 2.1.

La variable u, llamada **término de error**, o **perturbación** en la relación, representa factores distintos a x que afectan a y. Un análisis de regresión simple en realidad trata a todos los factores que afectan a y, y que son distintos a x como factores no observados. Es útil considerar a u como abreviación de "unobserved" (no observado, en inglés).

La ecuación (2.1) también resuelve el problema de la relación funcional entre y y x. Si los demás factores en u permanecen constantes, de manera que el cambio en u sea cero, $\Delta u = 0$, entonces x tiene un efecto *lineal* sobre y:

$$\Delta y = \beta_1 \Delta x \text{ si } \Delta u = 0. \tag{2.2}$$

Por tanto, el cambio en y es simplemente β_1 multiplicado por el cambio en x. Esto significa que β_1 es el **parámetro de la pendiente** en la relación entre y y x, cuando todos los demás factores en u permanecen constantes; este parámetro es de interés primordial en la economía aplicada. El **parámetro del intercepto** β_0, algunas veces llamado *término constante*, tiene también su utilidad, aunque es raro que tenga una importancia central en el análisis.

TABLA 2.1

Terminología en la regresión simple

y	x
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictora
Regresando	Regresor

Ejemplo 2.1

[Rendimiento del frijol de soya y el fertilizante]

Suponga que el rendimiento del frijol de soya está determinado por el modelo

$$rendimiento = \beta_0 + \beta_1 fertilizante + u, \tag{2.3}$$

de manera que $y = rendimiento$ y $x = fertilizante$. Al investigador agrícola le interesa el efecto del fertilizante sobre el rendimiento, cuando todos los demás factores permanecen constantes. Este efecto está dado por β_1. El término del error u comprende factores como calidad de la tierra, precipitación pluvial, etc. El coeficiente β_1 mide el efecto del fertilizante sobre el rendimiento, cuando todos los demás factores permanecen constantes: $\Delta rendimiento = \beta_1 \Delta\, fertilizante$.

Ejemplo 2.2

[Una ecuación sencilla para el salario]

Un modelo en el que se relaciona el salario de una persona con la educación observada y con otros factores no observados es

$$salario = \beta_0 + \beta_1 educ + u. \tag{2.4}$$

Si *salario* se mide en dólares por hora y *educ* se mide en años de educación, entonces β_1 mide la variación en el salario por hora por cada año más de educación, cuando todos los demás factores permanecen constantes. Entre estos factores se encuentran experiencia laboral, capacidades innatas, antigüedad en el empleo actual, ética laboral y otra gran cantidad de cosas.

La linealidad de la ecuación (2.1) implica que todo cambio de x en una unidad tiene siempre el *mismo* efecto sobre y, sin importar el valor inicial de x. En muchas aplicaciones de la economía esto no es muy realista. Así, en el ejemplo del salario y la educación, es deseable permitir que haya rendimientos *crecientes*: un año más en educación escolar debe tener un efecto *mayor* que el que tuvo el año anterior. En la sección 2.4 se verá cómo tener estas posibilidades.

El problema más difícil de abordar es si el modelo dado por la ecuación (2.1) en realidad permite formular conclusiones *ceteris paribus* acerca de cómo afecta x a y. Se acaba de ver que en la ecuación (2.2) β_1 *mide* el efecto de x sobre y, cuando todos los demás factores (en u) permanecen constantes. ¿Resuelve esto el problema de la causalidad? Por desgracia, no. ¿Cómo esperar conocer el efecto *ceteris paribus* de x sobre y, cuando todos los demás factores permanecen constantes, si se ignoran esos otros factores?

En la sección 2.5 se muestra que la única manera de obtener estimadores confiables de β_0 y β_1 a partir de los datos de una muestra aleatoria, es haciendo una suposición que restrinja la manera en que la variable no observable u está relacionada con la variable explicativa x. Sin esta restricción, no es posible estimar el efecto *ceteris paribus*, β_1. Como u y x son variables aleatorias, se necesita un concepto basado en la probabilidad.

Antes de establecer esta suposición clave acerca de la relación entre x y u se puede hacer una suposición acerca de u. En tanto el intercepto β_0 aparezca en la ecuación, nada se altera al suponer que el valor promedio de u en la población, es cero. Matemáticamente,

$$\mathrm{E}(u) = 0. \tag{2.5}$$

El supuesto (2.5) no dice nada acerca de la relación entre u y x, sólo afirma algo acerca de la distribución de los efectos no observables en la población. Al usar como ilustración los ejemplos anteriores, puede verse que el supuesto (2.5) no es muy restrictivo. En el ejemplo 2.1, no se modifica nada al normalizar los factores no observados que afectan el rendimiento del frijol de soya, por ejemplo la calidad de la tierra, para hacer que en la población de todas las parcelas cultivadas su promedio sea cero. Lo mismo ocurre con los factores no observados del ejemplo 2.2. Sin pérdida de la generalidad, se puede suponer que en la población de todas las personas trabajadoras, cosas como la capacidad promedio sea cero. Si no se está convencido de esto, puede resolverse el problema 2.2 para comprobar que en la ecuación (2.1) siempre es posible redefinir el intercepto con el eje Y para hacer que la ecuación (2.5) sea verdadera.

Ahora, volvamos al supuesto crucial sobre la manera en que están relacionadas u y x. Una medida natural de la relación entre dos variables aleatorias es el *coeficiente de correlación*. (Vea definición y propiedades en el apéndice B.) Si u y x *no están correlacionadas*, entonces, como variables aleatorias, no están relacionadas *linealmente*. Suponer que u y x no están correlacionadas es un avance para definir el sentido en el que u y x estarán relacionadas en la ecuación (2.1). Sin embargo, el avance no es suficiente, ya que la correlación sólo mide dependencia lineal entre u y x. La correlación tiene una propiedad un poco contraintuitiva: es posible que u no esté correlacionada con x y que, sin embargo, esté correlacionada con funciones de x como, por ejemplo, x^2. (Vea una mayor explicación en la sección B.4.) Esta posibilidad no es aceptable para la mayoría de los propósitos de la regresión, ya que causa problemas para interpretar el modelo y obtener propiedades estadísticas. Un supuesto mejor envuelve el *valor esperado de u dado x*.

Como u y x son variables aleatorias, se puede definir la distribución condicional de u dado cualquier valor de x. En particular, para cada x, se puede obtener el valor esperado (o promedio) de u en la porción de la población descrita por el valor de x. El supuesto crucial es que el valor promedio de u *no* depende del valor de x. Este supuesto se expresa como

$$\mathrm{E}(u|x) = \mathrm{E}(u). \tag{2.6}$$

La ecuación (2.6) indica que el valor promedio de los factores no observables es el mismo en todas las fracciones de la población determinados por los valores de x y que este promedio común es necesariamente igual al promedio de u en toda la población. Cuando se satisface el supuesto (2.6) se dice que u es **media independiente** de x. (Por supuesto, la independencia de la media es una consecuencia de la independencia entre u y x, un supuesto usado frecuentemente en probabilidad y estadística básicas.) Combinando la independencia de la media con el supuesto (2.5), se obtiene el **supuesto de media condicional cero**, $\mathrm{E}(u|x) = 0$. Es vital recordar que la ecuación (2.6) es el supuesto importante; el supuesto (2.5) sólo define el intercepto, β_0.

¿Qué conlleva la ecuación (2.6) en el ejemplo del salario? Para simplificar el análisis, supóngase que u es capacidad innata. Entonces la ecuación (2.6) requiere que el promedio de la capacidad sea el mismo sin importar los años de educación escolar. Por ejemplo, si $\mathrm{E}(capcaci|8)$ denota las capacidades promedio en el grupo de personas con ocho años de educación escolar y $\mathrm{E}(capcaci|16)$ denota las capacidades promedio entre todas las personas con 16 años de educación escolar, entonces la ecuación (2.6) implica que estos valores deben ser iguales. En efecto, el promedio de capacidad debe ser el mismo en *todos* los niveles de educación. Si, por ejemplo, se piensa que la capacidad promedio aumenta con los años de educación, entonces la ecuación (2.6) es falsa. (Esto ocurriría si, en promedio, las personas con mayor capacidad eligieran tener más educación.) Como las capacidades innatas no pueden ser observadas, no hay manera de saber si la capacidad promedio es la misma en todos los niveles de educación. Pero esta es una cuestión que debe ser tomada en consideración antes de confiar en el análisis de regresión simple.

Pregunta 2.1

Suponga que la calificación en un examen final, *score*, depende de las asistencias a las clases (*attend*) y de factores no observados que influyen sobre el desempeño en el examen (por ejemplo, habilidad del estudiante). Entonces

$$score = \beta_0 + \beta_1 attend + u. \tag{2.7}$$

¿En qué casos se esperará que este modelo satisfaga la ecuación (2.6)?

En el ejemplo del fertilizante, si las cantidades de fertilizante se eligen independientemente de otras características de las parcelas, entonces la ecuación (2.6) es válida: la calidad promedio de la tierra no dependerá de la cantidad de fertilizante. Pero, si a las parcelas de mejor calidad se les aplica más fertilizante, entonces el valor esperado de u variará de acuerdo con el nivel del fertilizante y la ecuación (2.6) no es válida.

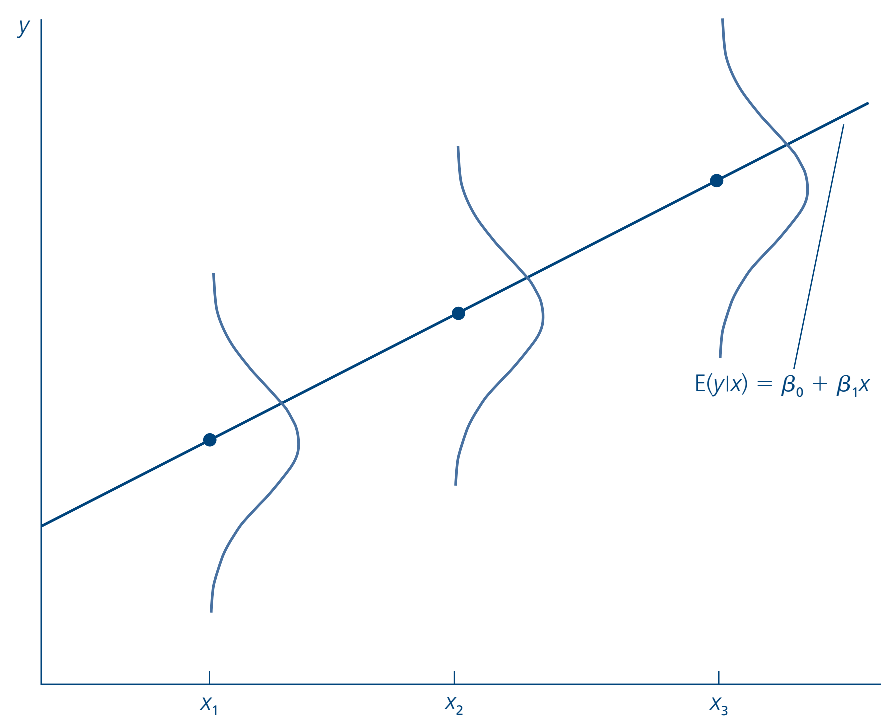
El supuesto de media condicional cero proporciona otra interpretación de β_1 que suele ser útil. Tomando el valor esperado de (2.1) condicionado a x usando $E(u|x) = 0$ se tiene

$$E(y|x) = \beta_0 + \beta_1 x. \tag{2.8}$$

La ecuación (2.8) muestra que la **función de regresión poblacional** (**FRP**), $E(y|x)$, es una función lineal de x. La linealidad significa que por cada aumento de una unidad en x el *valor esperado* de y se modifica en la cantidad β_1. Dado cualquier valor de x, la distribución de y está centrada en $E(y|x)$, como se ilustra en la figura 2.1.

FIGURA 2.1

E(*y*|*x*) como función lineal de *x*.



Es importante entender que la ecuación (2.8) dice cómo varía el valor *promedio* de y de acuerdo con la variación de x; esta ecuación no dice que y sea igual a $\beta_0 + \beta_1 x$ para cada una de las unidades de la población. Suponga, por ejemplo, que x sea el promedio obtenido en el bachillerato, y y sea el promedio obtenido en la universidad y que, además, se sepa que $E(promUniv|promBach) = 1.5 + 0.5\ promBach$. [Claro que, en la práctica, nunca se conocen ni el intercepto ni la pendiente poblacionales, pero para entender la ecuación (2.8) resulta útil suponer, por un momento, que se conocen.] Esta ecuación sobre calificaciones proporciona el *promedio* de las calificaciones de universidad de entre todos los estudiantes que tienen una determinada calificación de bachillerato. De esta manera, suponga que $Prombach = 3.6$. Entonces, el promedio de *Promuniv* de todos los que terminan el bachillerato y asisten a la universidad y que en el bachillerato tuvieron $Prombach = 3.6$ es $1.5 + 0.5(3.6) = 3.3$. *No* se está diciendo que todos los estudiantes que tengan $Prombach = 3.6$ tendrán 3.3 como promedio en la universidad; es claro que esto es falso. La FRP da una relación entre el promedio de y y diferentes valores de x. Algunos de los estudiantes que tengan $Prombach = 3.6$ obtendrán en la universidad un promedio de calificaciones mayor a 3.3 y otros obtendrán promedios más bajos. Que el verdadero *Promuniv* sea mayor o menor a 3.3 depende de los factores no observables en u, y éstos varían entre los estudiantes aun entre los que pertenecen a la porción de la población con $Prombach = 3.6$.

Dado el supuesto de media condicional cero $E(u|x) = 0$, es útil ver la ecuación (2.1) como una que divide a y en dos componentes. A la parte $\beta_0 + \beta_1 x$, que representa $E(y|x)$, se le llama *parte sistemática* de y, es decir, es la parte de y explicada por x y a u se le llama la *parte no sistemática*, o la parte de y que no es explicada por x. En el capítulo 3, en donde se introducirá más de una variable explicativa, se analizará cómo determinar qué tan grande es la parte sistemática con relación a la parte no sistemática.

En la sección siguiente, se usarán los supuestos (2.5) y (2.6) para obtener estimadores de β_0 y β_1 a partir de una muestra aleatoria de datos dada. El supuesto de media condicional cero también tiene un papel crucial en el análisis estadístico de la sección 2.6.

2.2 Obtención de las estimaciones de mínimos cuadrados ordinarios

Una vez que se han analizado los ingredientes básicos del modelo de regresión simple, se abordará el tema de cómo estimar los parámetros β_0 y β_1 de la ecuación (2.1). Para esto se necesita tomar una muestra de la población. Sea $\{(x_i, y_i): i = 1, \ldots, n\}$ una muestra aleatoria de tamaño n tomada de la población. Como estos datos provienen de (2.1), para toda i puede escribirse

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2.9}$$

Aquí u_i es el término del error de la observación i porque contiene todos los demás factores distintos de x_i que afectan a y_i.

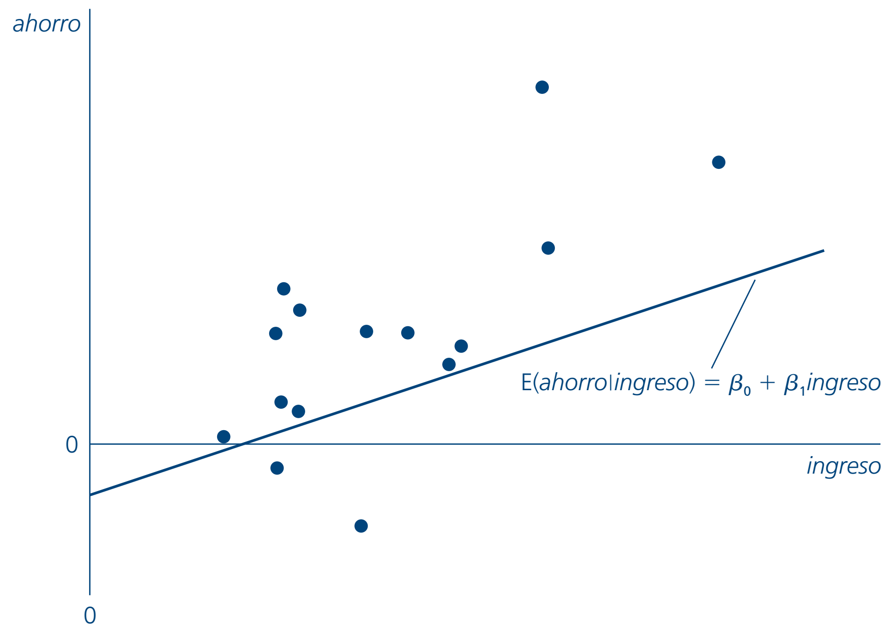
Por ejemplo, x_i puede ser el ingreso anual y y_i el ahorro anual de la familia i durante un determinado año. Si se recolectaron los datos de 15 familias, entonces $n = 15$. En la figura 2.2 se presenta el diagrama de dispersión de estos datos, así como la (necesariamente ficticia) función de regresión poblacional.

Hay que decidir cómo utilizar estos datos para obtener estimaciones para el intercepto y para la pendiente de la función de regresión poblacional de ahorro sobre ingreso.

Existen varias maneras de motivar el siguiente procedimiento de estimación. Aquí se usarán la ecuación (2.5) y una importante consecuencia del supuesto (2.6): en la población, u no está

FIGURA 2.2

Diagrama de dispersión de ingresos y ahorros de 15 familias y función de regresión poblacional $E(ahorros|ingreso) = \beta_0 + \beta_1 ingreso$.



correlacionada con x. Por tanto, se tiene que el valor esperado de u es cero y que la *covarianza* entre x y u es cero:

$$E(u) = 0 \tag{2.10}$$

y

$$\text{Cov}(x,u) = E(xu) = 0, \tag{2.11}$$

donde la primera igualdad en (2.11) sigue de (2.10). (Ver definición y propiedades de la covarianza en la sección B.4.) Las ecuaciones (2.10) y (2.11) pueden expresarse en términos de las variables observables x y y, y de los parámetros desconocidos β_0 y β_1, como sigue:

$$E(y - \beta_0 - \beta_1 x) = 0 \tag{2.12}$$

y

$$E[x(y - \beta_0 - \beta_1 x)] = 0, \tag{2.13}$$

respectivamente. Las ecuaciones (2.12) y (2.13) implican dos restricciones a la distribución de probabilidad conjunta de (x,y) en la población. Como hay que estimar dos parámetros desconocidos, se espera que las ecuaciones (2.12) y (2.13) puedan servir para obtener buenos estimadores de β_0 y β_1. En efecto, estas ecuaciones pueden servir para la estimación de estos parámetros.

Dada una muestra de datos, se eligen estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ que resuelvan las contrapartes *muestrales* de las ecuaciones (2.12) y (2.13):

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2.14}$$

y

$$n^{-1}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2.15}$$

Este es un ejemplo de empleo del *método de momentos* para la estimación. (Vea en la sección C.4 un análisis de los diferentes métodos de estimación.) De estas ecuaciones se pueden obtener soluciones para $\hat{\beta}_0$ y $\hat{\beta}_1$.

Aprovechando las propiedades básicas de la sumatoria, presentadas en el apéndice A, la ecuación (2.14) puede rescribirse como

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}, \tag{2.16}$$

donde $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ es el promedio muestral de las y_i y lo mismo ocurre con $\bar{x}$. Esta ecuación permite escribir $\hat{\beta}_0$ en términos de $\hat{\beta}_1$, $\bar{y}$ y $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{2.17}$$

Por tanto, una vez que se tiene la estimación de la pendiente $\hat{\beta}_1$, es fácil obtener la estimación del intercepto $\hat{\beta}_0$, dadas $\bar{y}$ y $\bar{x}$.

Eliminando el n^{-1} de la ecuación (2.15) (ya que esto no afecta a la solución) y empleando (2.17) para sustituir $\hat{\beta}_0$ en la ecuación (2.15) se obtiene

$$\sum_{i=1}^{n} x_i\,[y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i] = 0,$$

de donde, reordenando, se obtiene

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

De acuerdo con las propiedades básicas de la sumatoria [véase (A.7) y (A.8)],

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ y } \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Por tanto, siempre que

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0, \tag{2.18}$$

la pendiente estimada es

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.19}$$

La ecuación (2.19) es simplemente la covarianza muestral entre x y y dividida entre la varianza muestral de x. (Vea el apéndice C. Dividir tanto el numerador como el denominador entre $n - 1$ no modifica nada.) Esto tiene sentido porque, β_1 es igual a la covarianza poblacional dividida entre la varianza de x cuando $\mathrm{E}(u) = 0$ y $\mathrm{Cov}(x,u) = 0$. Como consecuencia inmediata se tiene que si en la muestra x y y están correlacionadas positivamente, entonces $\hat{\beta}_1$ es positiva; si x y y están correlacionadas negativamente, entonces $\hat{\beta}_1$ es negativa.

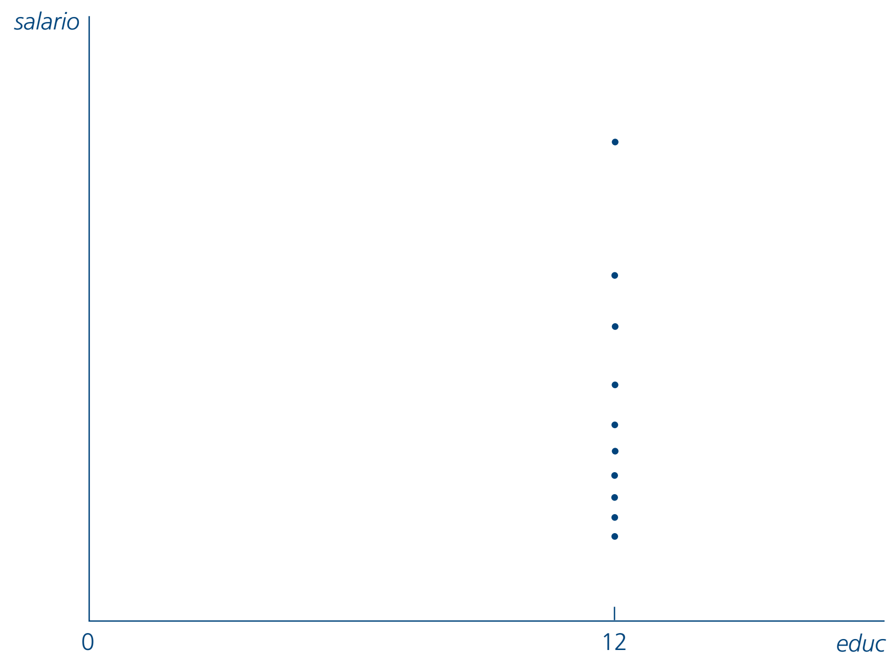
Aunque el método para obtener (2.17) y (2.19) está motivado por (2.6), el único supuesto necesario para calcular estas estimaciones, dada una muestra determinada, es (2.18). Esta ecuación difícilmente puede considerarse un supuesto: la ecuación (2.18) es verdadera siempre que las x_i de la muestra no sean todas iguales a un mismo valor. Si la ecuación (2.18) no es verdadera, no se ha tenido suerte al obtener la muestra de la población o se ha planteado un problema que no es interesante (x no varía en la población). Por ejemplo, si $y = salario$ y $x = educ$, entonces (2.18) no satisfará únicamente si todas las personas de la muestra tienen el mismo nivel de educación (por ejemplo, si todos tienen un título de bachillerato; vea la figura 2.3). Si una sola persona tiene un nivel de educación distinto, entonces la ecuación (2.18) será satisfecha y podrán calcularse las estimaciones.

A las estimaciones dadas por (2.17) y (2.19) se les llama estimaciones de **mínimos cuadrados ordinarios** (**MCO**) de β_0 y β_1. Para justificar este nombre, para todo $\hat{\beta}_0$ y $\hat{\beta}_1$ se define el **valor ajustado** para y cuando $x = x_i$ como

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2.20}$$

FIGURA 2.3

Diagrama de dispersión de salario contra educación cuando $educ_i = 12$ para toda i.



Este es el valor que se predice para y cuando $x = x_i$ con el intercepto y la pendiente dadas. Para cada observación de la muestra hay un valor ajustado. El **residual** de la observación i es la diferencia entre el verdadero valor y_i y su valor ajustado:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{2.21}$$

También hay n residuales. [Los residuales *no* son lo mismo que los errores en la ecuación (2.9), un punto al que se volverá en la sección 2.5.] En la figura 2.4 se muestran los valores ajustados y los residuales.

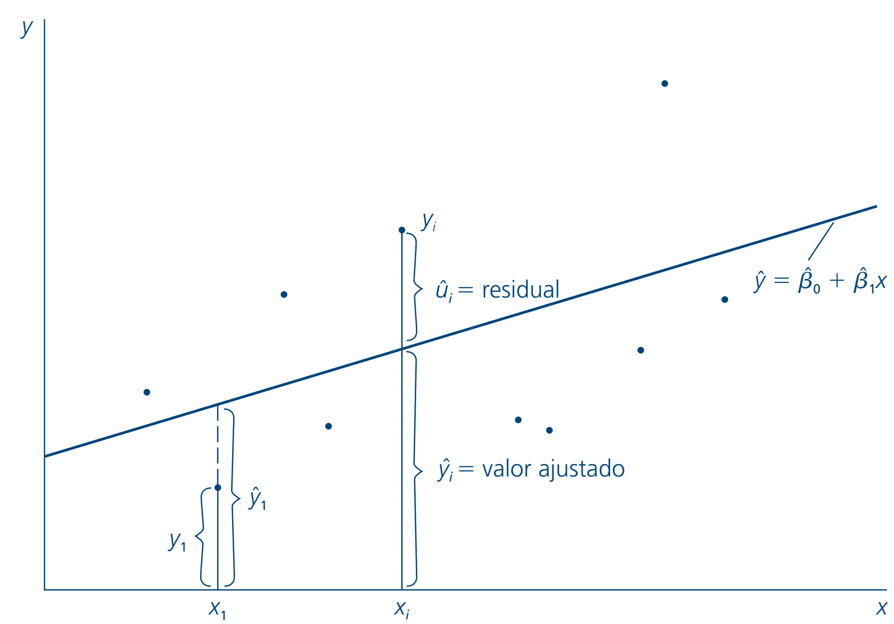
Ahora, suponga que $\hat{\beta}_0$ y $\hat{\beta}_1$ se eligen de manera que la **suma de residuales cuadrados**,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \tag{2.22}$$

sea tan pequeña como sea posible. En el apéndice de este capítulo se muestra que las condiciones necesarias para que $(\hat{\beta}_0, \hat{\beta}_1)$ minimicen (2.22) están dadas exactamente por las ecuaciones (2.14) y (2.15), sin n^{-1}. A las ecuaciones (2.14) y (2.15) se les suele llamar **condiciones de primer orden** para las estimaciones de MCO, término que proviene de la optimización empleando el cálculo (vea el apéndice A). Con base en los cálculos anteriores, se sabe que las soluciones de las condiciones de primer orden de MCO son las dadas por las ecuaciones (2.17) y (2.19). El nombre "mínimos cuadrados ordinarios" proviene del hecho de que estas estimaciones minimizan la suma de los residuales cuadrados.

FIGURA 2.4

Valores ajustados y residuales.



Cuando se ve que el método de mínimos cuadrados ordinarios minimiza la suma de los residuales cuadrados, es natural preguntar: ¿por qué no se minimiza alguna otra función de los residuales, por ejemplo, el valor absoluto de los residuales? En realidad, como se verá en la sección 9.4, minimizar la suma de los valores absolutos de los residuales algunas veces es muy útil; pero esto tiene algunos inconvenientes. Primero, no se pueden obtener fórmulas para los estimadores resultantes; dada una base de datos, las estimaciones deben obtenerse mediante rutinas de optimización numérica. Como consecuencia, la teoría estadística para estimadores que minimizan la suma de los residuales absolutos es muy complicada. Minimizar otras funciones de los residuales, por ejemplo, la suma de los residuales cada uno elevado a la cuarta potencia, tiene inconvenientes semejantes. (Nunca se pensaría en estimaciones para minimizar, por ejemplo, la suma de los residuales mismos, ya que los residuales grandes en magnitud pero de signos opuestos tienden a anularse.) Con MCO se podrá obtener insesgamiento, consistencia y otras propiedades estadísticas de una manera relativamente sencilla. Además, como sugiere la motivación en las ecuaciones (2.13) y (2.14) y como se verá en la sección 2.5, el método de MCO es adecuado para estimar los parámetros que aparecen en la función de la media condicional (2.8).

Una vez que se han determinado las estimaciones por MCO del intercepto y de la pendiente, se obtiene la **línea de regresión** de **MCO**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2.23}$$

donde se entiende que $\hat{\beta}_0$ y $\hat{\beta}_1$ han sido obtenidas empleando las ecuaciones (2.17) y (2.19). La notación $\hat{y}$, que se lee "*y* gorro" indica que los valores predichos por la ecuación (2.23) son estimaciones. El intercepto, $\hat{\beta}_0$, es el valor predicho de y cuando $x = 0$, aunque en algunos casos no tiene sentido hacer $x = 0$. En estas situaciones, $\hat{\beta}_0$ no tiene gran interés por sí misma. Cuando se emplea la ecuación (2.23) para calcular valores predichos de y para diversos valores de x, al hacer los cálculos hay que tomar en cuenta el intercepto. A la ecuación (2.23) también se le llama **función de regresión muestral** (**FRM**) debido a que es la versión estimada de la función de regresión poblacional $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$. Es importante recordar que la FRP es algo fijo, pero desconocido, de la población. Dado que la FRM se obtiene a partir de una muestra de datos, con una nueva muestra se obtendrán una pendiente e intercepto diferentes para la ecuación (2.23).

En la mayoría de los casos, la pendiente estimada, que se puede expresar como

$$\hat{\beta}_1 = \Delta\hat{y}/\Delta x, \tag{2.24}$$

es de primordial interés, pues indica la cantidad en la que cambia $\hat{y}$ cuando x se incrementa en una unidad. De manera equivalente,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x, \tag{2.25}$$

de manera que dado cualquier cambio en x (ya sea positivo o negativo), se puede calcular el cambio en y.

Ahora se presentarán varios ejemplos de regresión simple obtenidos empleando datos reales. En otras palabras, se encontrarán las estimaciones de la pendiente y del intercepto empleando las ecuaciones (2.17) y (2.19). Como en estos ejemplos se usan muchas observaciones, los cálculos fueron hechos empleando un software para econometría. Por ahora, hay que tener cuidado de no pretender obtener demasiada información de estas regresiones; éstas no necesariamente implican una relación causal. Hasta ahora no se ha dicho nada acerca de las propiedades estadísticas de los estimadores de MCO. En la sección 2.5 se consideran las propiedades estadísticas después de imponer explícitamente algunos supuestos a la ecuación del modelo poblacional (2.1).

Ejemplo 2.3

[Sueldo de los directores generales (CEO) y rendimiento sobre el capital]

En la población de los directores generales, sea y el sueldo anual (*salary*) en miles de dólares. De manera que $y = 856.3$ corresponde a un sueldo anual de \$856,300 y $y = 1,452.6$ corresponde a un sueldo de \$1,452,600. Sea x el promedio, en los últimos tres años, del rendimiento sobre el capital (*roe*) en las empresas de los CEO. (El rendimiento sobre el capital se define en términos de utilidad neta como porcentaje de acciones comunes.) Por ejemplo, si $roe = 10$, el rendimiento promedio sobre el capital es 10 por ciento.

Para estudiar la relación entre esta medida del desempeño de una empresa y el pago que reciben los CEO, se postula el modelo

$$salary = \beta_0 + \beta_1 roe + u.$$

El parámetro de la pendiente β_1 mide la variación del sueldo anual, en miles de dólares, que corresponde a un incremento de un punto porcentual en el rendimiento sobre capital. Debido a que un *roe* más alto es bueno para la empresa se considera que $\beta_1 > 0$.

La base de datos CEOSAL1.RAW contiene información, correspondiente al año 1990, sobre 209 CEO; estos datos fueron obtenidos de *Business Week* (5/6/91). En esta muestra, el sueldo anual promedio es \$1'281,120, y los sueldos menor y mayor son \$223,000 y \$14,822,000, respectivamente. La media del rendimiento sobre capital en 1988, 1989 y 1990 es 17.18% y los valores menor y mayor son 0.5 y 56.3%, respectivamente.

Usando los datos de CEOSAL1.RAW, la línea de regresión de MCO que relaciona *salary* y *roe* es

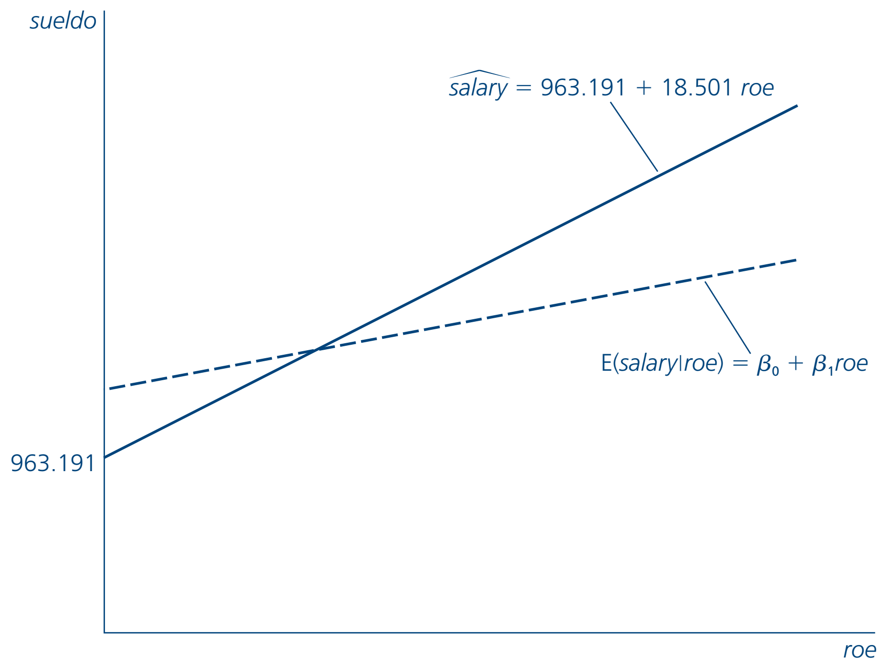
$$\widehat{salary} = 963.191 + 18.501\, roe, \tag{2.26}$$

en donde el intercepto y la pendiente estimadas se han redondeado a tres cifras decimales; "*salary* gorro" se usa para indicar que ésta es una ecuación estimada. ¿Cómo se interpreta esta ecuación? Primero, si el rendimiento sobre el capital es cero, $roe = 0$, entonces el sueldo que se predice corresponde al intercepto, 963.191, es decir, \$963,191, dado que *salary* se mide en miles. Luego, el cambio que se predice para el sueldo en función del cambio en el *roe* se expresa como: $\Delta\widehat{salary} = 18.501\,(\Delta roe)$. Esto significa que cuando el rendimiento sobre capital aumente en un punto porcentual, $\Delta roe = 1$, se predice que el *sueldo* variará aproximadamente 18.5, es decir \$18,500. Como la ecuación (2.26) es lineal, este es el cambio estimado sin importar cuál sea el sueldo inicial.

La ecuación (2.26) también se puede usar para comparar los sueldos que se predicen para diferentes valores del *roe*. Suponga que $roe = 30$. Entonces $\widehat{salary} = 963.191 + 18.501(30) = 1,518,221$, que es un poco más de \$1.5 millones. Sin embargo, esto *no* significa que un determinado CEO, para cuya empresa $roe = 30$ gane \$1,518,221. Hay otros muchos factores que afectan al sueldo. Esta es solamente una predicción a partir de la línea de regresión de MCO (2.26). En la figura 2.5 se grafica la línea estimada, así como la función de regresión poblacional $E(salary|roe)$. La FRP no podrá conocerse nunca, de manera que no se puede decir qué tan cerca están la FRM de la FRP. Con otros datos de muestra se obtendrá otra línea de regresión diferente, que podrá estar más o menos cerca de la línea de regresión poblacional.

FIGURA 2.5

Línea de regresión de MCO $\widehat{salary} = 963.191 + 18.501\, roe$ y la función de regresión poblacional (desconocida).



Ejemplo 2.4

[Salario y educación]

Para la población de personas en la fuerza de trabajo en 1976, sea $y = wage$ (salario), donde *wage* se mide en dólares por hora. Entonces, si para una determinada persona $wage = 6.75$, su salario por hora es $6.75. Sea $x = educ$ años de educación; por ejemplo, $educ = 12$ corresponde a haber terminado el bachillerato. Como el salario promedio de la muestra es $5.90, el índice de precios al consumidor indica que esta cantidad es equivalente a $19.06 dólares de 2003.

Usando los datos de WAGE1.RAW en los que $n = 526$ individuos, se obtiene la siguiente línea de regresión de MCO (o función de regresión muestral):

$$\widehat{wage} = -0.90 + 0.54\, educ. \tag{2.27}$$

Esta ecuación debe interpretarse con cuidado. El intercepto de -0.90 significa literalmente que para una persona sin ninguna educación el sueldo por hora que se predice es -90¢ por hora. Lo que, desde luego, no tiene ningún sentido. Resulta que, de las 526 personas que forman la muestra sólo 18 cuentan con menos de ocho años de educación. Por tanto, no es de sorprender que la línea de regresión no resulte adecuada para niveles de educación muy bajos. Para una persona que tiene ocho años de educación el salario que se predice es $\widehat{wage} = -0.90 + 0.54(8) = 3.42$, es decir $3.42 por hora (en dólares de 1976).

Pregunta 2.2

Con base en la ecuación (2.27), si $educ = 8$, el salario estimado es 3.42 en dólares de 1976. ¿A cuánto corresponde este valor en dólares de 2003? (*Indicación:* en el ejemplo 2.4 se da suficiente información para responder esta pregunta.)

La pendiente estimada de la ecuación (2.27) implica que un año adicional de educación hace que el salario por hora aumente 54¢ por hora. Por tanto, cuatro años más de educación hacen que el sueldo que se predice aumente $4(0.54) = 2.16$, es decir $2.16 por hora. Este efecto es bastante grande. Debido al carácter lineal de la ecuación (2.27), cada año adicional de educación hace que el salario aumente en una misma cantidad, independientemente del nivel inicial de educación. En la sección 2.4 se analizan algunos métodos que permiten considerar efectos marginales no constantes de las variables explicativas.

Ejemplo 2.5

[Resultados de una votación y gastos de campaña]

El archivo VOTE1.RAW contiene los resultados de las elecciones y los gastos de campaña correspondientes a 173 contiendas bipartidistas por la Cámara de Representantes de Estados Unidos, en 1988. En cada contienda hay dos candidatos, A y B. Sea *voteA* el porcentaje de votos obtenido por el candidato A y *shareA* el porcentaje del total de los gastos de campaña atribuido al candidato A. Además de *shareA* hay otros muchos factores que afectan los resultados de una elección (como la calidad de los candidatos y, probablemente, las cantidades en dólares gastadas por A y B). De cualquier manera, es posible estimar un modelo de regresión simple para determinar si gastar más con relación al rival implica obtener un porcentaje mayor de votos.

La ecuación estimada empleando las 173 observaciones es

$$\widehat{voteA} = 26.81 + 0.464\, shareA. \tag{2.28}$$

Esto significa que si la cantidad gastada por el candidato A aumenta en un punto porcentual, él obtendrá medio punto porcentual (0.464) más del total de los votos. No es claro si se trata o no de un efecto causal, pero no es inverosímil. Si $shareA = 50$, se predice que *voteA* será aproximadamente 50, es decir, la mitad de los votos.

En algunos casos el análisis de regresión no se usa para determinar causalidad sino sólo para ver si dos variables están relacionadas negativa o positivamente, de manera muy parecida a como se usa el análisis de correlación común. Ejemplo de esto es el ejercicio para computadora C2.3, donde se pide que usando los datos de Biddle y Hamermesh (1990) sobre cantidad de horas de sueño y de trabajo se investigue la relación inversa entre estos dos factores.

Pregunta 2.3

En el ejemplo 2.5, ¿cuál es la cantidad de votos que se predice para el candidato A si $shareA = 60$ (que significa 60 por ciento)? ¿Parece razonable esta respuesta?

Nota sobre la terminología

En la mayoría de los casos, la estimación de una relación mediante MCO se indicará escribiendo una ecuación como las ecuaciones (2.26), (2.27) o (2.28). Algunas veces, para abreviar, se indica que se ha hecho una regresión por MCO sin escribir la ecuación. Con frecuencia se indicará que la ecuación (2.23) se ha obtenido mediante MCO diciendo que *se ha hecho la regresión de*

$$y \text{ sobre } x, \quad \mathbf{2.29}$$

o simplemente que se *regresiona* y sobre x. Las posiciones de y y de x en (2.29) indican cuál es la variable dependiente y cuál la independiente: siempre se regresiona la variable dependiente sobre la independiente. En las aplicaciones específicas y y x se sustituyen por sus nombres. Así, para obtener la ecuación (2.26), se regresiona *salary* sobre *roe*, y al obtener la ecuación (2.28), se regresiona *voteA* sobre *shareA*.

Cuando se usa la terminología dada en (2.29), se quiere dar a entender que se van a estimar el intercepto, $\hat{\beta}_0$, y la pendiente, $\hat{\beta}_1$. Este es el caso en la gran mayoría de las aplicaciones. Algunas veces se desea estimar la relación entre y y x *suponiendo* que el intercepto es cero (de manera que $x = 0$ implica $\hat{y} = 0$); este caso se trata brevemente en la sección 2.6. A menos que se diga otra cosa, aquí siempre se estimarán el intercepto y la pendiente.

2.3 Propiedades de MCO en cualquier muestra de datos

En la sección anterior, se dedujeron las fórmulas para las estimaciones, por MCO, del intercepto y de la pendiente. En esta sección, se verán algunas otras propiedades algebraicas de la línea de regresión ajustada de MCO. Hay que recordar que estas propiedades, por construcción, son válidas para *cualquier* muestra de datos. La tarea más difícil —considerar las propiedades de MCO en todas las posibles muestras aleatorias de datos— se posponen hasta la sección 2.5.

Varias de las propiedades algebraicas que se van a deducir pueden parecer muy simples. Sin embargo, entenderlas ayudará a comprender lo que pasa con las estimaciones de MCO y con los estadísticos con ellos relacionados al manipular los datos de ciertas maneras, por ejemplo, cuando se modifican las unidades de medición de las variables dependiente o independiente.

Valores ajustados y residuales

Se supone que las estimaciones del intercepto y de la pendiente, $\hat{\beta}_0$ y $\hat{\beta}_1$, han sido obtenidas para los datos muestrales dados. Una vez que se tienen $\hat{\beta}_0$ y $\hat{\beta}_1$, se puede obtener el valor ajustado $\hat{y}_i$ correspondiente a cada observación. [Esto se indica en la ecuación (2.20).] Por definición, todos los valores ajustados $\hat{y}_i$ se encuentran sobre la línea de regresión de MCO. El residual de MCO correspondiente a la observación i, $\hat{u}_i$, es la diferencia entre y_i y su valor ajustado, como se indica en la ecuación (2.21). Si $\hat{u}_i$ es positivo, la línea predice un valor inferior al de y_i; si $\hat{u}_i$ es negativo, la línea predice un valor superior al de y_i. Lo ideal para la observación i es cuando $\hat{u}_i = 0$, pero en la mayoría de los casos, *todos* los residuales son distintos de cero. En otras palabras, no es necesario que ninguno de los puntos de los datos se encuentre exactamente sobre la línea de MCO.

Ejemplo 2.6

[Sueldo de los CEO y rendimiento sobre el capital]

La tabla 2.2 contiene una lista de las primeras 15 observaciones de la base de datos de los CEO, así como los valores ajustados, a los que se les llama *salarygorro* (*sueldogorro*), y los residuales, a los que se les llama *ugorro*.

TABLA 2.2

Valores ajustados y residuales de los primeros 15 CEO

obsno	*roe*	*salary (sueldo)*	*salarygorro*	*ugorro*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

Los primeros cuatro CEO tienen un sueldo menor que el que se predice empleando la línea de regresión de MCO (2.26); en otras palabras, dado únicamente el *roe* de las empresas, estos CEO ganan menos de lo que se predice. Como se puede ver, por el *ugorro* positivo, el quinto CEO gana más de lo que se predice de acuerdo con la línea de regresión de MCO.

Propiedades algebraicas de los estadísticos de MCO

Las estimaciones de MCO y sus correspondientes estadísticos tienen varias propiedades útiles. A continuación se verán las tres más importantes.

(1) La suma, y por tanto el promedio muestral de los residuales de MCO, es cero. Matemáticamente,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \tag{2.30}$$

Esta propiedad no necesita ser probada; es consecuencia inmediata de la condición de primer orden (2.14) de MCO, si se recuerda que los residuales están definidos por $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. En otras palabras, las estimaciones de MCO $\hat{\beta}_0$ y $\hat{\beta}_1$ se *eligen* de manera que la suma de los residuales sea cero (para cualquier base de datos). Esto no dice nada acerca del residual de una determinada observación *i*.

(2) La covarianza muestral entre los regresores y los residuales de MCO es cero. Esto es consecuencia de la condición de primer orden (2.15), que en términos de los residuales puede expresarse como

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \tag{2.31}$$

El promedio muestral de los residuales de MCO es cero, por lo que el lado izquierdo de la ecuación (2.31) es proporcional a la covarianza entre las x_i y los $\hat{u}_i$.

(3) El punto $(\bar{x},\bar{y})$ se encuentra siempre sobre la línea de regresión de MCO. En otras palabras, si en la ecuación (2.23) se sustituye $\bar{x}$ por x, el valor predicho es $\bar{y}$. Esto es exactamente lo que dice la ecuación (2.16).

Ejemplo 2.7

[Salario y educación]

En los datos contenidos en WAGE1.RAW, el promedio del salario muestral por hora es 5.90, redondeado a dos cifras decimales, y la educación promedio es 12.56. Si en la recta de regresión por MCO (2.27) se sustituye $educ = 12.56$ se obtiene $\widehat{wage} = -0.90 + 0.54(12.56) = 5.8824$, que redondeando a una cifra decimal es igual a 5.9. Estos números no coinciden exactamente debido a que se han redondeado el salario promedio y la educación, así como las estimaciones del intercepto y de la pendiente. No redondeando al inicio ninguno de estos valores se obtiene una mejor coincidencia, aunque de poca utilidad.

Escribiendo cada y_i como su valor ajustado, más su residual, se obtiene otra manera de interpretar la regresión de MCO. Para cada *i* se tiene

$$y_i = \hat{y}_i + \hat{u}_i. \tag{2.32}$$

De acuerdo con la propiedad (1), el promedio de los residuales es cero; lo que es equivalente a que el promedio muestral de los valores ajustados, $\hat{y}_i$, es igual al promedio muestral de las y_i, es decir $\bar{\hat{y}} = \bar{y}$. Además, con base en las propiedades (1) y (2) se puede mostrar que la covarianza muestral entre $\hat{y}_i$ y $\hat{u}_i$ es cero. Por tanto, se puede considerar que el método de MCO descompone cada y_i en dos partes, un valor ajustado y un residual. Los valores ajustados y los residuales no están correlacionados en la muestra.

Se definen la **suma total de cuadrados** (**STC**), la **suma explicada de cuadrados** (**SEC**) y la **suma residual de cuadrados** (**SRC**) (conocida también como suma de residuales cuadrados), como sigue:

$$\text{STC} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{2.33}$$

$$\text{SEC} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \qquad \textbf{2.34}$$

$$\text{SRC} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \qquad \textbf{2.35}$$

La STC es una medida de la variación muestral total en las y_i; es decir, mide qué tan dispersos están las y_i en la muestra. Si se divide la STC entre $n - 1$, se obtiene la varianza muestral de y, que se analiza en el apéndice C. De manera similar, la SEC mide la variación muestral de las $\hat{y}_i$ (donde se usa el hecho de que $\bar{\hat{y}} = \bar{y}$), y la SRC mide la variación muestral de los $\hat{u}_i$. La variación total de y puede expresarse como la suma de la variación explicada más la variación no explicada SRC. Por tanto,

$$\text{STC} = \text{SEC} + \text{SRC}. \qquad \textbf{2.36}$$

Probar (2.36) no es difícil, pero se requiere el uso de las propiedades de la sumatoria, vistas en el apéndice A. Al escribir

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n}[\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
&= \text{STC} + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \text{SEC}.
\end{aligned}$$

Ahora, (2.36) es verdadera si se muestra que

$$\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) = 0. \qquad \textbf{2.37}$$

Pero ya se ha dicho que la covarianza muestral entre los residuales y los valores ajustados es cero y esta covarianza es precisamente (2.37) dividida entre $n-1$. Por tanto, (2.36) queda probada.

Sólo una advertencia acerca de STC, SEC y SRC. No hay un acuerdo general para los nombres y las siglas que se emplean para las tres cantidades definidas en las ecuaciones (2.33), (2.34) y (2.35). Para la suma total de cuadrados se usa STC o SCT, de manera que hay un poco de confusión. Desafortunadamente, a la suma explicada de cuadrados suele llamársele también "suma de cuadrados de la regresión". Si se emplea su abreviación natural para este término, con facilidad puede confundirse con el término "suma residual de cuadrados". En algunos paquetes para regresión a la suma explicada de cuadrados se le llama "suma de cuadrados del modelo".

Para complicar las cosas, a la suma residual de cuadrados se le suele llamar "suma de cuadrados de los errores". Esto es en especial desafortunado ya que, como se verá en la sección 2.5, los errores y los residuales son cantidades diferentes. Por tanto, aquí a (2.35) se le llamará la suma residual de cuadrados o la suma de residuales cuadrados. Es preferible emplear la abreviación SRC para denotar la suma de residuales cuadrados, debido a que ésta es más común en los paquetes para econometría (SSR en los paquetes en inglés).

Bondad de ajuste

Hasta ahora, no se tiene una manera de medir qué tan bien la variable explicativa o independiente, x, explica a la variable dependiente, y. Suele ser útil calcular un número que resuma qué tan bien se ajusta la línea de regresión de MCO a los datos. En el siguiente análisis hay que recordar que se supone que se estiman la pendiente y el intercepto.

Suponiendo que la suma total de cuadrados, STC, no sea igual a cero —lo que siempre es así, salvo en el muy remoto caso de que todas las y_i tengan el mismo valor— puede dividirse (2.36) entre la STC para obtener $1 = \text{SEC/STC} + \text{SRC/STC}$. La ***R*-cuadrada** de la regresión, también llamada **coeficiente de determinación**, se define como

$$R^2 \equiv \text{SEC/STC} = 1 - \text{SRC/STC}. \tag{2.38}$$

R^2 es el cociente de la variación explicada entre la variación total; por tanto, se interpreta como la *proporción de la variación muestral de y que es explicada por x*. La segunda igualdad en la ecuación (2.38) proporciona otra manera de calcular R^2.

De acuerdo con la ecuación (2.36), el valor de R^2 está siempre entre cero y uno, ya que SEC no puede ser mayor a STC. Al interpretar R^2, se acostumbra multiplicarla por 100 para transformarla en un porcentaje: $100 \cdot R^2$ es el *porcentaje de la variación muestral de y que es explicada por x*.

Si todos los puntos de los datos se encuentran sobre una misma línea, los MCO proporcionan un ajuste perfecto a los datos. En este caso, $R^2 = 1$. Si el valor de R^2 es casi igual a cero, esto indica un ajuste pobre de la línea de MCO: muy poco de la variación de las y_i es captado por la variación de las $\hat{y}_i$ (que se encuentran, todas, sobre la línea de regresión de MCO). Se puede mostrar que R^2 es igual al *cuadrado* del coeficiente de correlación muestral entre y_i y $\hat{y}_i$. Es de aquí de donde proviene el término "*R*-cuadrada". (La letra R ha sido tradicionalmente empleada para denotar la estimación del coeficiente de correlación poblacional y su uso ha pasado al análisis de regresión.)

Ejemplo 2.8

[Sueldo de los CEO y rendimiento sobre el capital]

En la regresión del sueldo de los CEO, se obtiene lo siguiente:

$$\widehat{salary} = 963.191 + 18.501\ roe \tag{2.39}$$

$$n = 209,\ R^2 = 0.0132.$$

Para mayor claridad se dan de nuevo la línea de regresión de MCO y la cantidad de observaciones. Empleando la *R*-cuadrada (redondeada a cuatro cifras decimales) dada para esta ecuación, se puede ver en realidad cuánto de la variación en el salario es explicado por el rendimiento sobre el capital. La respuesta es: no mucho. El rendimiento sobre el capital de las empresas sólo explica, aproximadamente, 1.3 por ciento de la variación en los salarios de esta muestra de 209 CEO. Esto significa que ¡98.7 por ciento de la variación en los salarios de estos CEO no queda explicada! Esta falta de poder explicativo no debe sorprender, pues hay otras múltiples características, tanto de la empresa como de cada CEO, que influyen en el sueldo; estos factores quedan necesariamente incluidos en los errores en un análisis de regresión simple.

En las ciencias sociales, R-cuadradas bajas en las ecuaciones de regresión no son poco comunes en especial en análisis con datos de corte transversal. Esto se discutirá con más amplitud en el análisis de regresión múltiple, pero ahora vale la pena hacer hincapié en que una R-cuadrada muy baja no necesariamente significa que la ecuación de regresión de MCO sea inútil. Puede que, a pesar de todo, la ecuación (2.39) sea una buena estimación de la relación *ceteris paribus* entre *salary* y *roe*; que esto sea o no verdad, *no* depende de manera directa del tamaño de R-cuadrada. Los estudiantes que empiezan a aprender econometría tienden a darle mucha importancia al valor de R-cuadrada al evaluar una ecuación de regresión. Por ahora, hay que estar conscientes de que usar R-cuadrada como principal medida del éxito de un análisis econométrico puede acarrear problemas.

Algunas veces, la variable independiente expresa una parte sustancial de la variable dependiente.

Ejemplo 2.9

[Resultados de una votación y gastos de campaña]

En la ecuación (2.28) sobre los resultados de una votación, $R^2 = 0.856$. Por tanto, en esta muestra, los gastos de campaña explican más de 85% de la variación en los resultados de la votación. Esta es una porción considerable.

2.4 Unidades de medición y forma funcional

Dos temas importantes en economía aplicada son 1) entender cómo el cambiar las unidades de medición de la variable dependiente o de la variable independiente afecta las estimaciones de MCO y 2) saber cómo incorporar las formas funcionales comúnmente usadas en economía en el análisis de regresión. En el apéndice A se presenta un repaso de las matemáticas necesarias para una buena comprensión de las formas funcionales.

Efectos de los cambios de unidades de medición sobre los estadísticos obtenidos de MCO

En el ejemplo 2.3, el salario anual se midió en miles de dólares y el rendimiento sobre capital, en forma porcentual (no decimal). Es muy importante saber cómo se miden *salary* y *roe* en este ejemplo para entender las estimaciones obtenidas con la ecuación (2.39).

También hay que saber que las estimaciones de MCO cambian de maneras totalmente esperadas cuando cambian las unidades de medición de la variable dependiente o de la independiente. Suponga que en el ejemplo 2.3, en lugar de medir el salario en miles de dólares se mide en dólares. Sea *salardol* el salario en dólares (*salardol* = 845,761 se interpretará como $845,761). Es claro que entre *salardol* y el sueldo medido en miles de dólares existe una relación sencilla:

salardol = 1,000·*salary*. No es necesario hacer la regresión de *salardol* sobre *roe* para saber que la ecuación estimada será:

$$\widehat{salardol} = 963{,}191 + 18{,}501\ roe. \qquad \mathbf{2.40}$$

El intercepto y la pendiente de la ecuación (2.40) se obtienen multiplicando por 1,000 el intercepto y la pendiente de la ecuación (2.39). De esta manera, las ecuaciones (2.39) y (2.40) tienen una *misma* interpretación. En la ecuación (2.40), si $roe = 0$, entonces $\widehat{salardol} = 963{,}191$, de manera que el sueldo que se predice es $963,191 [que es el mismo valor que se obtuvo con la ecuación (2.39)]. Además, por cada aumento de uno en *roe*, el aumento predicho en el sueldo es $18,501; otra vez, lo que se concluyó en el análisis anterior de la ecuación (2.39).

En general, es fácil explicar lo que pasa con las estimaciones de la pendiente y el intercepto cuando cambian las unidades de medición de la variable dependiente. Cuando se multiplica la variable dependiente por una constante *c* —lo que significa multiplicar cada valor de la muestra por *c*—, entonces las estimaciones de MCO del intercepto y de la pendiente también son multiplicadas por *c*. (Esto supone que no ha cambiado nada en la variable independiente.) En el ejemplo de los salarios de los CEO, al pasar de *salary* a *salardol*, $c = 1{,}000$

El ejemplo de los salarios de los CEO también puede usarse para ver lo que ocurre cuando cambian las unidades de medición de la variable independiente. Se define $roedec = roe/100$ que es el equivalente decimal de *roe*; por tanto, $roedec = 0.23$ significa un rendimiento sobre el capital de 23 por ciento. Para concentrar la atención en el cambio de las unidades de medición de la variable independiente, se vuelve a la variable dependiente original, *salary*, medida en miles de dólares. Cuando se hace la regresión de *salary* sobre *roedec*, se obtiene

Pregunta 2.4

Suponga que el salario se mide en cientos de dólares, *salarhun*, en lugar de en miles de dólares. ¿Cuáles serán las estimaciones de MCO del intercepto y de la pendiente, en la regresión de *salarhun* sobre *roe*?

$$\widehat{salary} = 963.191 + 1{,}850.1\ roedec. \qquad \mathbf{2.41}$$

El coeficiente de *roedec* es el coeficiente de *roe* de la ecuación (2.39) multiplicado por 100. Así debe ser. Un cambio de *roe* en un punto porcentual es equivalente a $\Delta roedec = 0.01$. De acuerdo con la ecuación (2.41), si $\Delta roedec = 0.01$, entonces $\Delta\widehat{salary} = 1{,}850.1(0.01) = 18.501$, que es lo mismo que se obtiene usando la ecuación (2.39). Observe que al pasar de (2.39) a (2.41), la variable independiente se dividió entre 100, por lo que la estimación de MCO de la pendiente se multiplicó por 100, conservándose así la interpretación de la ecuación. En general, si la variable independiente se divide o se multiplica por una constante distinta de cero, *c*, entonces el coeficiente de la pendiente de MCO se multiplica por *c* o se divide entre *c*, respectivamente.

En la ecuación (2.41) el intercepto no cambia ya que $roedec = 0$ sigue correspondiendo a cero rendimiento sobre capital. En general, cambiar sólo las unidades de medición de la variable independiente no afecta el intercepto.

En la sección anterior se definió *R*-cuadrada como una medida de la bondad de ajuste de la regresión de MCO. También se puede preguntar qué ocurre con R^2 cuando cambian las unidades de medición, ya sea de la variable independiente o de la variable dependiente. La respuesta se sabe sin necesidad de hacer ningún cálculo algebraico: la bondad de ajuste del modelo no depende de las unidades de medición de las variables. Por ejemplo, la cantidad de variación del sueldo que es explicada por el rendimiento sobre el capital no depende de si el sueldo se mide en dólares o en miles de dólares o de si el rendimiento sobre el capital es un porcentaje o un decimal. Esta idea intuitiva puede verificarse matemáticamente: empleando la definición de R^2 se puede mostrar que, en efecto R^2 no varía ante los cambios de unidades de y o de x.

Incorporación de no linealidades en la regresión simple

Hasta ahora, se ha fijado la atención en relaciones *lineales* entre las variables dependiente e independiente. Como se dijo en el capítulo 1, las relaciones lineales no son suficientemente generales para todas las aplicaciones económicas. Por fortuna, es bastante fácil incorporar muchas no linealidades en el análisis de regresión simple mediante una definición apropiada de las variables dependiente e independiente. Aquí se verán dos posibilidades que surgen con frecuencia en la práctica.

En la literatura de las ciencias sociales, con frecuencia se encuentran ecuaciones de regresión en las que la variable dependiente aparece en forma logarítmica. ¿A qué se debe esto? Recuerde el ejemplo sueldo-educación, en el que se hizo la regresión del salario por hora sobre años de educación. La pendiente estimada fue de 0.54 [vea la ecuación (2.27)], lo que significa que se predice que por cada año más de educación el salario por hora aumentará 54 centavos. Debido a que la ecuación (2.27) es lineal, 54 centavos es el aumento, ya sea por el primer o por el vigésimo año de educación; cosa que no parece razonable.

Una mejor caracterización para el cambio del salario de acuerdo con la educación puede que sea que por cada año más de educación el salario aumente un *porcentaje* constante. Por ejemplo, que un aumento en la educación de cinco a seis años haga que el salario aumente, por ejemplo, 8% (*ceteris paribus*), y que un aumento en la educación de 11 a 12 años, haga que el salario aumente también 8%. Un modelo con el que (aproximadamente) se obtiene un efecto porcentual constante es

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{2.42}$$

donde $\log(\cdot)$ denota el logaritmo *natural*. (Vea en el apéndice A un repaso de los logaritmos.) En particular, si $\Delta u = 0$, entonces

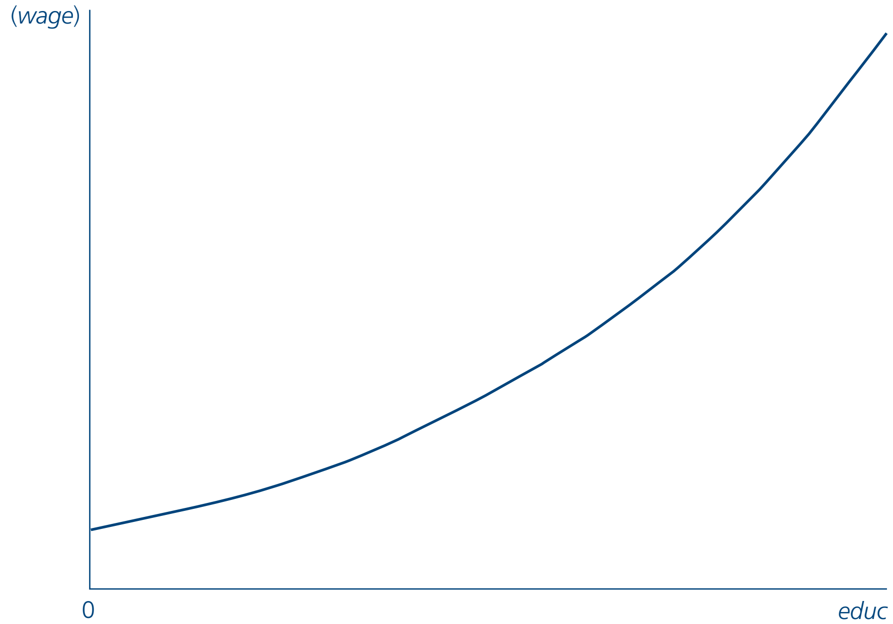
$$\%\Delta wage \approx (100 \cdot \beta_1)\Delta educ. \tag{2.43}$$

Observe que β_1 se multiplica por 100 para obtener el cambio porcentual de *wage* por un año más de educación. Como el cambio porcentual es el mismo por cada año adicional de educación, el cambio de *wage* por un año más de educación *aumenta* a medida que la educación lo hace; en otras palabras, (2.42) implica un rendimiento *creciente* de la educación. Exponenciando (2.42), se obtiene $wage = \exp(\beta_0 + \beta_1 educ + u)$. En la figura 2.6 se grafica esta ecuación, con $u = 0$.

La estimación de un modelo como el de la ecuación (2.42) es sencilla cuando se usa regresión simple. Sólo se define la variable dependiente, y, como $y = \log(wage)$. La variable independiente se representa por $x = educ$. La mecánica de MCO sigue siendo la misma que antes: las estimaciones del intercepto y de la pendiente están dadas por las fórmulas (2.17) y (2.19). En otras palabras, $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen mediante una regresión por MCO de $\log(wage)$ sobre $educ$.

FIGURA 2.6

$wage = \exp(\beta_0 + \beta_1 educ)$, con $\beta_1 > 0$.



Ejemplo 2.10

[Ecuación logarítmica del salario]

Empleando los mismos datos que en el ejemplo 2.4, pero usando log(*wage*) como variable dependiente, se obtiene la siguiente relación:

$$\widehat{\log(wage)} = 0.584 + 0.083\ educ \tag{2.44}$$

$$n = 526,\ R^2 = 0.186.$$

El coeficiente de *educ* tiene una interpretación porcentual multiplicándolo por 100: $\widehat{wage}$ aumenta 8.3% por cada año más de educación. Esto es a lo que los economistas se refieren cuando hablan de “rendimiento de un año más de educación”.

Es importante recordar que la principal razón para emplear el logaritmo de *wage* en la ecuación (2.42) es imponer a la educación un efecto porcentual constante sobre *wage*. Una vez obtenida la ecuación (2.42), el logaritmo natural de *wage* apenas se menciona. En particular, *no* es correcto decir que un año más de educación incrementa log(*wage*) 8.3 por ciento.

En la ecuación (2.42) el intercepto no tiene mucho significado, ya que da el log(*wage*) predicho, cuando *educ* = 0. La *R*-cuadrada muestra que *educ* explica cerca de 18.6% de la variación en log(*wage*) (*no* en *wage*). Por último, la ecuación (2.44) puede que no capte por completo la no linealidad de la relación entre salarios y escolaridad. Si hay “efectos de diploma”, entonces el duodécimo año de educación —terminación del bachillerato— deberá ser mucho más valioso que el undécimo año. En el capítulo 7 se verá cómo considerar esta clase de no linealidades.

Otro uso importante del logaritmo natural es la obtención de un **modelo de elasticidad constante**.

Ejemplo 2.11

[Sueldo de los CEO y ventas de la empresa]

Se va a estimar un modelo de elasticidad constante en el que relacione el sueldo de los CEO con las ventas de la empresa. El conjunto de datos es el mismo que se usó en el ejemplo 2.3, salvo que ahora se relaciona *sueldo* con *ventas*. Sea *sales* las ventas anuales de la empresa medidas en millones de dólares. Un modelo de elasticidad constante es

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \tag{2.45}$$

donde β_1 es la elasticidad de *salary* respecto a *sales*. Este modelo cae dentro de los modelos de regresión simple mediante la definición de la variable dependiente como $y = \log(salary)$ y de la variable independiente como $x = \log(sales)$. Estimando esta ecuación mediante MCO se obtiene

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \tag{2.46}$$

$$n = 209, R^2 = 0.211.$$

El coeficiente de log(*sales*) es la elasticidad estimada de *salary* (sueldo) respecto a *sales* (ventas). Esto implica que por cada aumento de 1% en las ventas de la empresa hay un aumento de aproximadamente 0.257% en el sueldo de los CEO —la interpretación usual de una elasticidad.

Las dos formas funcionales vistas en esta sección aparecerán con frecuencia en el resto de este libro. Estos modelos con logaritmos naturales se han visto aquí debido a que se encuentran con frecuencia en la práctica. En el caso de la regresión múltiple la interpretación de estos modelos no será muy diferente.

También es útil observar lo que ocurre con las estimaciones del intercepto y la pendiente cuando las unidades de medición de la variable dependiente cambian y esta variable aparece en forma logarítmica. Dado que el cambio a dicha forma aproxima un cambio proporcional, es razonable que no suceda *nada* con la pendiente. Esto se puede ver escribiendo cada observación i de la variable reescalada como $c_1 y_i$. La ecuación original es $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$. Si se agrega $\log(c_1)$ a ambos lados, se obtiene $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$, es decir $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. ((Recuerde que la suma de logaritmos es igual al logaritmo de sus productos, como se muestra en el apéndice A.) Por tanto, la pendiente sigue siendo β_1, pero el intercepto es ahora $\log(c_1) + \beta_0$. De manera similar, si la variable independiente es $\log(x)$ y se modifican las unidades de x antes de obtener el logaritmo, la pendiente sigue siendo la misma, pero el intercepto cambia. En el problema 2.9 se pide verificar esto.

Esta subsección se termina resumiendo las cuatro formas funcionales que se obtienen empleando ya sea la variable original o su logaritmo natural. En la tabla 2.3, x y y representan las variables en su forma original. Al modelo en el que y es la variable dependiente y x es la variable independiente se le llama modelo *nivel-nivel* debido a que las variables aparecen en sus unidades de medición original. Al modelo en el que $\log(y)$ es la variable dependiente y x la variable independiente se le llama modelo *log-nivel*. El modelo *log-nivel* no será analizado aquí explícitamente debido a que se encuentra con menos frecuencia en la práctica. No obstante, en capítulos posteriores se verán ejemplos de este modelo.

TABLA 2.3

Resumen de las formas funcionales en las que se emplean logaritmos

Modelo	Variable dependiente	Variable independiente	Interpretación de β_1
Nivel-nivel	y	x	$\Delta y = \beta_1 \Delta x$
Nivel-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-nivel	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

En la última columna de la tabla 2.3 se da la interpretación de β_1. En el modelo log-nivel, a $100 \cdot \beta_1$ se le suele conocer como la **semielasticidad** de y respecto a x. Como se dijo en el ejemplo 2.11, en el modelo log-log, β_1 es la **elasticidad** de y respecto a x. La tabla 2.3 merece cuidadoso estudio, ya que con frecuencia se hará referencia a ella en el resto del libro.

Significado de regresión "lineal"

Al modelo de regresión simple que se ha estudiado en este capítulo también se le llama modelo de regresión *lineal* simple. Sin embargo, como se acaba de ver, el modelo general también permite ciertas relaciones *no lineales*. Entonces, ¿qué significa aquí "lineal"? Observando la ecuación (2.1) se puede ver que $y = \beta_0 + \beta_1 x + u$. La clave es que esta ecuación es lineal en los *parámetros* β_0 y β_1. No hay restricción alguna en la manera en que y y x estén relacionadas con las variables originales, explicada y explicativa, de interés. Como se vio en los ejemplos 2.10 y 2.11, y y x pueden ser los logaritmos naturales de estas variables, lo que es muy común en las aplicaciones. Pero esto no es todo. Por ejemplo, nada impide que se use la regresión simple para estimar un modelo como el siguiente: $cons = \beta_0 + \beta_1\sqrt{inc} + u$, donde *cons* es el consumo anual e *inc* es el ingreso anual.

Mientras que la mecánica de la regresión simple no depende de la manera en que estén definidas y y x, la interpretación de los coeficientes sí depende de sus definiciones. Para realizar un trabajo empírico exitoso, es mucho más importante tener la capacidad de interpretar los coeficientes que tener destreza para calcular fórmulas como la (2.19). Cuando se estudie la regresión múltiple se logrará mucho más práctica en la interpretación de las estimaciones de la línea de regresión de MCO.

Hay cantidad de modelos que *no pueden* ser enmarcados en el modelo de regresión lineal ya que no son lineales en sus parámetros; un ejemplo es: $cons = 1/(\beta_0 + \beta_1 inc) + u$. La estimación de estos modelos nos lleva al campo de los *modelos de regresión no lineal*, que queda fuera del alcance de este libro. Para la mayoría de las aplicaciones, es suficiente elegir un modelo que se pueda situar dentro del marco de la regresión lineal.

2.5 Valores esperados y varianzas de los estimadores de MCO

En la sección 2.1 se definió el modelo poblacional $y = \beta_0 + \beta_1 x + u$, y se dijo que el principal supuesto para que el análisis de regresión simple sea útil es que el valor esperado de u dado cualquier valor de x sea cero. En las secciones 2.2, 2.3 y 2.4, se discutieron las propiedades algebraicas de las estimaciones de MCO. Ahora se vuelve al modelo poblacional para estudiar las propiedades *estadísticas* de MCO. En otras palabras, ahora $\hat{\beta}_0$ y $\hat{\beta}_1$ se considerarán como *estimadores* de los parámetros β_0 y β_1 que aparecen en el modelo poblacional. Esto significa que se estudiarán las propiedades de las distribuciones de los $\hat{\beta}_0$ y $\hat{\beta}_1$ que resultan de las diversas muestras aleatorias que es posible obtener de la población. (En el apéndice C se encuentra la definición de estimador, así como un repaso de sus propiedades.)

Insesgamiento de los estimadores MCO

Se empezará por demostrar el insesgamiento de los estimadores de MCO bajo un conjunto sencillo de supuestos. Para referencias futuras, estos supuestos se enumeran empleando el prefijo “RLS” como siglas de regresión lineal simple. El primer supuesto define el modelo poblacional.

Supuesto RLS.1 (Linealidad de los parámetros)

En el modelo poblacional, la variable dependiente, y, está relacionada con la variable independiente, x, y con el error (o perturbación), u, de la manera siguiente

$$y = \beta_0 + \beta_1 x + u, \tag{2.47}$$

donde β_0 y β_1 representan los parámetros poblacionales, del intercepto y pendiente, respectivamente.

Para ser realistas, al plantear el modelo poblacional, y, x, y u son consideradas como variables aleatorias. En la sección 2.1 se analizó, con cierta profundidad, la interpretación de este modelo y se dieron algunos ejemplos. En la sección anterior, se vio que la ecuación (2.47) no es tan restrictiva como a primera vista pareciera; eligiendo y y x de manera adecuada, se pueden obtener interesantes relaciones no lineales (por ejemplo, modelos de elasticidad constante).

La idea es usar los datos de y y de x para estimar los parámetros β_0 y, en especial, β_1. Se supone que los datos se obtienen de una muestra aleatoria. (Vea en el apéndice C un repaso sobre muestreo aleatorio.)

Supuesto RLS.2 (Muestreo aleatorio)

Se cuenta con una muestra aleatoria de tamaño n, $\{(x_i,y_i): i = 1, 2, \ldots, n\}$, que sigue el modelo poblacional de la ecuación.

En capítulos posteriores sobre el análisis de series de tiempo y problemas de selección de la muestra habrá que ocuparse de la falla del supuesto de muestreo aleatorio. No todas las muestras de corte transversal pueden considerarse como resultado de un muestreo aleatorio, aunque muchas pueden serlo.

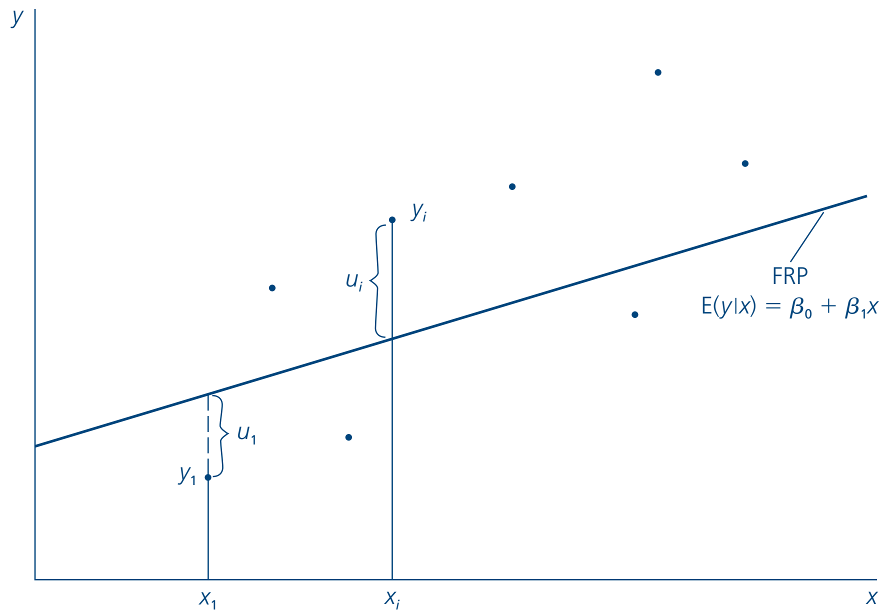
En términos de la muestra aleatoria, la ecuación (2.47) se expresa como

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, n, \tag{2.48}$$

donde u_i es el error o la perturbación para la observación i (por ejemplo, en la persona i, la empresa i, la ciudad i, etc.). Así que u_i contiene los efectos no observables de la observación i que afectan a y_i. Esta u_i no debe confundirse con los residuales, $\hat{u}_i$, definidos en la sección 2.3. Más

FIGURA 2.7

Gráfica de $y_i = \beta_0 + \beta_1 x_i + u_i$.



adelante se analizará la relación entre los errores y los residuales. Para interpretar β_0 y β_1 en una determinada aplicación, la ecuación (2.47) es la más informativa, pero la ecuación (2.48) es necesaria para algunas deducciones estadísticas.

Dado un conjunto de datos, la relación (2.48) puede graficarse como se muestra en la figura 2.7.

Con base en lo visto en la sección 2.2, las estimaciones de MCO del intercepto y de la pendiente sólo están definidas si en la muestra hay variación de la variable explicativa. A continuación se agrega a la lista de supuestos la variación de las x_i.

Supuesto RLS.3 (Variación muestral de la variable explicativa)

No todos los valores muestrales de x, a saber $\{x_i, i = 1, \ldots, n\}$, son iguales, es decir, no todos tienen el mismo valor.

Este es un supuesto muy débil —ciertamente poco útil de destacar, pero necesario de cualquier manera—. Si x varía en la población, entonces las muestras aleatorias típicamente mostrarán variación en x, a menos que la variación poblacional sea mínima o el tamaño de la muestra sea muy pequeño. Una sencilla inspección de los estadísticos de las x_i indicará si el supuesto RLS.3 se satisface o no: si la desviación estándar muestral de las x_i es cero, entonces el supuesto RLS.3 no se satisface; si no es así, este supuesto se satisface.

Por último, con objeto de obtener estimadores insesgados de β_0 y β_1, es necesario imponer el supuesto de media condicional cero, visto con cierto detalle en la sección 2.1. A continuación se agrega éste a la lista de supuestos.

Supuesto RLS.4 (Media condicional cero)

Para todo valor de la variable explicativa, el valor esperado del error u es cero. Es decir,

$$\mathrm{E}(u|x) = 0.$$

En una muestra aleatoria, este supuesto implica que $\mathrm{E}(u_i|x_i) = 0$, para toda $i = 1, 2, \ldots, n$.

Además de restringir la relación que hay en la población entre u y x, el supuesto de media condicional cero —junto con el supuesto del muestreo aleatorio— permite una conveniente simplificación técnica. En particular, es posible obtener las propiedades estadísticas de los estimadores de MCO como *condicionales* sobre los valores de x_i en la muestra. Técnicamente, al hacer cálculos estadísticos, condicionar sobre los valores muestrales de la variable independiente es lo mismo que tratar a las x_i como *fijas en muestras repetidas*, lo cual se entiende como sigue. Primero se escogen n valores muestrales para $x_1, x_2, \ldots, x_n$. (Éstos pueden repetirse.) Dados estos valores, se obtiene después una muestra de y (efectivamente al obtener una muestra aleatoria de las u_i). A continuación, se obtiene otra muestra de y, usando los *mismos* valores para $x_1, x_2, \ldots, x_n$. Después se obtiene otra muestra de y, usando los mismos $x_1, x_2, \ldots, x_n$. Y así sucesivamente.

La situación de que las x_i sean fijas en muestras repetidas no es muy realista en los contextos no experimentales. Por ejemplo, al muestrear individuos para el caso del salario y la educación, no tiene mucho sentido pensar en elegir de antemano los valores de *educ* y después muestrear individuos que tengan esos determinados niveles de educación. El muestreo aleatorio, en el que los individuos se eligen de manera aleatoria para conocer su salario y su educación, es representativo de la manera en que se obtienen las bases de datos para el análisis empírico en las ciencias sociales. Una vez que se *supone* que $\mathrm{E}(u|x) = 0$, y que se tiene un muestreo aleatorio, no cambia nada en los cálculos al tratar a las x_i como no aleatorias. El riesgo es que el supuesto de que las x_i sean fijas en muestras repetidas *siempre* implica que u_i y x_i sean independientes. Para decidir si con un análisis de regresión simple se van a obtener estimadores insesgados, es crucial pensar en los términos del supuesto RLS.4.

Ahora, ya se puede mostrar que los estimadores de MCO son insesgados. Para esto se usa el hecho de que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$ (vea el apéndice A) para escribir el estimador de la pendiente de MCO de la ecuación (2.19) como

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad \textbf{2.49}$$

Dado que ahora la atención se centra en el comportamiento de las $\hat{\beta}_1$ en todas las muestras posibles, es apropiado considerar a $\hat{\beta}_1$ como una variable aleatoria.

Sustituyendo el lado derecho de la ecuación (2.48) en la ecuación (2.49) se puede escribir $\hat{\beta}_1$ en términos de los coeficientes poblacionales y de los errores. Se tiene

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\mathrm{STC}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\mathrm{STC}_x}, \quad \textbf{2.50}$$

donde, para simplificar la notación, la variación total de las x_i se ha definido como $\text{STC}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$. (Esta no es exactamente la varianza muestral de las x_i ya que no está dividida entre $n - 1$.) Si se emplean las propiedades de la sumatoria, el numerador de $\hat{\beta}_1$ se puede escribir como

$$\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i$$

$$= \beta_0 \sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i. \quad \boxed{2.51}$$

Como se muestra en el apéndice A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ y $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{STC}_x$. Por tanto, el numerador de $\hat{\beta}_1$ se puede escribir como $\beta_1\text{STC}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Colocando esto sobre el denominador se obtiene

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{STC}_x} = \beta_1 + (1/\text{STC}_x)\sum_{i=1}^{n} d_i u_i, \quad \boxed{2.52}$$

donde $d_i = x_i - \bar{x}$. Ahora se puede ver que el estimador $\hat{\beta}_1$ es igual a la pendiente poblacional, β_1, más un término que es una combinación lineal de los errores $\{u_1, u_2, \ldots, u_n\}$. Condicionada sobre los valores de las x_i, la aleatoriedad de $\hat{\beta}_1$ se debe por completo a los errores de la muestra. El hecho de que estos errores en general sean distintos de cero es lo que hace que $\hat{\beta}_1$ sea diferente de β_1.

Empleando la representación dada en la ecuación (2.52) se puede demostrar la primera propiedad estadística importante de los estimadores de MCO.

Teorema 2.1 (Insesgamiento de los estimadores de MCO)

Empleando los supuestos RLS.1 a RLS.4,

$$\text{E}(\hat{\beta}_0) = \beta_0 \text{ y } \text{E}(\hat{\beta}_1) = \beta_1, \quad \boxed{2.53}$$

para cualquier valor de β_0 y β_1. Es decir, $\hat{\beta}_0$ es un estimador insesgado de β_0 y $\hat{\beta}_1$ es un estimador insesgado de β_1.

DEMOSTRACIÓN: En esta demostración, los valores esperados son condicionales sobre los valores muestrales de la variable independiente. Como STC_x y d_i son funciones sólo de las x_i, no son aleatorias bajo el condicionamiento. Por tanto, de acuerdo con la ecuación (2.52) y manteniendo implícito el condicionamiento sobre $\{x_1, x_2, \ldots, x_n\}$ se tiene

$$\text{E}(\hat{\beta}_1) = \beta_1 + \text{E}\Big[(1/\text{STC}_x)\sum_{i=1}^{n} d_i u_i\Big] = \beta_1 + (1/\text{STC}_x)\sum_{i=1}^{n}\text{E}(d_i u_i)$$

$$= \beta_1 + (1/\text{STC}_x)\sum_{i=1}^{n} d_i \text{E}(u_i) = \beta_1 + (1/\text{STC}_x)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

donde se ha usado el hecho de que el valor esperado de cada u_i (condicionado sobre $\{x_1, x_2, \ldots, x_n\}$) es cero bajo los supuestos RLS.2 y RLS.4. Dado que el insesgamiento es válido para cualesquiera valores de $\{x_1, x_2, \ldots, x_n\}$, la insesgadez también es válida sin el condicionamiento sobre $\{x_1, x_2, \ldots, x_n\}$.

La demostración para $\hat{\beta}_0$ es ahora inmediata. Se promedia (2.48) sobre todas las i y se obtiene $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, y esto se sustituye en la fórmula para $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Entonces, condicionado sobre los valores de las x_i,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

ya que $\mathrm{E}(\bar{u}) = 0$ de acuerdo con los supuestos RLS.2 y RLS.4. Pero, ya se mostró que $\mathrm{E}(\hat{\beta}_1) = \beta_1$, lo que implica que $\mathrm{E}[(\hat{\beta}_1 - \beta_1)] = 0$. Por tanto, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Estos dos argumentos son válidos para todos los valores de β_0 y β_1, con lo que queda demostrado el insesgamiento.

Hay que recordar que el insesgamiento es una propiedad de las distribuciones muestrales de $\hat{\beta}_1$ y $\hat{\beta}_0$, que no dice nada acerca de las estimaciones que se obtienen a partir de una determinada muestra. Se espera que, si la muestra es de alguna manera "representativa", la estimación deberá estar "cerca" del valor poblacional. Desafortunadamente, siempre es posible tener una muestra con la que la estimación puntual que se obtenga esté lejos de β_1, y *nunca* se podrá saber con seguridad si éste es el caso. En el apéndice C se presenta un repaso sobre estimadores insesgados y un ejercicio de simulación en la tabla C.1 que ilustra el concepto de insesgamiento.

En general, el insesgamiento no se cumple cuando no se satisface alguno de los cuatro supuestos. Esto significa que es muy importante reflexionar sobre la veracidad de cada supuesto en la aplicación particular de que se trate. El supuesto RLS.1 requiere que y y x estén relacionadas linealmente y tengan una perturbación aditiva. Es claro que esto puede no darse. Pero también se sabe que pueden escogerse y y x de manera que den interesantes relaciones no lineales. Cuando la ecuación (2.47) no se satisface, se requieren otros métodos más avanzados que quedan fuera del alcance de este libro.

Más adelante habrá que relajar el supuesto RLS.2, el del muestreo aleatorio, al tratar el análisis de las series de tiempo. ¿Y respecto a su uso en análisis de cortes transversales? El muestreo aleatorio puede no satisfacerse en un corte transversal si la muestra no es representativa de la población subyacente; en realidad, hay bases de datos que se obtienen muestreando exageradamente, de manera intencionada, diferentes partes de la población. En los capítulos 9 y 17 se analizarán los problemas del muestreo no aleatorio.

Como ya se ha visto, el supuesto RLS.3 es casi siempre satisfecho en las aplicaciones interesantes de la regresión. Sin este supuesto no es posible ni siquiera obtener los estimadores de MCO.

El supuesto a considerar ahora es RLS.4 si éste se satisface, los estimadores de MCO son insesgados. Y de manera similar, si RLS.4 no se satisface, los estimadores de MCO son, por lo general, *sesgados*. Hay maneras de determinar cuál puede ser la dirección y el tamaño de este sesgo; esto se estudiará en el capítulo 3.

La posibilidad de que x esté correlacionada con u es una preocupación común en el análisis de regresión simple con datos no experimentales, como se indicó mediante varios ejemplos en la sección 2.1. Cuando se usa regresión simple, si u contiene factores que afectan a y y que están correlacionados con x, puede obtenerse una *correlación espuria*: esto significa que se encuentra una relación entre y y x que en realidad se debe a factores no observados que afectan a y y que resultan estar correlacionados con x.

Ejemplo 2.12

[Desempeño de los estudiantes en matemáticas y el programa de desayunos escolares]

Sea *math10* el porcentaje de estudiantes que aprueban el examen estandarizado de matemáticas en el primer año de bachillerato de una escuela. Suponga que se desea estimar el efecto del programa federal de desayunos escolares sobre el desempeño de los estudiantes. Por supuesto, se espera que este programa tenga, un efecto *ceteris paribus*, positivo sobre el desempeño: si todos los demás factores permanecen constantes, si a un estudiante que es tan pobre como para no tener una buena alimentación se le beneficia con el programa de desayunos escolares, su desempeño deberá mejorar. Sea *lnchprg* el porcentaje de estudiantes beneficiados con el programa de desayunos escolares. Entonces, un modelo de regresión simple es

$$math10 = \beta_0 + \beta_1 lnchprg + u, \tag{2.54}$$

donde u contiene características de la escuela y del estudiante que afectan el desempeño general de la escuela. Empleando los datos de MEAP93.RAW, correspondientes a 408 escuelas de Michigan durante el ciclo escolar 1992-1993, se obtiene

$$\widehat{math10} = 32.14 - 0.319\ lnchprg$$
$$n = 408,\ R^2 = 0.171.$$

Esta ecuación predice que si el porcentaje de alumnos que reciben el desayuno escolar aumenta 10 puntos porcentuales, el porcentaje de alumnos que aprueban el examen de matemáticas *decrecerá* aproximadamente 3.2 puntos porcentuales. ¿Se puede creer que un aumento en el porcentaje de estudiantes que reciben el desayuno escolar *cause* un peor desempeño? Casi seguro que no. Una explicación es que el término del error u de la ecuación (2.54) esté correlacionado con *lnchprg*. En realidad, u contiene factores como la tasa de pobreza de los niños que asisten a la escuela, lo cual afecta el desempeño del estudiante y está fuertemente correlacionado con que se otorgue el programa del desayuno. Variables tales como la calidad de la escuela y los recursos de que dispone están contenidas en u, y pueden estar correlacionadas con *lnchprg*. Es importante recordar que la estimación -0.319 sólo es válida para esta muestra en particular, pero su signo y magnitud hacen sospechar que u y x estén correlacionadas, haciendo que la regresión simple esté sesgada.

Además de las variables omitidas, hay otras razones que hacen que x esté correlacionada con u en el modelo de regresión simple. Dado que este mismo tema surge en el análisis de regresión múltiple, el tratamiento sistemático de este problema se pospondrá hasta entonces.

Varianza de los estimadores de mínimos cuadrados ordinarios

Además de saber que la distribución muestral de $\hat{\beta}_1$ está centrada en β_1 ($\hat{\beta}_1$ es insesgado), también es importante saber qué tanto puede esperarse que $\hat{\beta}_1$ se aleje, en promedio, de β_1. Entre otras cosas, esto permite elegir el mejor estimador de todos, o por lo menos, de una amplia clase de estimadores insesgados. La medida de la dispersión de la distribución de $\hat{\beta}_1$ (y de $\hat{\beta}_0$) con la que es más fácil trabajar es con la varianza o su raíz cuadrada, la desviación estándar. (Vea en el apéndice C un análisis más detallado.)

La varianza de los estimadores de MCO puede ser calculada con los supuestos RLS.1 a RLS.4. Sin embargo, estas expresiones son un poco complicadas. En lugar de esto, se agregará un supuesto tradicional en el análisis de corte transversal y que establece que la varianza de los factores inobservables, u, condicionales en x, es constante. Esto se conoce como el supuesto de **homocedasticidad** o de "varianza constante".

Supuesto RLS.5 (Homocedasticidad)

El error u tiene la misma varianza para cualquier valor de la variable explicativa. En otras palabras,

$$\mathrm{Var}(u|x) = \sigma^2.$$

Hay que señalar que el supuesto de homocedasticidad es totalmente distinto al de media condicional cero, $\mathrm{E}(u|x) = 0$. El supuesto RLS.4 se refiere al *valor esperado* de u, mientras que el RLS.5 está relacionado con la *varianza* de u (ambos condicionales sobre x). Recuerde que el insesgamiento de los estimadores de MCO se demostró sin usar el supuesto RLS.5: el supuesto de homocedasticidad *no* se emplea en la demostración de que $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados. El supuesto RLS.5 se agrega debido a que simplifica los cálculos de las varianzas de $\hat{\beta}_0$ y $\hat{\beta}_1$ y a que implica que los mínimos cuadrados ordinarios tienen ciertas propiedades de eficiencia, que se verán en el capítulo 3. Si se supone que u y x son *independientes*, entonces la distribución de u dada x no depende de x, y entonces $\mathrm{E}(u|x) = \mathrm{E}(u) = 0$ y $\mathrm{Var}(u|x) = \sigma^2$. Pero la independencia es algunas veces muy fuerte de suponer.

Como $\mathrm{Var}(u|x) = \mathrm{E}(u^2|x) - [\mathrm{E}(u|x)]^2$ y $\mathrm{E}(u|x) = 0$, $\sigma^2 = \mathrm{E}(u^2|x)$, lo cual significa que σ^2 es también la esperanza *incondicional de* u^2. Por tanto, $\sigma^2 = \mathrm{E}(u^2) = \mathrm{Var}(u)$, debido a que $\mathrm{E}(u) = 0$. En otras palabras, σ^2 es la varianza *incondicional* de u y por esto a σ^2 también se le suele llamar la **varianza del error** o varianza de la perturbación. La raíz cuadrada de σ^2, σ, es la desviación estándar del error. Una σ mayor, indica que la distribución de los factores inobservables que afectan a y tiene una mayor dispersión.

Con frecuencia es útil escribir los supuestos RLS.4 y RLS.5 en términos de la media condicional y de la varianza condicional de y:

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x. \qquad \textbf{2.55}$$

$$\mathrm{Var}(y|x) = \sigma^2. \qquad \textbf{2.56}$$

En otras palabras, la esperanza condicional de y dada x es lineal en x, pero la varianza de y dada x es constante. Esta situación se grafica en la figura 2.8 donde $\beta_0 > 0$ y $\beta_1 > 0$.

Cuando $\mathrm{Var}(u|x)$ depende de x, se dice que el término del error muestra **heterocedasticidad** (o varianza no constante). Como $\mathrm{Var}(u|x) = \mathrm{Var}(y|x)$, la heterocedasticidad está presente siempre que $\mathrm{Var}(y|x)$ sea función de x.

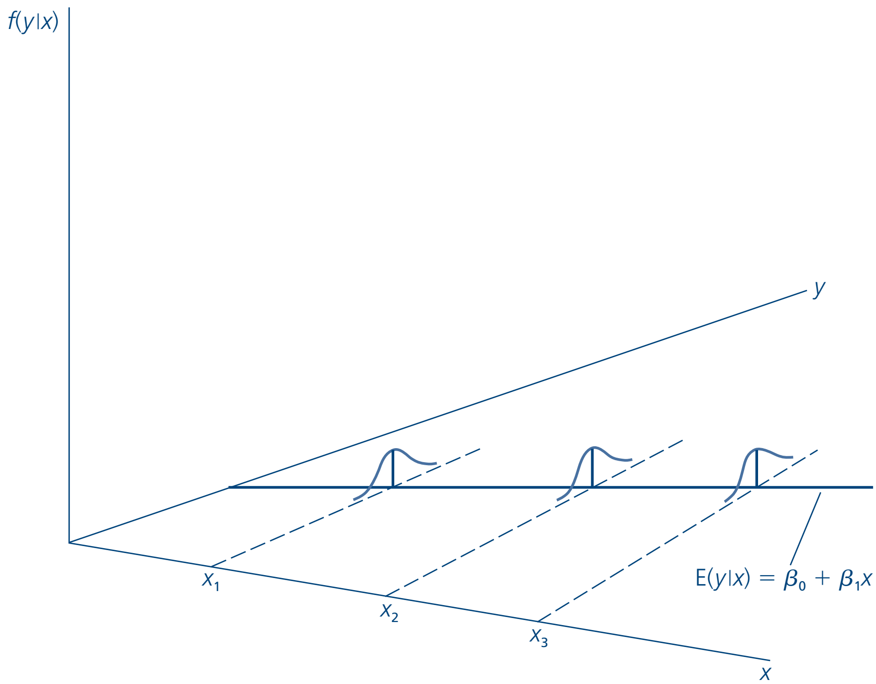
Ejemplo 2.13

[Heterocedasticidad en una ecuación del salario]

Con objeto de obtener un estimador insesgado del efecto *ceteris paribus* de *educ* sobre *wage,* es necesario suponer que $\mathrm{E}(u|educ) = 0$, y esto implica que $\mathrm{E}(wage|educ) = \beta_0 + \beta_1 educ$. Si a esto se agrega el supuesto de homocedasticidad, entonces $\mathrm{Var}(u|educ) = \sigma^2$ no depende del nivel de educación, lo que es lo mismo que suponer que $\mathrm{Var}(wage|educ) = \sigma^2$. Por tanto, mientras se permite que el salario promedio aumente con el nivel de educación —esta es la tasa de incremento que quiere estimar— se supone que la *variabilidad* del salario en torno a su media es constante en todos los niveles de educación. Esto no parece estar de acuerdo con la realidad. Parece ser más posible que las personas con más educación tengan más intereses y más oportunidades de trabajo, lo que puede hacer que a niveles de educación más altos haya mayor variabilidad en el salario. Las personas con niveles de educación muy bajos tienen menos oportunidades y suelen tener que trabajar por el salario mínimo; esto contribuye a reducir la variabilidad del salario en los niveles de educación bajos. Esta situación se muestra en la figura 2.9. Por último, si el supuesto RLS.5 se satisface o no es una cuestión empírica y en el capítulo 8 se muestra cómo probar el supuesto RLS.5.

FIGURA 2.8

Modelo de regresión simple bajo homocedasticidad.



Una vez que se asume la homocedasticidad se está listo para establecer lo siguiente:

Teorema 2.2 (Varianza de muestreo de los estimadores de MCO)

Bajo los supuestos RLS.1 a RLS.5,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\mathrm{STC}_x, \tag{2.57}$$

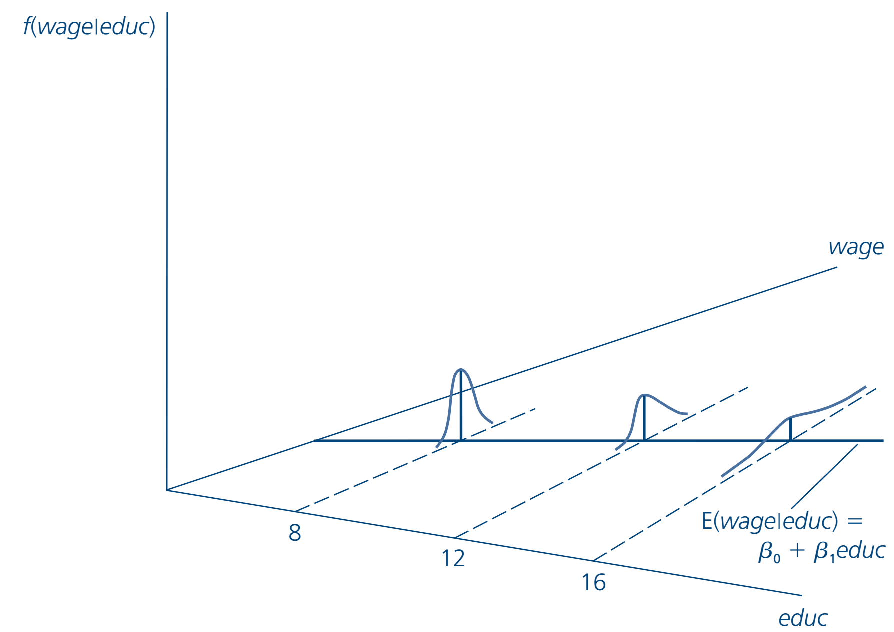
y

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.58}$$

donde éstos son condicionales sobre los valores muestrales $\{x_1, \dots, x_n\}$.

FIGURA 2.9

La Var(*wage*|*educ*) aumenta con *educ*.



DEMOSTRACIÓN: A continuación se deduce la fórmula para la Var($\hat{\beta}_1$), y la otra deducción se deja como ejercicio (problema 2.10). El punto de partida es la ecuación (2.52): $\hat{\beta}_1 = \beta_1 + (1/\text{STC}_x)\sum_{i=1}^{n} d_i u_i$. Dado que β_1 es una constante y que se está condicionando sobre las x_i, STC_x y $d_i = x_i - \bar{x}$ también son no aleatorios. Además, como las u_i son variables aleatorias independientes para toda i (debido al muestreo aleatorio), la varianza de la suma es la suma de las varianzas. Usando estos hechos, se tiene

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= (1/\text{STC}_x)^2 \text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/\text{STC}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \text{Var}(u_i)\right) \\
&= (1/\text{STC}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \sigma^2\right) \qquad [\text{ya que Var}(u_i) = \sigma^2 \text{ para toda } i] \\
&= \sigma^2 (1/\text{STC}_x)^2 \left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2 (1/\text{STC}_x)^2 \text{STC}_x = \sigma^2/\text{STC}_x,
\end{aligned}$$

lo cual es lo que se quería demostrar.

Las ecuaciones (2.57) y (2.58) son las fórmulas "estándar" para el análisis de regresión simple, y no son válidas en presencia de heterocedasticidad. Esto será importante cuando se estudien los intervalos de confianza y las pruebas de hipótesis en el análisis de regresión múltiple.

En la mayoría de los casos, lo que interesa es la $\text{Var}(\hat{\beta}_1)$. Se puede resumir fácilmente cómo esta varianza depende de la varianza del error, σ^2, y de la variación total en $\{x_1, x_2, \ldots, x_n\}$, STC_x. Primero, cuanto mayor sea la varianza del error, mayor es $\text{Var}(\hat{\beta}_1)$. Esto tiene sentido, ya que una mayor variación en los factores no observables que afectan a y hace más difícil estimar β_1 con precisión. Por otro lado, en la variable independiente se prefiere mayor variabilidad: a medida que aumenta la variabilidad de las x_i la varianza de $\hat{\beta}_1$ disminuye. También esto es intuitivamente correcto, ya que entre más dispersa sea la muestra de las variables independientes, más fácil será hallar la relación entre $\text{E}(y|x)$ y x. Es decir, será más sencillo estimar β_1. Si hay poca variación en las x_i, entonces puede ser difícil hallar cómo varía $\text{E}(y|x)$ con la variación en x. A medida que se incrementa el tamaño de la muestra, también aumenta la variación total de las x_i. Por tanto, un tamaño de muestra mayor da como resultado una varianza menor de las $\hat{\beta}_1$.

Este análisis muestra que, si lo que interesa es β_1 y si se tiene la posibilidad de elegir, entonces se debe escoger que las x_i estén tan dispersas como sea posible. Esto es factible algunas veces, cuando se trata de datos experimentales, pero en las ciencias sociales rara vez se puede tener esos lujos: por lo general, hay que conformarse con las x_i que se obtengan mediante un muestreo aleatorio. Algunas veces, se tiene la oportunidad de contar con tamaños de muestra mayores, aunque esto puede ser costoso.

Pregunta 2.5

Muestre que, para estimar β_0, lo mejor es que $\bar{x} = 0$. ¿Cuál es la $\text{Var}(\hat{\beta}_0)$ en este caso? [*Sugerencia:* en cualquier muestra de números, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, donde la igualdad se da sólo si $\bar{x} = 0$.]

Para la construcción de intervalos de confianza y para la obtención de estadísticos de prueba, será necesario trabajar con las desviaciones estándar de $\hat{\beta}_1$ y $\hat{\beta}_0$, $\text{de}(\hat{\beta}_1)$ y $\text{de}(\hat{\beta}_0)$. Recuerde que éstas se obtienen al calcular la raíz cuadrada de las varianzas dadas en las ecuaciones (2.57) y (2.58). En particular, $\text{de}(\hat{\beta}_1) = \sigma/\sqrt{\text{STC}_x}$, donde σ es la raíz cuadrada de σ^2, y $\sqrt{\text{STC}_x}$ es la raíz cuadrada de STC_x.

Estimación de la varianza del error

Las fórmulas de las ecuaciones (2.57) y (2.58) permiten aislar los factores que contribuyen a la $\text{Var}(\hat{\beta}_1)$ y a la $\text{Var}(\hat{\beta}_0)$. Pero estas fórmulas son desconocidas, salvo en el raro caso en que se conozca σ^2. No obstante, usando los datos puede estimarse σ^2, lo que entonces permite estimar la $\text{Var}(\hat{\beta}_1)$ y la $\text{Var}(\hat{\beta}_0)$.

Esta es una buena ocasión para hacer hincapié en la diferencia entre los *errores* (o perturbaciones) y los *residuales*, ya que es crucial para construir un estimador de σ^2. La ecuación (2.48) muestra cómo escribir el modelo poblacional en términos de una observación muestral aleatoria como $y_i = \beta_0 + \beta_1 x_i + u_i$, donde u_i es el error en la observación i. También se puede expresar y_i en términos de su valor ajustado y su residual como en la ecuación (2.32): $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$. Comparando estas dos ecuaciones, se ve que el error aparece en la ecuación que contiene los parámetros *poblacionales*, β_0 y β_1. Por otro lado, los residuales aparecen en la ecuación *estimada* con $\hat{\beta}_0$ y $\hat{\beta}_1$. Los errores jamás son observables, mientras que los residuales se calculan a partir de los datos.

Usando las ecuaciones (2.32) y (2.48) se pueden expresar los residuales en función de los errores:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

o

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i. \tag{2.59}$$

Aunque el valor esperado de $\hat{\beta}_0$ es igual a β_0, y lo mismo ocurre con $\hat{\beta}_1$, $\hat{u}_i$ no es lo mismo que u_i. Pero el *valor esperado* de la diferencia entre ellos sí es cero.

Una vez comprendida la diferencia entre los errores y los residuales, se puede volver a la estimación de σ^2. Primero, $\sigma^2 = \mathrm{E}(u^2)$, de manera que un "estimador" insesgado de σ^2 es $n^{-1}\sum_{i=1}^{n} u_i^2$. Por desgracia, éste no es un verdadero estimador, ya que los errores u_i no pueden observarse. Pero, se tienen estimaciones de las u_i, a saber, los residuales de MCO $\hat{u}_i$. Sustituyendo los errores por los residuales de MCO, se tiene $n^{-1}\sum_{i=1}^{n}\hat{u}_i^2 = \mathrm{SRC}/n$. Éste *es* un verdadero estimador, porque da una regla de cómo calcular su valor a partir de cualquier muestra de datos de x y y. Una pequeña desventaja de este estimador es que es sesgado (aunque si n es grande el sesgo es pequeño). Como es fácil obtener un estimador insesgado, se usará uno de este tipo.

El estimador SRC/n es sesgado debido esencialmente a que no toma en cuenta dos restricciones que deben satisfacer los residuales de MCO. Estas restricciones están dadas por las dos condiciones de primer orden de MCO:

$$\sum_{i=1}^{n}\hat{u}_i = 0, \quad \sum_{i=1}^{n}x_i\hat{u}_i = 0. \tag{2.60}$$

Una manera de ver estas restricciones es: si se conocen $n - 2$ residuales, entonces los otros dos pueden obtenerse empleando las restricciones que implican las condiciones de primer orden de las ecuaciones en (2.60). Así, para los residuales de MCO hay sólo $n - 2$ **grados de libertad**, a diferencia de los n grados de libertad de los errores. Si en las ecuaciones en (2.60) se sustituye $\hat{u}_i$ por u_i las restricciones ya no son válidas. En el estimador insesgado de σ^2 que se usará aquí, se hace un ajuste para los grados de libertad:

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{u}_i^2 = \mathrm{SRC}/(n-2). \tag{2.61}$$

(Este estimador suele en ocasiones denotarse como s^2, pero se continuaría usando la convención de colocar un "gorro" sobre los estimadores.)

Teorema 2.3 (Estimación insesgada de σ^2)

Bajo los supuestos RLS.1 a RLS.5,

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2.$$

DEMOSTRACIÓN: Si en la ecuación (2.59) se determina el promedio sobre todas las i y se usa el hecho de que el promedio de los residuales de MCO es cero, se obtiene $0 = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x}$; restando esto de la ecuación (2.59) se llega a $\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Por tanto, $\hat{u}_i^2 = (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Sumando sobre todas las i se obtiene $\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}u_i(x_i - \bar{x})$. Ahora, el valor esperado del primer término es $(n-1)\sigma^2$, lo cual se muestra en el apéndice C. El valor esperado del segundo término es simplemente σ^2 dado que $\mathrm{E}[(\hat{\beta}_1 - \beta_1)^2] = \mathrm{Var}(\hat{\beta}_1) = \sigma^2/s_x^2$. Y por último, el tercer término se puede escribir como $2(\hat{\beta}_1 - \beta_1)^2 s_x^2$; tomando la esperanza se obtiene $2\sigma^2$. Estos tres términos juntos dan $\mathrm{E}\left(\sum_{i=1}^{n}\hat{u}_i^2\right) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$, de manera que $\mathrm{E}[\mathrm{SRC}/(n-2)] = \sigma^2$.

Si en las fórmulas de la varianza (2.57) y (2.58) se sustituye $\hat{\sigma}^2$, se obtienen estimadores insesgados de Var($\hat{\beta}_1$) y Var($\hat{\beta}_0$). Más tarde se necesitarán estimadores de las desviaciones estándar de $\hat{\beta}_1$ y $\hat{\beta}_0$ y, para esto, se necesita estimar σ. El estimador natural de σ es

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{2.62}$$

al que se le llama **error estándar de la regresión** (**EER**). (Otros nombres para $\hat{\sigma}$ son *error estándar de la estimación* y *raíz del error cuadrático medio*, pero no se usaran aquí.) Aunque $\hat{\sigma}$ no es un estimador insesgado de σ, se puede mostrar que es un estimador *consistente* de σ (vea el apéndice C) y será útil para nuestros propósitos.

La $\hat{\sigma}$ estimada es interesante porque es una estimación de la desviación estándar de los factores no observables que afectan a *y*; de manera equivalente, es una estimación de la desviación estándar de *y* después de haber eliminado el efecto de *x*. La mayoría de los paquetes para regresión dan el valor de $\hat{\sigma}$ junto con la *R*-cuadrada, el intercepto, la pendiente y otros estadísticos MCO (bajo alguno de los nombres dados arriba). Por ahora, el interés primordial es usar $\hat{\sigma}$ para estimar las desviaciones estándar de $\hat{\beta}_0$ y $\hat{\beta}_1$. Como de($\hat{\beta}_1$) $= \sigma/\sqrt{\text{STC}_x}$, el estimador natural de de($\hat{\beta}_1$) es

$$\text{de}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{STC}_x} = \hat{\sigma}\Big/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2};$$

al que se le llama **error estándar de $\hat{\beta}_1$**. Observe que ee($\hat{\beta}_1$) se ve como una variable aleatoria cuando se piensa en aplicar MCO a diferentes muestras de *y*; esto es cierto porque $\hat{\sigma}$ varía en las distintas muestras. En una muestra dada, se($\hat{\beta}_1$) es un número, de la misma manera que $\hat{\beta}_1$ es sólo un número cuando se calcula a partir de los datos dados.

De manera similar, ee($\hat{\beta}_0$) se obtiene de de($\hat{\beta}_0$) sustituyendo σ por $\hat{\sigma}$. El error estándar de cualquier estimación da una idea de qué tan preciso es el estimador. Los errores estándar son de gran importancia en todo este libro; se usarán para construir estadísticos de prueba e intervalos de confianza para cada uno de los procedimientos econométricos que se estudien, a partir del capítulo 4.

2.6 Regresión a través del origen

En raros casos, se impone la restricción de que, cuando $x = 0$, el valor esperado de *y* sea cero. Hay ciertas relaciones para las que esto es razonable. Por ejemplo, si el ingreso (*x*) es cero, entonces la recaudación de impuestos al ingreso (*y*) deberá ser cero. Además hay ocasiones en las que un modelo que originalmente tiene un intercepto distinto de cero se transforma en un modelo sin intercepto.

Formalmente, ahora se elige un estimador de la pendiente, al que se le llamará $\tilde{\beta}_1$, y una línea de la forma

$$\tilde{y} = \tilde{\beta}_1 x, \tag{2.63}$$

donde las tildes sobre $\tilde{\beta}_1$ y $\tilde{y}$ se emplean para distinguir este problema del más común en el que se estima un intercepto y una pendiente. A (2.63) se le llama **regresión a través del origen** debido a que la recta de la ecuación (2.63) pasa a través del punto $x = 0$, $\tilde{y} = 0$. Para obtener la estimación de la pendiente de la ecuación (2.63), también se emplea el método de mínimos cuadrados ordinarios, el cual en este caso minimiza la suma de los cuadrados de los residuales:

$$\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2. \tag{2.64}$$

Empleando cálculo de una sola variable, se puede mostrar que $\tilde{\beta}_1$ debe ser solución de la condición de primer orden:

$$\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0. \quad \textbf{2.65}$$

En esta ecuación se puede despejar $\tilde{\beta}_1$:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \quad \textbf{2.66}$$

siempre que no todas las x_i sean cero, caso que se descarta aquí.

Observe la diferencia entre $\tilde{\beta}_1$ y la estimación de la pendiente cuando también se estima el intercepto (en lugar de igualarlo a cero). Estas dos estimaciones son iguales si y sólo si, $\bar{x} = 0$. [Vea la ecuación (2.49) de $\hat{\beta}_1$]. Obtener una estimación de β_1 cuando se usa la regresión a través del origen, no suele usarse en la práctica por buenas razones: si el intercepto $\beta_0 \neq 0$, entonces $\tilde{\beta}_1$ es un estimador sesgado de β_1. En el problema 2.8 se le pide al lector demostrar esto.

RESUMEN

En este capítulo se introdujo el modelo de regresión lineal simple y se vieron sus propiedades básicas. Dada una muestra aleatoria, se usó el método de mínimos cuadrados ordinarios para estimar los parámetros correspondientes a la pendiente y al intercepto en el modelo poblacional. Se demostró el álgebra de la línea de regresión de MCO, comprendiendo el cálculo de los valores ajustados y de los residuales, así como la obtención de cambios que se predicen en la variable dependiente para un cambio dado de la variable independiente. En la sección 2.4, se analizaron dos temas de importancia práctica: 1) el comportamiento de las estimaciones de MCO al cambiar las unidades de medición de la variable dependiente o de la variable independiente y 2) el uso del logaritmo natural para poder tener modelos de elasticidad y de semielasticidad constante.

En la sección 2.5, se mostró que, bajo los cuatro supuestos RLS.1 a RLS.4, los estimadores de MCO son insesgados. El supuesto principal es que dado cualquier valor de la variable independiente x, el término del error u tiene media cero. Desafortunadamente, hay razones para pensar que esto es falso en muchas aplicaciones de la regresión simple a las ciencias sociales, donde los factores omitidos que están en u suelen estar correlacionados con x. Cuando se agrega el supuesto de que la varianza del error, dada x, es constante, se obtienen fórmulas sencillas para las varianzas de muestreo de los estimadores de MCO. Como se vio, la varianza del estimador de la pendiente $\hat{\beta}_1$ aumenta a medida que la varianza del error aumenta y disminuye cuando hay más variación muestral en la variable independiente. También se determinó un estimador insesgado de $\sigma^2 = \mathrm{Var}(u)$.

En la sección 2.6 se analizó brevemente la regresión a través del origen, en donde el estimador de la pendiente se obtiene bajo el supuesto de que el intercepto es cero. Algunas veces esto es útil, pero en la práctica se presenta con muy poca frecuencia.

Aún queda mucho por hacer. Por ejemplo, falta saber cómo probar hipótesis acerca de los parámetros poblacionales, β_0 y β_1. Por tanto, aunque se sabe que bajo los supuestos RLS.1 a RLS.4 los estimadores MCO de los parámetros poblacionales son insesgados, aún no se tiene manera de hacer inferencias acerca de la población. Otros temas, como la eficiencia del método de MCO en relación con otros procedimientos posibles, también han sido omitidos.

Las cuestiones relacionadas con intervalos de confianza, pruebas de hipótesis y eficiencia también tienen central importancia en el análisis de regresión múltiple. Dado que la manera en que se construyen intervalos de confianza y estadísticos de prueba es muy parecida a la de la regresión múltiple —y dado que la regresión simple es sólo un caso especial de la regresión múltiple— para aprovechar más el tiempo es mejor pasar a la regresión múltiple, que tiene mucha más aplicación que la regresión simple. El propósito del capítulo 2 fue hacer pensar al lector en las cuestiones que surgen en el análisis econométrico en un escenario bastante sencillo.

Supuestos de Gauss-Markov para la regresión simple

A continuación se resumen los **supuestos de Gauss-Markov** empleados en este capítulo. Es importante recordar que para mostrar que $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados sólo se necesitan los supuestos RLS.1 a RLS.4. El supuesto de homocedasticidad, RLS.5 se agregó para obtener las fórmulas usuales (2.57) y (2.58) de las varianzas de MCO.

Supuesto RLS.1 (Lineal en los parámetros)

En el modelo poblacional, la variable dependiente, y, está relacionada con la variable independiente, x, y con el error (o perturbación), u, de la manera siguiente:

$$y = \beta_0 + \beta_1 x + u,$$

donde β_0 y β_1 son los parámetros poblacionales correspondientes al intercepto y a la pendiente, respectivamente.

Supuesto RLS.2 (Muestreo aleatorio)

Se tiene una muestra aleatoria de tamaño n, $\{(x_i,y_i): i = 1, 2, \ldots, n\}$, que sigue el modelo poblacional del supuesto RLS.1.

Supuesto RLS.3 (Variación muestral en la variable explicativa)

Los valores muestrales de x, a saber, $\{x_i, i = 1, \ldots, n\}$, no son todos iguales.

Supuesto RLS.4 (Media condicional cero)

Dado cualquier valor de la variable explicativa, el valor esperado de error u es cero. En otras palabras,

$$\mathrm{E}(u|x) = 0.$$

Supuesto RLS.5 (Homocedasticidad)

Para cualquier valor de la variable explicativa, el error u tienen la misma varianza. En otras palabras,

$$\mathrm{Var}(u|x) = \sigma^2.$$

TÉRMINOS CLAVE

Coeficiente de determinación
Condiciones de primer orden
Covariada
Elasticidad
Error estándar de $\hat{\beta}_1$
Error estándar de la regresión (EER)
Función de regresión muestral (FRM)
Función de regresión poblacional (FRP)
Grados de libertad
Heterocedasticidad
Homocedasticidad

Línea de regresión de MCO
Media independiente
Mínimos cuadrados ordinarios (MCO)
Modelo de elasticidad constante
Modelo de regresión lineal simple
Parámetro del intercepto
Parámetro de la pendiente
R-cuadrada
Regresando
Regresión a través del origen
Regresor
Residual
Semielasticidad
Suma explicada de cuadrados (SEC)
Suma de residuales cuadrados (SRC)
Suma total de cuadrados (STC)
Supuesto de media condicional cero
Supuestos de Gauss-Markov
Término de error (perturbación)
Valor ajustado
Variable de control
Variable de respuesta
Variable dependiente
Variable explicada
Variable explicativa
Variable independiente
Variable predicha
Variable predictora
Varianza del error

PROBLEMAS

2.1 Sea *niños* la cantidad de hijos que ha tenido una mujer, y *educ* los años de educación que tiene esta mujer. Un modelo sencillo para relacionar fertilidad con años de educación es

$$niños = \beta_0 + \beta_1 educ + u,$$

donde u es el error no observado.

i) ¿Qué tipo de factores son los contenidos en u? ¿Es posible que estos factores estén correlacionados con el nivel de educación?

ii) ¿Es posible que con un análisis de regresión simple se halle el efecto *ceteris paribus* de educación sobre fertilidad? Explique.

2.2 En el modelo de regresión lineal simple $y = \beta_0 + \beta_1 x + u$, suponga que $\mathrm{E}(u) \neq 0$. Sea $\alpha_0 = \mathrm{E}(u)$, muestre que siempre es posible reescribir el modelo con la misma pendiente, pero con otro intercepto y otro error, de manera que el nuevo error tenga valor esperado cero.

2.3 En la tabla siguiente se presentan las puntuaciones obtenidas en el examen de ingreso a la universidad en Estados Unidos, *ACT* (American College Test), y en el *GPA* (promedio escolar) por ocho estudiantes universitarios. El *GPA* está medido en una escala de cuatro puntos y se ha redondeado a un dígito después del punto decimal.

Estudiante	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

i) Estime la relación entre *GPA* y *ACT* empleando MCO; es decir, obtenga las estimaciones para la pendiente y para el intercepto en la ecuación

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comente la dirección de la relación ¿tiene, en este caso, el intercepto una interpretación útil? Explique, ¿qué tanto más alto será el *GPA* predicho si *ACT* aumenta cinco puntos?

ii) Calcule los valores ajustados y los residuales para cada observación y verifique que los residuales (aproximadamente) sumen cero.

iii) ¿Cuál es el valor que se predice para el *GPA* si $ACT = 20$?

iv) ¿Qué tanto de la variación en el *GPA* de estos ocho estudiantes es explicada por el *ACT*? Explique.

2.4 La base de datos BWGHT.RAW contiene cifras sobre los hijos nacidos de mujeres en Estados Unidos. Las dos variables de interés son la variable independiente, peso en onzas del niño al nacer (*bwght*) y la variable explicativa, cantidad promedio diaria de cigarros consumidos por la madre durante el embarazo (*cigs*). La siguiente ecuación de regresión simple se estimó con datos de $n = 1{,}388$ nacimientos:

$$\widehat{bwght} = 119.77 - 0.514\, cigs$$

i) ¿Cuál es el peso al nacer que se predice si $cigs = 0$? ¿Y cuando $cigs = 20$ (un paquete por día)? Analice la diferencia.

ii) ¿Capta esta ecuación de regresión simple una relación causal entre el peso del niño al nacer y el hábito de fumar de la madre? Explique.

iii) Para que el peso al nacer predicho sea de 125 onzas, ¿cuál tiene que ser el valor de *cigs*? Explique.

iv) La proporción de mujeres en la muestra que no fumaron durante el embarazo es aproximadamente 0.85. ¿Ayuda esto a entender sus hallazgos del inciso iii)?

2.5 En la función lineal de consumo

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 inc,$$

la *propensión marginal a consumir* estimada (PMgC) del ingreso no es más que la pendiente, $\hat{\beta}_1$, mientras que la *propensión media a consumir* (PMeC) es $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Usando las observaciones sobre ingreso anual y consumo de 100 familias (ambos medidos en dólares), se obtiene la ecuación siguiente:

$$\widehat{cons} = -124.84 + 0.853\, inc$$
$$n = 100,\ R^2 = 0.692.$$

i) Interprete el intercepto en esta ecuación y analice su signo y su magnitud.

ii) ¿Cuál es el consumo que se predice si el ingreso familiar es de $30,000?

iii) Con *inc* en el eje *x*, trace una gráfica la PMgC estimada y de la PMeC estimada.

2.6 Usando los datos de Kiel y McClain (1995) sobre las casas vendidas en 1988 en Andover, Massachusetts, en la ecuación siguiente se relaciona el precio de las casas (*price*) con la distancia a un incinerador de basura construido recientemente (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$
$$n = 135,\ R^2 = 0.162.$$

i) Interprete el coeficiente de log(*dist*). ¿Es el signo de esta estimación el que se esperaba?

ii) ¿Considera que la regresión simple proporciona un estimador insesgado de la elasticidad *ceteris paribus* de *price* (precio) respecto a *dist*? (Reflexione en la decisión de la ciudad sobre dónde colocar el incinerador.)

iii) ¿Qué otros factores relacionados con una casa afectan su precio? ¿Pueden estos factores estar correlacionados con la distancia al incinerador?

2.7 Considere la función de ahorro

$$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc} \cdot e,$$

donde e es una variable aleatoria con $E(e) = 0$ y $Var(e) = \sigma_e^2$. Suponga que e es independiente de inc.

i) Muestre que $E(u|inc) = 0$, de manera que el supuesto clave de media condicional cero (supuesto RLS.4) se satisface. [*Sugerencia:* Si e es independiente de inc, *entonces* $E(e|inc) = E(e)$.]

ii) Muestre que $Var(u|inc) = \sigma_e^2 inc$, de manera que se viola el supuesto RLS.5 de homocedasticidad. En particular, la varianza de sav (ahorro) aumenta con inc. [*Sugerencia:* $Var(e|inc) = Var(e)$, si inc y e son independientes.]

iii) Proporcione un análisis para sostener el supuesto de que la varianza de los ahorros aumenta con el ingreso familiar.

2.8 Considere el modelo estándar regresión simple $y = \beta_0 + \beta_1 x + u$ bajo los supuestos RLS.1 a RLS.5 de Gauss-Markov. Los estimadores usuales de MCO $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados para sus respectivos parámetros poblacionales. Sea $\tilde{\beta}_1$ el estimador de β_1 obtenido suponiendo que el intercepto es cero (vea la sección 2.6).

i) Determine $E(\tilde{\beta}_1)$ en términos de x_i β_0 y β_1. Verifique que $\tilde{\beta}_1$ es insesgado respecto a β_1 cuando el intercepto poblacional (β_0) es cero. ¿Hay otros casos en los que $\tilde{\beta}_1$ sea insesgado?

ii) Determine la varianza de $\tilde{\beta}_1$. (*Sugerencia:* La varianza no depende de β_0.)

iii) Muestre que $Var(\tilde{\beta}_1) \leq Var(\hat{\beta}_1)$. [*Sugerencia:* En cualquier muestra de datos, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, donde la desigualdad es estricta a menos que $\bar{x} = 0$.]

iv) Analice el efecto de sustitución que existe entre sesgo y varianza al elegir entre $\hat{\beta}_1$ y $\tilde{\beta}_1$.

2.9 i) Sean $\hat{\beta}_0$ y $\hat{\beta}_1$ el intercepto y la pendiente en la regresión de y_i sobre x_i, usando n observaciones. Sean c_1 y c_2 constantes, con $c_2 \neq 0$. Sean $\tilde{\beta}_0$ y $\tilde{\beta}_1$ el intercepto y la pendiente de la regresión de $c_1 y_i$ sobre $c_2 x_i$. Muestre que $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_0$ y $\tilde{\beta}_0 = c_1\hat{\beta}_0$, verificando con esto lo que se afirma en la sección 2.4 sobre las unidades de medición. [*Sugerencia:* Para obtener $\tilde{\beta}_1$, sustituya las versiones escaladas de x y de y en la ecuación (2.19).] Después, use la ecuación (2.17) para $\tilde{\beta}_0$, usando la x y la y escaladas y la pendiente correcta.]

ii) Ahora, sean $\tilde{\beta}_0$ y $\tilde{\beta}_1$ la intersección con el eje Y y la pendiente en la regresión de $(c_1 + y_i)$ sobre $(c_2 + x_i)$ (sin ninguna restricción sobre c_1 o c_2). Muestre que $\tilde{\beta}_1 = \hat{\beta}_1$ y $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

iii) Ahora, sean $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de MCO en la regresión de $\log(y_i)$ sobre x_i, donde hay que suponer que $y_i > 0$ para toda i. Para $c_1 > 0$, sean $\tilde{\beta}_0$ y $\tilde{\beta}_1$ el intercepto y la pendiente en la regresión de $\log(c_1 y_i)$ sobre x_i. Muestre que $\tilde{\beta}_1 = \hat{\beta}_1$ y $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

iv) Ahora, suponiendo que $x_i > 0$ para toda i, sean $\tilde{\beta}_0$ y $\tilde{\beta}_1$ el intercepto y la pendiente de y_i sobre $\log(c_2 x_i)$. ¿Cuál es la relación de $\tilde{\beta}_0$ y $\tilde{\beta}_1$ con el intercepto y la pendiente de la regresión de y_i sobre $\log(x_i)$?

2.10 Sean $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de MCO del intercepto y de la pendiente, respectivamente, y sea $\bar{u}$ la media muestral de los errores (¡no de los residuales!).

i) Muestre que $\hat{\beta}_1$ se puede expresar como $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$ donde $w_i = d_i/\text{STC}_x$ y $d_i = x_i - \bar{x}$.

ii) Use el inciso i), así como $\sum_{i=1}^{n} w_i = 0$, para mostrar que $\hat{\beta}_1$ y $\bar{u}$ no están correlacionados. [*Sugerencia*: Se pide que demuestre que $\text{E}[(\hat{\beta}_1 - \beta_1) \cdot \bar{u}] = 0$.]

iii) Muestre que $\hat{\beta}_0$ se puede expresar como $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

iv) Use los incisos ii) y iii) para mostrar que $\text{Var}(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2/\text{STC}_x$.

v) Simplifique la expresión del inciso iv) para llegar a la ecuación (2.58). [*Sugerencia*: $\text{STC}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$.]

2.11 Suponga que desea estimar el efecto de las horas invertidas en un curso de preparación para el examen estándar de admisión (SAT) sobre la puntuación obtenida en este examen (*sat*). La población es la de todos los alumnos de último año de bachillerato que están por ingresar a la universidad en un determinado año.

i) Suponga que se le otorga un subsidio para realizar un experimento controlado. Explique cómo estructuraría el experimento con objeto de estimar el efecto causal de *hours* sobre *sat*.

ii) Considere el caso más real en el que los estudiantes deciden cuánto tiempo invertir en el curso de preparación para el examen, y usted sólo puede muestrear de la población, en forma aleatoria, *sat* y *hours*. Escriba un modelo poblacional de la forma

$$sat = \beta_0 + \beta_1 hours = u$$

donde, como de costumbre en un modelo con intercepto, puede suponerse que $\text{E}(u) = 0$. Enumere por lo menos dos factores contenidos en u. ¿Es posible que éstos estén correlacionados positiva o negativamente con *hours*?

iii) En la ecuación del inciso ii), ¿cuál debe ser el signo de β_1 si el curso de preparación es efectivo?

iv) En la ecuación del inciso ii), ¿cuál es la interpretación de β_0?

EJERCICIOS EN COMPUTADORA

C2.1 La base de datos 401K.RAW es un subconjunto de los datos analizados por Papke (1995) para estudiar la relación entre la participación en un plan de pensión y la generosidad del plan. La variable *prate* es el porcentaje de trabajadores que están inscritos en el plan y que tienen cuenta activa; ésta es la variable que se quiere explicar. La medida de la generosidad es la tasa de contribución (de la empresa) al plan, *mrate*. Esta variable es la cantidad promedio con la que la empresa contribuye al plan de cada trabajador por cada $1 que aporte el trabajador. Por ejemplo, si $mrate = 0.50$, entonces a una contribución de $1 del trabajador corresponde una contribución de 50 centavos de la empresa.

i) Encuentre el promedio de la tasa de participación y el promedio de la tasa de contribución para la muestra.

ii) Ahora, estime la ecuación de regresión simple

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

y dé los resultados, el tamaño de la muestra y R-cuadrada.

iii) Interprete el intercepto de la ecuación. Interprete también el coeficiente de *mrate*.

iv) Determine la *prate* que se predice para = 3.5. ¿Es razonable esta predicción? Explique qué ocurre aquí.

v) ¿Qué tanto de la variación en *prate* es explicada por *mrate*? En su opinión, ¿es mucho?

C2.2 La base de datos CEOSAL2.RAW contiene información sobre directores generales de empresas (CEO) estadounidenses. La variable *salary* es el sueldo anual, en miles de dólares, y *ceoten* son los años de antigüedad como CEO de la empresa.

i) Determine el sueldo y la antigüedad promedio en esta muestra.

ii) ¿Cuántos de estos directivos se encuentran en su primer año como CEO (es decir, $ceoten = 0$)? ¿Cuál es la mayor antigüedad entre estos CEO?

iii) Estime el modelo de regresión simple

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

y dé los resultados en la forma usual. ¿Cuál es el aumento porcentual (aproximado) que se pronostica en el sueldo por cada año adicional de antigüedad como CEO?

C2.3 Utilice los datos de SLEEP75.RAW de Biddle y Hamermesh (1990) para analizar si existe una relación inversa entre las horas de sueño por semana y las horas de trabajo pagado por semana. Cualquiera de las variables puede usarse como la variable dependiente. Estime el modelo

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

donde *sleep* corresponde a minutos de sueño por semana durante la noche y *totwrk* corresponde al total de minutos de trabajo por semana.

i) Dé sus resultados en forma de ecuación, además de la cantidad de observaciones y la R^2. ¿Qué significa el intercepto de la ecuación?

ii) Si *totwrk* aumenta 2 horas, ¿cuánto se estima que disminuirá *sleep*? ¿Le parece que este efecto sea grande?

C2.4 Use la base de datos WAGE2.RAW para estimar una regresión simple que explique el salario mensual (*wage*) en términos de la puntuación del coeficiente intelectual (*IQ*).

i) Determine el promedio muestral del salario y de IQ. ¿Cuál es la desviación estándar muestral de IQ? (La puntuación del coeficiente intelectual está estandarizada, de manera que el promedio de la población es 100 y la desviación estándar es 15.)

ii) Estime un modelo de regresión simple en el que un aumento de un punto en *IQ* modifique *wage* en una cantidad de dólares constante. Use este modelo para determinar el aumento que se predice en *wage* para un aumento de 15 puntos en *IQ*. ¿Explica *IQ* la mayor parte de la variación en *wage*?

iii) Ahora, estime un modelo en el que cada aumento de un punto en *IQ* tenga un mismo efecto porcentual sobre *wage*. Si *IQ* aumenta 15 puntos, ¿cuál es el aumento porcentual pronosticado para *wage*?

C2.5 En la población formada por las empresas de la industria química, sea *rd* gastos anuales en investigación y desarrollo y *sales* ventas anuales (ambos en millones de dólares).

i) Dé un modelo (no una ecuación estimada) que implique una elasticidad constante entre *rd* y *sales*. ¿Qué parámetro es la elasticidad?

ii) Ahora, estime el modelo usando la base de datos RDCHEM.RAW. Escriba la ecuación estimada en su forma usual. ¿Cuál es la elasticidad estimada para *rd* respecto a *sales*? Explique qué significa esta elasticidad.

C2.6 En el ejemplo 2.12 se usan los datos de MEAP93.RAW. Ahora se quiere explorar la relación entre la tasa de aprobados en matemáticas (*math10*) y el gasto por estudiante (*expend*).

i) ¿Piensa que cada dólar más que se gasta tiene un mismo efecto en la tasa de aprobados o parece más apropiado que haya un efecto decreciente? Explique.

ii) En el modelo poblacional

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

justifique que $\beta_1/10$ es el cambio en puntos porcentuales en *math10* dado un aumento de 10% en *expend*.

iii) Use los datos de MEAP93.RAW para estimar el modelo del inciso ii). Dé la ecuación estimada de la manera usual, incluyendo el tamaño de la muestra y R-cuadrada.

iv) ¿Qué tan grande es el efecto estimado del gasto? Es decir, si el gasto aumenta 10%, ¿cuál es el aumento de *math10* en puntos porcentuales?

v) Puede ser preocupante que el análisis de regresión produzca valores ajustados para *math10* que sean mayores a 100. ¿Por qué esto no es de preocupar en esta base estándar de datos?

C2.7 Use la base de datos CHARITY.RAW [obtenidos de Franses y Paap (2001)] para responder a las preguntas siguientes:

i) ¿Cuál es el donativo (*gift*) promedio en esta muestra de 4,268 personas (en florines holandeses)? ¿Qué porcentaje de estas personas no dio ningún donativo?

ii) ¿Cuál es el promedio de envíos por año (*mailsyear*)? ¿Cuáles son los valores mínimo y máximo?

iii) Estime el modelo

$$gift = \beta_0 + \beta_1 mailsyear + u$$

mediante MCO y dé los resultados de la manera usual, incluidos el tamaño de la muestra y la R-cuadrada.

iv) Interprete el coeficiente de la pendiente. Si cada envío cuesta un florín, ¿espera la beneficencia obtener una ganancia neta por cada envío? ¿Significa esto que la beneficencia obtiene una ganancia neta en cada envío? Explique.

v) En esta muestra, ¿cuál es el menor donativo? Usando el análisis de regresión simple, ¿se puede predecir que *gift* sea igual a cero?

Apéndice 2A

Minimización de la suma de los residuales cuadrados

Se mostrará que $\hat{\beta}_0$ y $\hat{\beta}_1$ estimados de MCO minimizan la suma de los residuales cuadrados como se afirma en la sección 2.2. Formalmente, el problema es encontrar las soluciones $\hat{\beta}_0$ y $\hat{\beta}_1$ del problema de minimización

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

donde b_0 y b_1 son argumentos ficticios en el problema de optimización; para simplificar, llámesele a esta función $Q(b_0, b_1)$. De acuerdo con una fórmula fundamental del cálculo multivariado (vea el apéndice A), una condición necesaria para que $\hat{\beta}_0$ y $\hat{\beta}_1$ sean soluciones del problema de minimización es que las derivadas parciales de $Q(b_0, b_1)$ respecto a b_0 y b_1 evaluadas en $\hat{\beta}_0$,

$\hat{\beta}_1$: $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ y $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. Con ayuda de la regla de la cadena del cálculo, estas dos ecuaciones se convierten en

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

$$-2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Estas dos ecuaciones son precisamente (2.14) y (2.15) multiplicadas por $-2n$ y, por tanto, son resueltas por los mismos $\hat{\beta}_0$ y $\hat{\beta}_1$.

¿Cómo se sabe que efectivamente se ha minimizado la suma de los residuales cuadrados? Las condiciones de primer orden son necesarias pero no suficientes. Una manera de comprobar que se ha minimizado la suma de los residuales cuadrados es expresando, para cualquier b_0 y b_1,

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2 \\ &= \sum_{i=1}^{n}[\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2 \\ &= \sum_{i=1}^{n}\hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2\sum_{i=1}^{n}x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1)\sum_{i=1}^{n}x_i, \end{aligned}$$

donde se han empleado las ecuaciones (2.30) y (2.31). El primer término no depende de b_0 ni de b_1, mientras que la suma de los últimos tres términos se puede expresar como

$$\sum_{i=1}^{n}[(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2,$$

como puede comprobarse mediante manipulaciones algebraicas sencillas. Dado que ésta es una suma de términos al cuadrado, el menor valor que puede tener es cero. Por tanto, tendrá el menor valor posible cuando $b_0 = \hat{\beta}_0$ y $b_1 = \hat{\beta}_1$.

CAPÍTULO 3

Análisis de regresión múltiple: estimación

En el capítulo 2 se vio cómo usar el análisis de regresión simple para explicar una variable dependiente, *y*, como función de una sola variable independiente, *x*. El principal inconveniente del análisis de regresión simple en el trabajo empírico es que es muy difícil obtener conclusiones *ceteris paribus* de cómo afecta *x* a *y*: el supuesto clave RLS.4 —de que todos los demás factores que afectan a *y* no están correlacionados con *x*— a menudo no es realista.

El **análisis de regresión múltiple** es más adecuado para un análisis *ceteris paribus* debido a que permite controlar *de manera explícita* muchos otros factores que afectan en forma simultánea a la variable dependiente. Esto es importante tanto para probar teorías económicas como para evaluar los efectos de una política cuando hay que apoyarse en datos no experimentales. Debido a que los modelos de regresión múltiple pueden aceptar diversas variables explicativas que tal vez estén correlacionadas, puede esperarse inferir causalidad en casos en los que el análisis de regresión simple podría no dar buenos resultados.

Si al modelo se le agregan factores que pueden ser útiles para explicar *y*, entonces puede explicarse más de la variación en *y*. Por tanto, el análisis de regresión múltiple puede emplearse para construir mejores modelos para predecir la variable dependiente.

Otra ventaja del análisis de regresión múltiple es que puede incorporar relaciones con formas funcionales muy generales. En el modelo de regresión simple, en la ecuación únicamente puede aparecer una función de una sola variable explicativa. Como se verá, el modelo de regresión múltiple permite más flexibilidad.

En la sección 3.1 se introduce de manera formal el modelo de regresión múltiple y se analizan las ventajas de la regresión múltiple sobre la simple. En la sección 3.2 se demuestra cómo estimar los parámetros del modelo de regresión múltiple usando el método de mínimos cuadrados ordinarios. En las secciones 3.3, 3.4 y 3.5 se describen varias propiedades estadísticas de los estimadores de MCO, como el insesgamiento y la eficiencia.

El modelo de regresión múltiple sigue siendo el vehículo más empleado para el análisis empírico en la economía y en otras ciencias sociales. Asimismo, el método de mínimos cuadrados ordinarios se usa de manera general para estimar los parámetros del modelo de regresión múltiple.

3.1 Motivación para la regresión múltiple

El modelo con dos variables independientes

Se empezará con algunos ejemplos sencillos para mostrar el uso del análisis de regresión lineal múltiple para resolver problemas que no es posible resolver mediante regresión simple.

El primer ejemplo es una sencilla variación de la ecuación del salario, presentada en el capítulo 2, para obtener el efecto de la educación sobre el salario por hora:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u, \tag{3.1}$$

donde *exper* es años de experiencia en el mercado de trabajo. Por tanto, *wage* (salario) está determinada por las dos variables independientes o explicativas, educación y experiencia, y por otros factores no observados, contenidos en u. El interés principal sigue siendo el efecto de *educ* (educación) sobre *wage* (salario 1), manteniendo constantes todos los otros factores que afectan a *wage*; es decir, lo que interesa es el parámetro β_1.

Comparada con un análisis de regresión simple, en el que se relaciona *wage* con *educ*, la ecuación (3.1) extrae *exper* del término del error y la coloca de manera explícita en la ecuación. Dado que *exper* aparece en la ecuación, su coeficiente, β_2, mide el efecto *ceteris paribus* de *exper* sobre *wage*, que también es de cierto interés.

Como en la regresión simple, aquí también habrá que hacer supuestos acerca de la relación de u en la ecuación (3.1) con las variables independientes *educ* y *exper*. Pero, como se verá en la sección 3.2, hay algo de lo que se puede estar seguro: como en la ecuación (3.1) aparece la experiencia de manera explícita, se podrá medir el efecto de la educación sobre el salario, manteniendo constante la experiencia. Con un análisis de regresión simple —en el cual *exper* forma parte del término del error— hay que suponer que la experiencia no está correlacionada con la educación, un supuesto cuestionable.

Como segundo ejemplo, considere el problema de explicar el efecto del gasto por estudiante (*expend*) sobre la calificación promedio en el examen estandarizado (*avgscore*) a nivel de bachillerato. Suponga que la calificación promedio en el examen depende del financiamiento, del ingreso familiar promedio (*avginc*) y de otros factores no observables:

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u. \tag{3.2}$$

El coeficiente de interés para los propósitos de las políticas es β_1, el efecto *ceteris paribus* de *expend* sobre *avgscore*. Incluir *avginc* de manera explícita en el modelo permite controlar su efecto sobre *avgscore*. Esto puede ser importante porque el ingreso familiar promedio tiende a estar correlacionado con el gasto por estudiante, el cual suele estar determinado tanto por el impuesto sobre las propiedades inmuebles como por el impuesto local sobre la renta. En un análisis de regresión simple, *avginc* quedaría incluido en el término de error, que es posible que esté correlacionado con *expend*, lo que ocasionaría que en el modelo de dos variables el estimador MCO de β_1 sea sesgado.

En los dos ejemplos anteriores se muestra cómo incluir en el modelo de regresión otros factores observables [*educ* en la ecuación (3.1) y *expend* en la ecuación (3.2)], además de la variable de principal interés. Un modelo con dos variables independientes puede expresarse en general como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.3}$$

donde

β_0 es el intercepto.
β_1 mide el cambio en y respecto a x_1, manteniendo constantes todos los demás factores.
β_2 mide el cambio en y respecto a x_2, manteniendo constantes todos los demás factores.

El análisis de regresión múltiple es útil también para generalizar relaciones funcionales entre variables. Por ejemplo, suponga que el consumo familiar (*cons*) sea una función cuadrática del ingreso familiar (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u, \tag{3.4}$$

donde u contiene otros factores que afectan el consumo. En este modelo, el consumo sólo depende de un factor observado, el ingreso, por lo que parece que puede tratarse en el marco de la regresión simple. Pero este modelo cae fuera de la regresión simple, porque contiene dos funciones del ingreso, *inc* e inc^2 (y por tanto tres parámetros: β_0, β_1 y β_2). Sin embargo, la función consumo puede expresarse de manera sencilla como un modelo de regresión con dos variables independientes haciendo $x_1 = inc$ y $x_2 = inc^2$.

De forma mecánica, *no* habrá *ninguna* diferencia al usar el método de mínimos cuadrados ordinarios (presentado en la sección 3.2) para estimar ecuaciones tan diferentes como la (3.1) y la (3.4). Cada una de ellas puede escribirse como la ecuación (3.3), que es lo único que interesa para los cálculos. Sin embargo, hay una diferencia importante en la *interpretación* de los parámetros. En la ecuación (3.1), β_1 es el efecto *ceteris paribus* de *educ* sobre *wage*. En la ecuación (3.4) no es esta la interpretación del parámetro β_1. En otras palabras, no tiene sentido medir el efecto de *inc* sobre *cons* cuando inc^2 se mantiene constante, porque si *inc* cambia, ¡también cambia inc^2! En lugar de esto, el cambio en consumo respecto al cambio en ingreso —la propensión marginal a consumir— se aproxima mediante

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc.$$

Vea en el apéndice A el cálculo requerido para obtener esta ecuación. En otras palabras, el efecto marginal del ingreso sobre el consumo depende tanto de β_2 como de β_1 y del nivel de ingreso. Este ejemplo muestra que, en cualquier aplicación particular, la definición de las variables independientes es crucial. Pero para el desarrollo teórico de la regresión múltiple, no es necesario ser tan preciso acerca de estos detalles. Ejemplos como éste se estudiarán de manera más cabal en el capítulo 6.

En el modelo con dos variables independientes, el supuesto clave acerca de cómo está relacionado u con x_1 y x_2 es

$$\mathrm{E}(u|x_1, x_2) = 0. \tag{3.5}$$

La interpretación de la condición (3.5) es similar a la del supuesto RLS.4 en el análisis de regresión lineal simple. Esta condición significa que, para cualesquiera valores de x_1 y x_2 en la población, el promedio del efecto de los factores no observables es igual a cero. Como en la regresión simple, la parte importante de este supuesto es que el valor esperado de u es el mismo para todas las combinaciones de x_1 y x_2; que este valor común es cero no es ningún supuesto siempre que el intercepto β_0 se incluya en el modelo (vea la sección 2.1).

¿Cómo puede interpretarse el supuesto de media condicional cero en los ejemplos anteriores? En la ecuación (3.1), este supuesto es $\mathrm{E}(u|educ,exper) = 0$. Esto significa que los otros factores que afectan *wage* no están relacionados en promedio con *educ* y *exper*. Por tanto, si se piensa que la capacidad innata es parte de u, entonces se necesita que los niveles promedio de capacidad sean iguales para todas las combinaciones de educación y experiencia en la población trabajadora. Esto puede ser cierto o no, pero como se verá en la sección 3.3, hay que formular esta pregunta para determinar si el método de mínimos cuadrados ordinarios produce estimadores insesgados.

El ejemplo en el que se mide el desempeño de un estudiante [ecuación (3.2)] es parecido a la ecuación del salario. El supuesto de media condicional cero es $E(u|expend, avginc) = 0$, lo cual significa que los otros factores que afectan las calificaciones —características de la escuela o del estudiante— no están, en promedio, relacionadas con el financiamiento por estudiante y con el ingreso familiar promedio.

Pregunta 3.1

Un modelo sencillo para explicar la tasa de homicidios urbanos (*murdrate*) en términos de la probabilidad de ser convicto (*prbconv*) y la duración promedio de la sentencia (*avgsen*) es

$$murdrate = \beta_0 + \beta_1 prbconv + \beta_2 avgsen + u.$$

¿Cuáles son algunos de los factores contenidos en *u*? ¿Considera probable que se satisfaga el supuesto clave (3.5)?

Aplicado a la función cuadrática de consumo en (3.4), el supuesto de media condicional cero tiene una interpretación un poco diferente. Expresada en forma literal, la ecuación (3.5) se convierte en $E(u|inc,inc^2) = 0$. Como cuando se conoce *inc*, también se conoce inc^2, resulta redundante incluir inc^2 en la esperanza: $E(u|inc,inc^2) = 0$ es lo mismo que $E(u|inc) = 0$. No hay problema si en la esperanza se coloca *inc* e inc^2 al establecer el supuesto, pero $E(u|inc) = 0$ es más concisa.

Modelo con *k* variables independientes

Una vez en el contexto de la regresión múltiple, no es necesario quedarse con sólo dos variables independientes. El análisis de regresión múltiple permite muchos factores observados que afecten a *y*. En el ejemplo del salario también pueden incluirse cantidad de capacitación laboral, años de antigüedad en el empleo actual, mediciones de la capacidad e incluso variables demográficas como cantidad de hermanos o educación de la madre. En el ejemplo del financiamiento de la escuela, otras variables pueden ser mediciones de la calidad de los maestros y tamaño de la escuela.

El **modelo** general **de regresión lineal múltiple** (también llamado *modelo de regresión múltiple*) poblacional puede expresarse como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u, \tag{3.6}$$

donde

β_0 es el **intercepto**.
β_1 es el parámetro asociado con x_1.
β_2 es el parámetro asociado con x_2, y así sucesivamente.

Como hay *k* variables independientes y un intercepto, la ecuación (3.6) contiene $k + 1$ parámetros poblacionales (desconocidos). Por brevedad, a los parámetros distintos del intercepto se les llamará **parámetros de pendiente**, incluso aunque no siempre es esto lo que literalmente son. [Vea la ecuación (3.4), en donde ni β_1 ni β_2 son pendientes, pero juntos determinan la pendiente de la relación entre consumo e ingreso.]

En la regresión múltiple, la terminología es similar a la de la regresión simple y se presenta en la tabla 3.1. Como en la regresión simple, la variable *u* es el **término de error** o **perturbación**. Este término contiene los otros factores distintos de $x_1, x_2, \dots, x_k$ que afectan a *y*. No importa cuántas variables explicativas se incluyan en el modelo, siempre habrá factores que no se pueden incluir y todos ellos juntos están contenidos en *u*.

Cuando se emplea el modelo general de regresión múltiple, hay que saber cómo interpretar los parámetros. En este capítulo y en los siguientes se obtendrá suficiente práctica, pero en este

TABLA 3.1

Terminología de la regresión múltiple

y	$x_1, x_2, \ldots, x_k$
Variable dependiente	Variables independientes
Variable explicada	Variables explicativas
Variable de respuesta	Variables de control
Variable predicha	Variables predictoras
Regresando	Regresores

punto es útil recordar algunas cosas ya sabidas. Suponga que el sueldo (*salary*) de un director general o CEO está relacionado con las ventas de la empresa (*sales*) y su antigüedad en la organización mediante (*ceoten*)

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u. \quad \textbf{3.7}$$

Esta ecuación encaja en el modelo de regresión múltiple (con $k = 3$) definiendo $y = \log(salary)$, $x_1 = \log(sales)$, $x_2 = ceoten$ y $x_3 = ceoten^2$. Como se sabe, por el capítulo 2, el parámetro β_1 es la *elasticidad* (*ceteris paribus*) del sueldo (*salary*) respecto a las ventas (*sales*). Si $\beta_3 = 0$, entonces $100\beta_2$ es aproximadamente el incremento porcentual *ceteris paribus* de *salary* cuando *ceoten* aumenta en un año. Cuando $\beta_3 \neq 0$, el efecto de *ceoten* sobre *salary* es más complicado. El tratamiento más general de modelos con términos cuadráticos se pospondrá hasta el capítulo 6.

La ecuación (3.7) proporciona un aviso importante acerca del análisis de regresión múltiple. La palabra “lineal” en el modelo de regresión lineal múltiple significa que la ecuación (3.6) es lineal en los *parámetros*, β_j. La ecuación (3.7) es un ejemplo de modelo de regresión múltiple que, aunque lineal en las β_j, es una relación no lineal entre *salary* y las variables *sales* y *ceoten*. En muchas aplicaciones de la regresión lineal múltiple hay relaciones no lineales entre las variables subyacentes.

El supuesto clave en el modelo general de regresión múltiple se establece con facilidad en términos de una esperanza condicional:

$$E(u|x_1, x_2, \ldots, x_k) = 0. \quad \textbf{3.8}$$

Como mínimo, la ecuación (3.8) requiere que ninguno de los factores en el término de error no observado esté correlacionado con las variables explicativas. También significa que se ha entendido de manera correcta la relación funcional entre la variable explicada y las variables explicativas. Cualquier problema que cause que u esté correlacionada con cualquiera de las variables independientes hace que (3.8) no se satisfaga. En la sección 3.3 se mostrará que el supuesto (3.8) implica que los estimadores de MCO son insesgados y se obtendrá el sesgo que surge cuando una variable clave se omite de la ecuación. En los capítulos 15 y 16 se estudiarán otras razones que pueden hacer que (3.8) no se satisfaga y se mostrará qué puede hacerse en los casos en que no se satisface.

3.2 Mecánica e interpretación de los mínimos cuadrados ordinarios

A continuación se resumen algunas características algebraicas y de cálculo del método de mínimos cuadrados ordinarios aplicado a un conjunto de datos determinado. También se analiza cómo interpretar la ecuación estimada.

Obtención de las estimaciones de MCO

Primero se considera la estimación del modelo con dos variables independientes. La ecuación estimada de MCO se escribe de manera similar a la de regresión simple:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \tag{3.9}$$

donde

$\hat{\beta}_0$ = la estimación de β_0.
$\hat{\beta}_1$ = la estimación de β_1.
$\hat{\beta}_2$ = la estimación de β_2.

Pero, ¿cómo se obtienen $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$? El método de **mínimos cuadrados ordinarios** elige las estimaciones que minimizan la suma de los residuales cuadrados. Es decir, dadas n observaciones sobre y, x_1 y x_2, $\{(x_{i1}, x_{i2}, y_i): i = 1, 2, \ldots, n\}$, las estimaciones $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ se eligen de manera simultánea para que

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \tag{3.10}$$

sea tan pequeña como sea posible.

Para entender lo que hacen los MCO, es importante dominar el significado de los índices de las variables independientes en la ecuación (3.10). Aquí las variables independientes tienen dos subíndices, i seguido ya sea de un 1 o de un 2. El subíndice i se refiere al número de la observación. De manera que la suma en la ecuación (3.10) corresponde a todas las observaciones, desde $i = 1$ hasta n. El segundo índice es un método que sirve para distinguir entre las diferentes variables independientes. En el ejemplo en que se relacionan *wage* (salario) con *educ* (educación) y *exper* (experiencia), $x_{i1} = educ_i$ es la educación de la persona i de la muestra, y $x_{i2} = exper_i$ es la experiencia de la persona i. En la ecuación (3.10) la suma de los residuales cuadrados es $\sum_{i=1}^{n} (wage_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. En lo que se presenta a continuación, el subíndice i se reserva como índice del número de la observación. Si se escribe x_{ij}, esto significa la observación i-ésima sobre la variable independiente j-ésima. (Algunos autores prefieren intercambiar el orden del número de observación y el número de variable, de manera que x_{1i} es la observación i de la variable uno. Pero esto es sólo cuestión de gustos respecto a la notación.)

En el caso general, con k variables independientes, se buscan las estimaciones $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_k$ en la ecuación

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.11}$$

Estas $k + 1$ estimaciones de MCO se eligen de manera que minimicen la suma de los residuales cuadrados:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2. \tag{3.12}$$

Este problema de minimización se resuelve empleando cálculo multivariable (vea el apéndice 3A). Esto lleva a $k + 1$ ecuaciones lineales en $k + 1$ incógnitas $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) &= 0.
\end{aligned} \tag{3.13}$$

A estas ecuaciones se les suele llamar las **condiciones de primer orden** de MCO. Como en el caso del modelo de regresión simple de la sección 2.2, las condiciones de primer orden de MCO se obtienen con el método de los momentos: bajo el supuesto (3.8), $\mathrm{E}(u) = 0$ y $\mathrm{E}(x_j u) = 0$, donde $j = 1, 2, \ldots, k$. Las ecuaciones en (3.13) son las contrapartes muestrales de estos momentos poblacionales, aunque se ha omitido la división entre el tamaño n de la muestra.

Resolver a mano las ecuaciones en (3.13) es tedioso aun cuando n y k sean de tamaño moderado. Sin embargo, las computadoras modernas que cuenten con software para estadística y econometría resuelven estas ecuaciones muy rápido, incluso con n y k grandes.

Sólo una pequeña advertencia: se debe suponer que las ecuaciones en (3.13) tienen soluciones *únicas* para las $\hat{\beta}_j$. Por ahora, sólo se supondrá esto, como suele ser el caso en los modelos bien definidos. En la sección 3.3 se establece el supuesto necesario para que existan las estimaciones únicas de MCO (vea el supuesto RLM.3).

Como en el análisis de regresión simple, a la ecuación (3.11) se le llama **línea de regresión de MCO** o **función de regresión muestral (FRM)**. A $\hat{\beta}_0$ se le llama **estimación del intercepto de MCO** y a $\hat{\beta}_1, \ldots, \hat{\beta}_k$, **estimaciones de las pendientes de MCO** (correspondientes a las variables independientes $x_1, x_2, \ldots, x_k$).

Para indicar que se ha realizado una regresión por MCO, se escribe la ecuación (3.11) reemplazando la y y las $x_1, \ldots, x_k$ por los nombres de las variables (por ejemplo, *wage*, *educ* y *exper*), o se dice que "se realizó una regresión por MCO de y sobre $x_1, x_2, \ldots, x_k$" o que "se ha regresado y sobre $x_1, x_2, \ldots, x_k$". Estas son maneras abreviadas de decir que se ha usado el método de mínimos cuadrados ordinarios para obtener la ecuación (3.11) de MCO. A menos que de manera explícita se diga otra cosa, aquí siempre se estimará el intercepto junto con las pendientes.

Interpretación de la ecuación de regresión de MCO

Más importante que los detalles relacionados con el cálculo de los $\hat{\beta}_j$ es la *interpretación* de la ecuación estimada. Para empezar se verá el caso en que se tienen dos variables independientes:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.14}$$

En la ecuación (3.14) el intercepto $\hat{\beta}_0$ es el valor predicho para y cuando $x_1 = 0$ y $x_2 = 0$. Algunas veces es interesante hacer $x_1 = 0$ y $x_2 = 0$, otras no tiene sentido. Sin embargo, el intercepto siempre es necesario para obtener una predicción de y mediante la línea de regresión de MCO, como se indica en la ecuación (3.14).

Las estimaciones $\hat{\beta}_1$ y $\hat{\beta}_2$ se interpretan como **efectos parciales** o **efectos *ceteris paribus***. De acuerdo con la ecuación (3.14), se tiene

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2,$$

de manera que dados los cambios en x_1 y x_2. se puede obtener el cambio predicho para y. (Observe cómo el intercepto no tiene nada que ver con el cambio en y.) En particular, cuando x_2 se mantiene constante, de manera que $\Delta x_2 = 0$, entonces

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1,$$

con x_2 constante. El punto clave es que, al incluir x_2 en el modelo, se obtiene un coeficiente para x_1 con una interpretación *ceteris paribus*. A esto se debe que el análisis de regresión múltiple sea tan útil. De manera similar,

$$\Delta\hat{y} = \hat{\beta}_2\Delta x_2,$$

x_1 constante.

Ejemplo 3.1

[Determinantes del promedio en la universidad]

Las variables en la base de datos GPA1.RAW incluyen el promedio general de calificaciones en la universidad (*colGPA*), el promedio general de calificaciones en el bachillerato (*hsGPA*) y la puntuación en el examen de admisión a la universidad (*ACT*) para una muestra de 141 estudiantes de una universidad grande; los promedios generales de calificaciones tanto del bachillerato como de la universidad se dan en una escala de cuatro puntos. Para predecir el promedio general de calificaciones en la universidad, a partir del promedio general de calificaciones en el bachillerato y de la calificación en el examen de admisión se obtiene la siguiente línea de regresión de MCO:

$$\widehat{colGPA} = 1.29 + .453\ hsGPA + .0094\ ACT. \tag{3.15}$$

¿Cómo se interpreta esta ecuación? Primero, el intercepto 1.29 es la predicción del promedio general de calificaciones en la universidad si *hsGPA y ACT* son ambos cero. Dado que ninguna persona que asista a la universidad tiene cero como promedio general de calificaciones de bachillerato ni cero en el examen de admisión a la universidad, el intercepto, en este caso, no tiene en sí ningún significado.

Estimaciones más interesantes son las de los coeficientes de pendiente de *hsGPA* y *ACT*. Como era de esperarse, existe una relación parcial positiva entre *colGPA* y *hsGPA*: con *ACT* constante, cada punto más en *hsGPA* se relaciona con .453 adicional en el promedio general de la universidad, es decir, casi medio punto. En otras palabras, si se eligen dos estudiantes, A y B, y éstos tienen la misma puntuación en el examen de admisión (ACT), pero el promedio general en el bachillerato del estudiante A es un punto superior al del estudiante B, entonces se predice que en la universidad el estudiante A tendrá un promedio general de calificaciones .453 más alto que el estudiante B. (Esto no dice nada acerca de dos personas reales, es sólo la mejor predicción.)

El signo de *ACT* implica que, si *hsGPA* permanece constante, un cambio de 10 puntos en el examen de admisión (ACT) —un cambio muy grande, ya que en la muestra la puntuación promedio es de 24 con una desviación estándar menor a tres— tendrá un efecto sobre *colGPA* de menos de una décima de punto. Este es un efecto pequeño que indica que, una vez que se ha tomado en cuenta el promedio general del bachillerato, la puntuación en el examen de admisión (ACT) no es un fuerte predictor del promedio general en la universidad. (Naturalmente, hay muchos otros factores que contribuyen al promedio general de calificaciones en la universidad, pero aquí nos concentramos en los estadísticos disponibles para los estudiantes de bachillerato.) Más adelante, después de que se analice la inferencia estadística, se mostrará que el coeficiente de ACT no sólo es pequeño para fines prácticos, sino que es estadísticamente insignificante.

Centrándose en el análisis de regresión simple que relaciona *colGPA* sólo con *ACT* se obtiene

$$\widehat{colGPA} = 2.40 + .0271\ ACT;$$

de manera que el coeficiente de *ACT* es casi el triple del estimado en (3.15). Pero esta ecuación *no* permite comparar dos personas con el mismo promedio general en el bachillerato; esta ecuación corresponde a otro experimento. Después se comentará más acerca de las diferencias entre la regresión múltiple y la simple.

El caso con más de dos variables independientes es similar. La línea de regresión de MCO es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.16}$$

Expresada en términos de cambios,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k. \tag{3.17}$$

El coeficiente de x_1 mide el cambio en $\hat{y}$ por un aumento de x_1 de una unidad, manteniendo constantes todas las demás variables independientes. Es decir,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, \tag{3.18}$$

manteniendo constantes $x_2, x_3, \ldots, x_k$. Por tanto, las variables $x_2, x_3, \ldots, x_k$ han sido *controladas* al estimar el efecto de x_1 sobre y. Los demás coeficientes tienen una interpretación similar.

A continuación se presenta un ejemplo con tres variables independientes.

Ejemplo 3.2

[Ecuación para el salario por hora]

Empleando las 526 observaciones sobre trabajadores en la base de datos WAGE1.RAW, las variables *educ* (años de educación), *exper* (años de experiencia en el mercado laboral) y *tenure* (años de antigüedad en el empleo actual) se incluyen en una ecuación para explicar log(*wage*). La ecuación estimada es

$$\widehat{\log(wage)} = .284 + .092\ educ + .0041\ exper + .022\ tenure. \tag{3.19}$$

Como en el caso de la regresión simple, los coeficientes tienen una interpretación porcentual. Aquí la única diferencia es que también tienen una interpretación *ceteris paribus*. El coeficiente .092 significa que manteniendo *exper* y *tenure* constantes, se predice que un año más de educación incrementa log(*wage*) en .092, lo que se traduce en un aumento aproximado de 9.2% [100(.092)] en *wage*. Es decir, si se toman dos personas con los mismos niveles de experiencia y antigüedad laboral, el coeficiente de *educ* es la diferencia proporcional con el salario predicho cuando en sus niveles de educación hay una diferencia de un año.

Esta medida del rendimiento de la educación al menos mantiene constantes dos factores importantes de la productividad; para saber si ésta es un buen estimado del rendimiento *ceteris paribus* de un año más de educación es necesario estudiar las propiedades estadísticas de los estimadores de MCO (vea la sección 3.3).

El significado de "mantener todos los demás factores constantes" en la regresión múltiple

La interpretación del efecto parcial de los coeficientes de pendiente en el análisis de regresión múltiple puede causar cierta confusión, por lo que a continuación se presenta un análisis más amplio.

En el ejemplo 3.1, se observó que el coeficiente de *ACT* mide la diferencia que se predice para *colGPA* cuando *hsGPA* se mantiene constante. El poder del análisis de regresión múltiple es que proporciona esta interpretación *ceteris paribus* incluso cuando los datos *no* hayan sido recolectados de manera *ceteris paribus*. Al darle al coeficiente de *ACT* una interpretación de efecto parcial, puede parecer que se salió y se muestrearon personas con el mismo promedio general en el bachillerato pero con puntuaciones diferentes en el examen de admisión (ACT). Este no es el caso. Los datos son una muestra aleatoria tomada de una universidad grande: para obtener los datos no se pusieron restricciones sobre los valores muestrales de *hsGPA* o de *ACT*. Es muy raro que al obtener una muestra pueda uno darse el lujo de mantener constantes ciertas variables. *Si* se pudiera obtener una muestra de individuos con un mismo promedio general en el bachillerato, entonces se podría realizar un análisis de regresión simple relacionando *colGPA* con *ACT*. La regresión múltiple permite imitar esta situación sin restringir los valores de ninguna de las variables independientes.

El poder del análisis de regresión múltiple es que permite hacer en un ambiente no experimental, lo que en las ciencias naturales puede hacerse con experimentos controlados de laboratorio: mantener constantes otros factores.

Cambiar de manera simultánea más de una variable independiente

Algunas veces se desea cambiar a la vez más de una variable independiente para determinar el efecto resultante sobre la variable dependiente. Esto es fácil de hacer usando la ecuación (3.17). Por ejemplo, en la ecuación (3.19) se puede obtener el efecto estimado sobre *wage* cuando una persona permanece un año más en una misma empresa: tanto *exper* (experiencia general en la fuerza laboral) como *tenure* (antigüedad en el empleo actual) aumentan en un año. El efecto total (manteniendo *educ* constante) es

$$\Delta\widehat{\log(wage)} = .0041\ \Delta exper + .022\ \Delta tenure = .0041 + .022 = .0261,$$

es decir, aproximadamente 2.6%. Dado que tanto *exper* como *tenure* aumentan un año, simplemente se suman los coeficientes de *exper* y *tenure* y se multiplica por 100 para convertir el efecto en un porcentaje.

Valores ajustados y residuales de MCO

Después de obtener la línea de regresión de MCO (3.11), para cada observación puede obtenerse el *valor ajustado* o *valor predicho*. Para la observación i, el valor ajustado es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}, \tag{3.20}$$

que es el valor predicho que se obtiene sustituyendo en la ecuación (3.11) los valores de las variables independientes de la observación i. Al obtener los valores ajustados, no debe olvidarse el intercepto; de no ser así, el resultado puede ser erróneo. Por ejemplo, si en la ecuación (3.15) $hsGPA_i = 3.5$ y $ACT_i = 24$, $\widehat{colGPA}_i = 1.29 + .453(3.5) + .0094(24) = 3.101$ (redondeado a tres cifras decimales).

En general, el valor real de y_i para cualquier observación i no será igual al valor predicho, $\hat{y}_i$: el método de MCO minimiza el *promedio* del error de predicción cuadrado, lo cual no dice nada acerca del error de predicción de cualquier observación particular. El **residual** de la observación i está definido como en el caso de la regresión simple,

$$\hat{u}_i = y_i - \hat{y}_i. \tag{3.21}$$

Hay un residual para cada observación. Si $\hat{u}_i > 0$, entonces $\hat{y}_i$ es menor que y_i, lo que significa que, para esta observación, el valor predicho para y_i es menor al valor de y_i. Si $\hat{u}_i < 0$, entonces $y_i < \hat{y}_i$ y el valor predicho para y_i es mayor al valor de y_i.

Los valores ajustados y los residuales de MCO tienen algunas propiedades que son extensiones inmediatas del caso de una variable:

1. El promedio muestral de los residuales es cero y de esta manera $\bar{y} = \bar{\hat{y}}$.
2. La covarianza muestral entre cada una de las variables independientes y los residuales de MCO es cero. Por consiguiente, la covarianza muestral entre los valores ajustados de MCO y los residuales de MCO es cero.
3. El punto $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ se encuentra siempre sobre la línea de regresión de MCO: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k$.

Pregunta 3.2

En el ejemplo 3.1, la línea ajustada de MCO que explica *colGPA* en términos de *hsGPA* y de la puntuación en *ACT* es

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT.$$

Si el promedio de *hsGPA* es 3.4 y el promedio *ACT* es aproximadamente 24.2, ¿cuál es el promedio muestral de *colGPA*?

Las dos primeras propiedades son consecuencia inmediata del conjunto de ecuaciones empleadas para obtener las estimaciones de MCO. La primera ecuación en (3.13) dice que la suma de los residuales es cero. Las ecuaciones restantes son de la forma $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, lo que significa que la covarianza entre cada variable independiente y $\hat{u}_i$ es cero. La propiedad (3) es consecuencia inmediata de la propiedad (1).

Una interpretación de descuento de efectos parciales de la regresión múltiple

Cuando se aplica el método de MCO, no es necesario conocer las fórmulas explícitas para obtener los $\hat{\beta}_j$ que resuelven el sistema de ecuaciones dado en (3.13). Sin embargo, para ciertas deducciones, sí se necesitan las fórmulas explícitas para los $\hat{\beta}_j$. Estas fórmulas aclaran, además, el funcionamiento de los MCO.

Considérese nuevamente el caso con $k = 2$ variables independientes, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2$. Para mayor concreción, la atención se concentrará en $\hat{\beta}_1$. Una manera de expresar $\hat{\beta}_1$ es

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1}y_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right), \tag{3.22}$$

donde los $\hat{r}_{i1}$ son los residuales de MCO de una regresión simple de x_1 sobre x_2, usando la muestra presente. Se regresa la primera variable independiente, x_1, sobre la segunda variable independiente, x_2, y después se obtienen los residuales (aquí y no interviene). La ecuación (3.22) muestra que después se puede hacer una regresión simple de y sobre $\hat{r}_1$ para obtener $\hat{\beta}_1$. (Observe que los residuales $\hat{r}_{i1}$ tienen media muestral cero y de esta manera $\hat{\beta}_1$ es la estimación habitual de la pendiente en la regresión simple.)

La representación de la ecuación (3.22) da otra demostración de la interpretación del efecto parcial de $\hat{\beta}_1$. Los residuales $\hat{r}_{i1}$ son la parte de x_{i1} que no está correlacionada con x_{i2}. Otra manera de decir esto es que $\hat{r}_{i1}$ es x_{i1} después de que los efectos parciales de x_{i2} han sido descontados o deducidos. De esta manera, $\hat{\beta}_1$ mide la relación muestral entre y y x_1 después descontar los efectos parciales de x_2.

En el análisis de regresión simple no se descuentan los efectos parciales de otras variables porque en la regresión no se incluyen otras variables. El ejercicio de computadora C3.5 lo llevará a través del proceso de deducción de los efectos parciales empleando los datos de salario del ejemplo 3.2. Para fines prácticos, lo importante es que en la ecuación $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, $\hat{\beta}_1$ mide el cambio en y por cada incremento de una unidad en x_1, manteniendo x_2 constante.

En el modelo general con k variables explicativas, $\hat{\beta}_1$ también puede escribirse como en la ecuación (3.22), pero los residuales $\hat{r}_{i1}$ provienen de la regresión de x_1 sobre $x_2, \ldots, x_k$. Por tanto, $\hat{\beta}_1$ mide el efecto de x_1 sobre y después de que los efectos parciales de $x_2, \ldots, x_k$ han sido descontados o deducidos.

Comparación entre las estimaciones de la regresión simple y de la regresión múltiple

Existen dos casos especiales en los que la regresión simple de y sobre x_1 produce la *misma* estimación de MCO para el coeficiente de x_1 que la regresión de y sobre x_1 y x_2. Para ser más precisos, escriba la regresión simple de y sobre x_1 como $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, y la regresión múltiple como $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Se sabe que, en general, el coeficiente $\tilde{\beta}_1$ de la regresión simple no es igual al coeficiente $\hat{\beta}_1$ de la regresión múltiple. Resulta que existe una relación sencilla entre $\tilde{\beta}_1$ y $\hat{\beta}_1$, que permite hacer interesantes comparaciones entre la regresión simple y la múltiple:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1, \tag{3.23}$$

donde $\tilde{\delta}_1$ es el coeficiente de pendiente de la regresión simple de x_{i2} sobre x_{i1}, $i = 1, \ldots, n$. Esta ecuación muestra cómo difiere $\tilde{\beta}_1$ del efecto parcial de x_1 sobre $\hat{y}$. El término de confusión es el efecto parcial de x_2 sobre $\hat{y}$ multiplicado por la pendiente de la regresión muestral de x_2 sobre x_1 (vea una verificación más general en la sección 3A.4 del apéndice de este capítulo).

La relación entre $\tilde{\beta}_1$ y $\hat{\beta}_1$ muestra, además, que hay dos casos en los que $\tilde{\beta}_1$ y $\hat{\beta}_1$ son iguales:

1. El efecto parcial de x_2 sobre $\hat{y}$ es cero en la muestra. Es decir, $\hat{\beta}_2 = 0$.
2. x_1 y x_2 no están correlacionadas en la muestra. Es decir, $\tilde{\delta}_1 = 0$.

Aun cuando las estimaciones de regresión simple y de regresión múltiple casi nunca son idénticas, la fórmula anterior puede emplearse para identificar por qué estas estimaciones pueden ser o muy diferentes o bastante similares. Por ejemplo, si $\hat{\beta}_2$ es pequeño, puede esperarse que las estimaciones de β_1 en la regresión simple y en la regresión múltiple sean parecidos. En el ejemplo 3.1, la correlación muestral entre *hsGPA* y *ACT* es aproximadamente .346, que es una correlación no trivial. Pero el coeficiente de *ACT* es bastante pequeño. No extrañará encontrar que la regresión simple de *colGPA* sobre *hsGPA* produzca una estimación para la pendiente de .482, que no difiere mucho de la estimación .453 en (3.15).

Ejemplo 3.3

[Participación en los planes de pensión 401(k)]

Se usará la base de datos 401K.RAW para estimar el efecto de la tasa de aportación de la empresa al plan (*mrate*) sobre la tasa de participación (*prate*) en ese plan de pensión 401(k). La tasa de aportación es el monto que la empresa aporta al plan de ahorro de un trabajador por cada dólar que éste invierte (hasta un cierto límite); así *mrate* = .75 significa que la empresa aporta 75¢ por cada dólar que invierte el trabajador. La tasa de participación es el porcentaje de trabajadores con derecho al plan que en efecto están inscritos en un plan de pensión 401(k). La variable *age* es la antigüedad del plan 401(k). En esta base de datos hay 1,534 planes, el *prate* promedio es 87.36, el *mrate* promedio es .732 y la *age* promedio es 13.2.

Regresando *prate* sobre *mrate* y *age* se obtiene

$$\widehat{prate} = 80.12 + 5.52\, mrate + .243\, age.$$

De manera que tanto *mrate* como *age* tienen el efecto esperado. ¿Qué ocurre si no se controla *age*? El efecto estimado de *age* no es trivial y, por tanto, puede esperarse una variación grande en el efecto estimado de *mrate* si se elimina *age* de la regresión. Sin embargo, la regresión simple de *prate* sobre *mrate* da $\widehat{prate} = 83.08 + 5.86\, mrate$. En la regresión simple, el efecto estimado de *mrate* sobre *prate* es claramente diferente del que se obtiene en la regresión múltiple, pero la diferencia no es muy grande. (La estimación en la regresión simple es apenas cerca de 6.2% mayor que la estimación en la regresión múltiple.) Esto puede explicarse por el hecho de que la correlación muestral entre *mrate* y *age* es sólo .12.

En el caso con k variables independientes, la regresión simple de y sobre x_1 y la regresión múltiple de y sobre $x_1, x_2, \ldots, x_k$ se obtiene una estimación idéntica de x_1 sólo si 1) todos los coeficientes de MCO, desde x_2 hasta x_k son cero *o si* 2) x_1 *no* está correlacionada con *ninguna* de las $x_2, \ldots, x_k$. En la práctica ninguna de estas dos cosas es muy probable. Pero si los coeficientes de x_2 a x_k son pequeños o si las correlaciones muestrales entre x_1 y las demás variables independientes son insignificantes, entonces las estimaciones para el efecto de x_1 sobre y obtenidos mediante regresión simple y regresión múltiple pueden ser similares.

Bondad de ajuste

Como en el caso de la regresión simple, se definen la **suma total de cuadrados (STC)**, la **suma explicada de cuadrados (SEC)** y la **suma residual de cuadrados** o **suma de residuales cuadrados (SRC)** como

$$\text{STC} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \textbf{3.24}$$

$$\text{SEC} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \textbf{3.25}$$

$$\text{SRC} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \qquad \textbf{3.26}$$

Empleando el mismo argumento que en el caso de la regresión simple, se puede mostrar que

$$\text{STC} = \text{SEC} + \text{SRC}. \qquad \textbf{3.27}$$

Es decir, la variación total en $\{y_i\}$ es la suma de la variación total en $\{\hat{y}_i\}$ y en $\{\hat{u}_i\}$.

Suponiendo que la variación total en y no sea cero, como es el caso, a menos que las y_i sean constantes en la muestra, se puede dividir (3.27) entre STC para obtener

$$\text{SRC/STC} + \text{SEC/STC} = 1.$$

Como en el caso de la regresión simple, la R-cuadrada se define como

$$R^2 \equiv \text{SEC/STC} = 1 - \text{SRC/STC}, \tag{3.28}$$

y se interpreta como la proporción de la variación muestral en y_i que es explicada por la línea de regresión de MCO. Por definición, R^2 es un número entre cero y uno.

También se puede mostrar que R^2 es igual al cuadrado del coeficiente de correlación entre las y_i reales y los valores ajustados $\hat{y}_i$. Es decir,

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right)}. \tag{3.29}$$

[En (3.29) se ha empleado del promedio de las $\hat{y}_i$ para ser fiel a la fórmula del coeficiente de correlación; se sabe que este promedio es igual a $\bar{y}$ debido a que el promedio muestral de los residuales es cero y $y_i = \hat{y}_i + \hat{u}_i$.]

Un hecho importante acerca de R^2 es que nunca disminuye y, en general, aumenta cuando se agrega otra variable independiente a la regresión. Este hecho algebraico se debe a que, por definición, la suma de los residuales cuadrados nunca aumenta cuando se añaden regresores al modelo. Por ejemplo, el último dígito del número de seguridad social de una persona no tiene nada que ver con su salario por hora, pero añadir ese dígito a una ecuación de salario hará que aumente R^2 (un poco, por lo menos).

El hecho de que la R^2 nunca disminuya cuando se agrega *cualquier* variable a la regresión hace de R^2 un instrumento poco confiable para decidir si agregar una o varias variables al modelo. El factor que debe determinar si una variable explicativa pertenece a un modelo es si ésta tiene un efecto parcial distinto de cero sobre y en la *población*. En el capítulo 4, cuando se estudie la inferencia estadística, se mostrará cómo probar esta hipótesis. Se verá también que, usada de manera adecuada, la R^2 permite *probar* si un grupo de variables es importante para explicar y. Por ahora, la R^2 se usará como una medida de la bondad de ajuste de un modelo determinado.

Ejemplo 3.4

[Determinantes de las calificaciones promedio de universidad]

De la regresión hecha antes con los promedios de calificaciones se obtuvo la ecuación y la R^2 siguientes

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$
$$n = 141,\ R^2 = .176.$$

Esto significa que en esta muestra de estudiantes *hsGPA* y *ACT* juntos, explican de manera aproximada 17.6% de la variación en el promedio de calificaciones en la universidad. Puede que esto no parezca ser un porcentaje muy alto, pero se debe tener en cuenta que hay muchos otros factores —antecedentes familiares, personalidad, calidad de la educación en el bachillerato, afinidad con la universidad— que contribuyen al desempeño de un estudiante en la universidad. Si *hsGPA* y *ACT* explicaran casi toda la variación en *colGPA* entonces, ¡el desempeño en la universidad estaría determinado por el desempeño en el bachillerato!

Ejemplo 3.5

[Explicación de los historiales de arrestos]

La base de datos CRIME1.RAW contiene información sobre arrestos durante 1986, y otros datos, sobre 2,725 hombres nacidos en California en 1960 o 1961. Cada hombre de la muestra fue arrestado al menos una vez antes de 1986. La variable *narr86* indica el número de veces que un hombre fue arrestado durante 1986: esta variable es cero para la mayoría de los hombres de la muestra (72.29%) y varía desde 0 hasta 12 (El porcentaje de hombres detenidos una sola vez durante 1986 es 20.51). La variable *pcnv* es la proporción (no el porcentaje) de detenciones anteriores a 1986 que condujeron a una condena, *avgsen* es la duración promedio de las condenas anteriores cumplidas (cero para la mayoría de las personas), *ptime86* es el tiempo en meses pasado en prisión durante 1986 y *qemp86* es la cantidad de trimestres que la persona tuvo empleo en 1986 (de cero a cuatro).

Un modelo lineal para explicar el número de arrestos es

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 ptime86 + \beta_4 qemp86 + u,$$

donde *pcnv* es una variable proxy para representar la probabilidad de ser condenado por un delito y *avgsen* es una medida de la severidad esperada del castigo, si es condenado. La variable *ptime86* capta el efecto del encarcelamiento sobre la delincuencia: si un individuo está en prisión, no puede ser detenido por un delito fuera de la misma. Las oportunidades en el mercado laboral son captadas en *qemp86*.

Primero se estima el modelo sin la variable *avgsen*. Se obtiene

$$\widehat{narr86} = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$
$$n = 2{,}725,\ R^2 = .0413.$$

Esta ecuación indica que, juntas, estas tres variables *pcnv*, *ptime86* y *qemp86* explican de manera aproximada 4.1% de la variación en *narr86*.

Todos los coeficientes de pendiente de MCO tienen el signo esperado. El aumento en la proporción de condenas disminuye el número predicho de detenciones. Si se aumenta *pcnv* .50 (un aumento grande en la probabilidad de ser condenado), entonces, manteniendo todos los demás factores constantes, $\Delta\widehat{narr86} = -.150(.50) = -.075$. Esto puede parecer extraño porque la variación en una detención no puede ser una cantidad fraccionaria. Pero este valor puede usarse para predecir la variación esperada en la cantidad de detenciones de un grupo grande de hombres. Por ejemplo, dados 100 hombres, la disminución predicha en las detenciones cuando *pcnv* aumenta .50 es -7.5.

De manera similar, una estancia más larga en prisión conduce a que el número de detenciones predichas sea menor. En efecto, si *ptime86* aumenta de 0 a 12, las detenciones predichas para un determinado hombre disminuyen .034(12) = .408. Cada trimestre más de empleo legal hace que las detenciones que se predicen disminuyan .104, lo que corresponde a 10.4 detenciones en 100 hombres.

Se sabe que si se añade *avgsen* al modelo, aumentará la R^2. La ecuación estimada es

$$\widehat{narr86} = .707 - .151\, pcnv + .0074\, avgsen - .037\, ptime86 - .103\, qemp86$$
$$n = 2{,}725,\ R^2 = .0422.$$

De manera que, al agregar la variable de la condena promedio la R^2 aumenta de .0413 a .0422, un efecto en realidad muy pequeño. También el signo del coeficiente de *avgsen* es inesperado: dice que cuanto más larga en promedio sea la condena aumenta la actividad delictiva.

El ejemplo 3.5 merece una advertencia final. El hecho de que las cuatro variables explicativas de la segunda regresión expliquen sólo 4.2% de la variación en *narr86* no necesariamente significa que esta ecuación no sea útil. Aun cuando estas cuatro variables juntas no expliquen gran parte de la variación en los arrestos, es posible, sin embargo, que las estimaciones de MCO sean estimaciones confiables del efecto *ceteris paribus* de cada variable independiente sobre *narr86*. Se verá más adelante que este sea o no el caso no depende de manera directa de la magnitud de la R^2. En general, que la R^2 sea pequeña indica que es difícil predecir con mucha exactitud resultados individuales para y, tema que se estudiará en el capítulo 6. En el ejemplo sobre las detenciones, esa R^2 tan pequeña refleja lo que se sospecha en las ciencias sociales: en general, es muy difícil predecir el comportamiento de los individuos.

Regresión a través del origen

Hay ocasiones en las que una teoría económica o el sentido común hacen pensar que β_0 debe ser cero y, por esta razón, habrá que mencionar brevemente la estimación de MCO cuando el intercepto es cero. En concreto, se busca una ecuación de la forma

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k, \tag{3.30}$$

donde el símbolo "~" sobre los coeficientes estimados se emplea para distinguirlos de los coeficientes estimados de MCO obtenidos tomando en cuenta el intercepto [como en (3.11)]. En (3.30), cuando $x_1 = 0$, $x_2 = 0$, …, $x_k = 0$, el valor predicho es cero. En este caso se dice que $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ son los coeficientes estimados de MCO para la regresión *a través del origen* de y sobre $x_1, x_2, \ldots, x_k$.

Las estimaciones de MCO en (3.30), como siempre, minimizan la suma de los residuales cuadrados, pero con el intercepto igualado a cero. Hay que hacer la advertencia de que en la regresión a través del origen no se tienen las propiedades de los MCO obtenidas antes. En particular, los residuales de MCO ya no tienen media muestral cero. Además, si R^2 se define como $1 - \text{SRC/STC}$, donde STC está dada en (3.24) y SRC es ahora $\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik})^2$, entonces R^2 puede ser negativa. Esto significa que la media muestral, $\bar{y}$, "explica" más de la variación en las y_i que las variables explicativas. Se debe incluir un intercepto en la regresión o concluir que las variables explicativas no explican bien y. Para tener siempre una R-cuadrada no negativa, algunos economistas prefieren calcular R^2 como el cuadrado del coeficiente de correlación entre los valores reales y los ajustados de y, como en (3.29). (En este caso, el valor ajustado promedio debe calcularse directamente dado que no será igual a $\bar{y}$.) Sin embargo, no hay una regla fija para calcular R-cuadrada en la regresión a través del origen.

Una seria desventaja de la regresión a través del origen es que, si en el modelo poblacional el intercepto β_0 es diferente de cero, entonces los estimadores de MCO de los parámetros de la pendiente estarán sesgados. En algunos casos este sesgo puede ser severo. El costo de estimar el intercepto cuando β_0 es realmente cero es que las varianzas de los estimadores de pendiente de MCO son mayores.

3.3 Valor esperado de los estimadores de MCO

Ahora se verán las propiedades estadísticas del método de MCO para estimar los parámetros del modelo poblacional. En esta sección se obtienen los valores esperados de los estimadores de MCO. En particular, se establecen y se analizan cuatro supuestos, que son extensiones directas de los supuestos del modelo de regresión simple, bajo el cual los estimadores de MCO de los parámetros poblacionales son insesgados. Se obtiene también de manera explícita el sesgo de MCO cuando se omite una variable importante para la regresión.

Hay que recordar que las propiedades estadísticas no tienen nada que ver con la muestra de que se trate, sino con la propiedad de los estimadores cuando el muestreo aleatorio se hace repetidas veces. Así, las secciones 3.3, 3.4 y 3.5 son un poco abstractas. Aunque se dan ejemplos de la obtención del sesgo en modelos específicos, no tiene sentido hablar de las propiedades estadísticas de un conjunto de estimaciones obtenidas de una sola muestra.

El primer supuesto sólo define el modelo de regresión lineal múltiple (RLM).

Supuesto RLM.1 (Lineal en los parámetros)

El modelo poblacional puede expresarse como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u, \tag{3.31}$$

donde $\beta_0, \beta_1, \dots, \beta_k$ son los parámetros (constantes) desconocidos de interés y u es un error aleatorio o término de perturbación no observable.

La ecuación (3.31) expresa formalmente el **modelo poblacional**, llamado algunas veces el **modelo verdadero**, para permitir la posibilidad de estimar un modelo que difiera de (3.31). La característica clave es que este modelo es lineal en los parámetros $\beta_0, \beta_1, \dots, \beta_k$. Como se sabe, (3.31) es bastante flexible porque tanto y como las variables independientes pueden ser funciones arbitrarias de las variables de interés, tales como logaritmos naturales y cuadrados [vea, por ejemplo, la ecuación (3.7)].

Supuesto RLM.2 (Muestreo aleatorio)

Se tiene una muestra aleatoria de n observaciones, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i): i = 1, 2, \dots, n\}$, que sigue el modelo poblacional del supuesto RLM.1.

Algunas veces se necesita dar la ecuación de una determinada observación i: dada una observación obtenida de manera aleatoria de la población, se tiene

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i. \tag{3.32}$$

Recuerde que i se refiere a una observación y que el segundo subíndice de las x es el número de la variable. Por ejemplo, se puede escribir la ecuación del sueldo del director general o CEO para un determinado CEO i como

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + u_i. \tag{3.33}$$

El término u_i contiene los factores no observados del CEO i que afectan su sueldo. Para las aplicaciones, suele ser más fácil dar el modelo en forma poblacional, como en (3.31). Este modelo contiene menos desorden y hace énfasis en el hecho de que interesa estimar una relación poblacional.

A la luz del modelo (3.31), los estimadores de MCO $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ de la regresión de y sobre $x_1, \ldots, x_k$ se consideran como estimadores de $\beta_0, \beta_1, \ldots, \beta_k$. En la sección 3.2 se vio que, dada una muestra, MCO elige las estimaciones de intercepto y de las pendientes de manera que el promedio de los residuales sea cero y que la correlación muestral entre cada variable independiente y los residuales sea cero. Sin embargo, no se han dado las condiciones bajo las cuales, dada una muestra, las estimaciones de MCO están bien definidas. El supuesto siguiente llena esta brecha.

Supuesto RLM.3 (No hay colinealidad perfecta)

En la muestra (y por tanto en la población), ninguna de las variables independientes es constante y no hay ninguna relación *lineal exacta* entre las variables independientes.

El supuesto RLM.3 es más complicado que su contraparte para la regresión simple, porque ahora hay que considerar la relación entre todas las variables independientes. Si una variable independiente en (3.31) es una combinación lineal exacta de las otras variables independientes, entonces se dice que el modelo sufre de **colinealidad perfecta** y que no puede ser estimado por el método de MCO.

Es importante observar que el supuesto RLM.3 *sí* permite que las variables independientes estén correlacionadas; lo único que no permite es que estén *perfectamente* correlacionadas. Si no se permitiera ninguna correlación entre las variables independientes, entonces la regresión múltiple sería de muy poca utilidad para el análisis econométrico. Por ejemplo, en el modelo en el que se relacionan las puntuaciones de exámenes con los gastos en educación y el ingreso familiar promedio,

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

se espera que *expend* y *avginc* estén correlacionados: los distritos escolares en los que el ingreso familiar promedio es alto tienden a gastar más en educación por estudiante. De hecho, la principal motivación para incluir *avginc* en la ecuación es que se sospecha que está relacionado con *expend*, y por esto se desea mantenerlo constante en el análisis. El supuesto RLM.3 sólo descarta la correlación *perfecta*, en nuestra muestra, entre *expend* y *avginc*. Sería muy mala suerte obtener una muestra en la que los gastos por estudiante estuvieran correlacionados de manera perfecta con el ingreso familiar promedio. Pero una cierta correlación, quizá en una cantidad importante es esperada y en efecto permitida.

El caso más sencillo en que dos variables independientes pueden estar correlacionadas de manera perfecta es aquel en el que una variable sea un múltiplo constante de otra. Esto puede ocurrir cuando el investigador, en forma inadvertida, coloca en una ecuación de regresión la misma variable medida en diferentes unidades. Por ejemplo, al estimar la relación entre consumo e ingreso no tiene sentido incluir como variables independientes ingreso medido en dólares e ingreso medido en miles de dólares. Una de estas dos variables es redundante. ¿Qué sentido tendría mantener constante el ingreso medido en dólares y variar el ingreso medido en miles de dólares?

Como se sabe, entre los regresores *puede* haber diferentes funciones lineales de una misma variable. Por ejemplo, el modelo $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ no viola el supuesto RLM.3: aun cuando $x_2 = inc^2$ es una función exacta de $x_1 = inc$, inc^2 no es una función *lineal* exacta de *inc*. Incluir inc^2 en el modelo es una manera útil de generalizar la forma funcional, a diferencia de incluir el ingreso medido en dólares y en miles de dólares.

El sentido común indica no incluir en una misma ecuación de regresión la misma variable explicativa medida en diferentes unidades. Existen también situaciones más sutiles en las que una variable independiente puede ser múltiplo de otra. Suponga que se desea estimar una extensión de una función de consumo de elasticidad constante. Parecería natural especificar un modelo como el siguiente

$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u, \tag{3.34}$$

donde $x_1 = \log(inc)$ y $x_2 = \log(inc^2)$. Al utilizar las propiedades básicas de los logaritmos naturales (véase apéndice A), $\log(inc^2) = 2\cdot\log(inc)$. Es decir, $x_2 = 2x_1$, y esto es válido para todas las observaciones de la muestra. Esto viola el supuesto RLM.3. En lugar de esto hay que incluir $[\log(inc)]^2$, y no $\log(inc^2)$, junto con $\log(inc)$. Esta es una extensión razonable del modelo de elasticidad constante, y en el capítulo 6 se verá cómo interpretar tales modelos.

Otro caso en el que las variables independientes pueden ser perfectamente colineales es cuando una variable independiente puede expresarse como una función lineal exacta de otras dos o más variables independientes. Por ejemplo, suponga que se quiere estimar el efecto de los gastos de campaña sobre los resultados de la misma. Para simplificar, suponga que en cada elección hay dos candidatos. Sea *voteA* el porcentaje de votos obtenidos por el candidato A, sea *expendA* los gastos de campaña del candidato A, sea *expendB* los gastos de campaña del candidato B y sea *totexpend* el total de los gastos de campaña; las últimas tres variables se dan en dólares. Parece natural especificar un modelo como

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u, \tag{3.35}$$

con objeto de aislar los efectos de los gastos de cada candidato y la cantidad total gastada. Pero este modelo viola el supuesto RLM.3 porque por definición $x_3 = x_1 + x_2$. Al tratar de interpretar esta ecuación de una manera *ceteris paribus* se ve con claridad el problema. Se supone que el parámetro β_1 de la ecuación (3.35) mide el efecto de aumentar un dólar en los gastos del candidato A sobre los votos obtenidos por el candidato A, manteniendo constantes los gastos del candidato B y los gastos totales. Esto no tiene sentido, porque si *expendB* y *totexpend* se mantienen constantes, entonces no se puede incrementar *expendA*.

La solución a la colinealidad perfecta en (3.35) es sencilla: eliminar del modelo una de las tres variables. Tal vez se eliminará *totexpend* y entonces el coeficiente de *expendA* medirá el efecto del aumento en los gastos de A sobre el porcentaje de votos obtenido por A manteniendo fijos los gastos de B.

Los ejemplos anteriores muestran que el supuesto RLM.3 puede no satisfacerse si se descuida especificar el modelo. El supuesto RLM.3 tampoco se satisface si el tamaño de la muestra, *n*, es demasiado pequeño en relación con el número de parámetros que se estiman. En general, en el modelo de regresión de la ecuación (3.31), hay $k + 1$ parámetros y RLM.3 no se satisface si $n < k + 1$. De manera intuitiva, esto es razonable: para estimar $k + 1$ parámetros, se necesitan por lo menos $k + 1$ observaciones. Es claro que es mejor tener tantas observaciones como sea posible, cosa que se notará al ver el cálculo de la varianza en la sección 3.4.

Pregunta 3.3

En el ejemplo anterior, si como variables explicativas se usan *expendA*, *expendB* y *shareA*, donde *shareA* = 100·(*expendA*/*totexpend*) es la participación porcentual de los gastos del candidato A en el total de los gastos de campaña, ¿se viola el supuesto RLM.3?

Si el modelo se ha especificado con cuidado y $n \geq k + 1$, el supuesto RLM.3 puede no satisfacerse en casos raros debido a mala suerte al recolectar la muestra. Por ejemplo, en una ecuación para el salario en que la educación y experiencia sean las variables, es posible que se obtenga una muestra aleatoria en la que cada individuo tenga exactamente el doble de años de educación que años de experiencia. Esta situación hará que el supuesto RLM.3 no se satisfaga, pero esta situación es muy poco probable a menos que se tenga un tamaño de muestra en extremo pequeño.

El último supuesto, y el más importante, para el insesgamiento es una extensión directa del supuesto RLS.4.

Supuesto RLM.4 (Media condicional cero)

El valor esperado del error u, dados los valores de las variables independientes, es cero. En otras palabras

$$\mathrm{E}(u|x_1, x_2, ..., x_k) = 0. \tag{3.36}$$

El supuesto RLM.4 puede no satisfacerse si en la ecuación (3.31) la relación funcional entre las variables explicada y explicativas está mal especificada: por ejemplo, si se olvida incluir el término cuadrático inc^2 en la función de consumo $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ al estimar el modelo. Otra especificación errónea de la forma funcional se presenta cuando se emplea una variable en su nivel original siendo que en el modelo poblacional se emplea el logaritmo de la variable, o viceversa. Por ejemplo, si el verdadero modelo tiene log(*wage*) como variable dependiente, pero en el análisis de regresión se usa *wage* como variable dependiente, entonces los estimadores estarán sesgados. De manera intuitiva esto es bastante claro. En el capítulo 9 se analizarán maneras de detectar formas funcionales mal especificadas.

Omitir un factor importante correlacionado con cualquiera de las $x_1, x_2, \ldots, x_k$ ocasiona también que el supuesto RLM.4 no se satisfaga. En el análisis de regresión múltiple pueden incluirse muchos factores entre las variables explicativas y es menos probable que las variables omitidas sean un problema en comparación con el análisis de regresión simple. De cualquier manera en toda aplicación, hay factores que, debido a las limitaciones de los datos o a ignorancia, no pueden incluirse. Si se cree que estos factores deben controlarse y están correlacionados con una o más de las variables independientes, se violará el supuesto RLM.4. Más adelante se verán estos sesgos.

Hay otras maneras en las que u puede estar correlacionada con una variable explicativa. En el capítulo 15 se analizará el problema del error de medición en una variable explicativa. En el capítulo 16 se verá el problema, conceptualmente más complicado, en el que una o más de las variables explicativas se determina conjuntamente con y. El estudio de estos problemas se pospondrá hasta que se tenga una comprensión más firme del análisis de regresión múltiple bajo un conjunto ideal de supuestos.

Cuando se satisface el supuesto RLM.4 se suele decir que se tienen **variables explicativas exógenas**. Si por alguna razón x_j está correlacionada con u, entonces se dice que x_j es una **variable explicativa endógena**. Los términos "exógena" y "endógena" son originarios del análisis de ecuaciones simultáneas (vea el capítulo 16), pero el término "variable explicativa endógena" ha evolucionado para abarcar cualquier caso en el que una variable explicativa esté correlacionada con el término del error.

Antes de mostrar el insesgamiento de los estimadores de MCO bajo RLM.1 a RLM.4, una advertencia. En econometría los estudiantes que inician suelen confundir el supuesto RLM.3 con el RLM.4, pero son totalmente diferentes. El supuesto RLM.3 descarta ciertas relaciones entre las variables independientes o explicativas y no tiene *nada* que ver con el error, u. Cuando se lleva a cabo una estimación por MCO, se sabe de inmediato si el supuesto RLM.3 se satisface o no. Por otro lado, el supuesto RLM.4 —el más importante de los dos— restringe la relación entre los factores no observables de u y las variables explicativas. Por desgracia, nunca se sabrá con seguridad si el valor promedio de los factores no observables en efecto no está relacionado con las variables explicativas. Pero este es el supuesto crítico.

Ahora ya es posible mostrar el insesgamiento de los estimadores de MCO bajo los primeros cuatro supuestos de la regresión lineal múltiple. Como en el caso de la regresión simple, las esperanzas son condicionales sobre los valores de las variables explicativas en la muestra, lo cual se indica de manera explícita en el apéndice 3A pero no a lo largo del libro.

Teorema 3.1 (Insesgamiento de los estimadores de MCO)

Bajo los supuestos RLM.1 a RLM.4,

$$\mathrm{E}(\hat{\beta}_j) = \beta_j,\ j = 0, 1, \ldots, k, \qquad \boxed{3.37}$$

para cualquier valor del parámetro poblacional β_j. En otras palabras, los estimadores de MCO son estimadores insesgados de los parámetros poblacionales.

En los ejemplos empíricos anteriores, el supuesto RLM.3 se ha satisfecho (porque se han podido calcular las estimaciones de MCO). Además, en su mayoría, las muestras han sido tomadas de manera aleatoria de una población bien definida. Si se cree que los modelos especificados son correctos bajo el supuesto clave RLM.4, entonces se puede concluir que en estos ejemplos el modelo de MCO es insesgado.

Como estamos llegando a un punto en el que se puede usar la regresión múltiple en trabajos empíricos serios, es útil recordar el significado del insesgamiento. Uno se siente tentado, en ejemplos como el de la ecuación del salario en (3.19), a decir algo así como "9.2% es una estimación insesgada del rendimiento de la educación". Como se sabe, una estimación no puede ser insesgada: una estimación es un número fijo, obtenido a partir de una determinada muestra que, por lo general, no es igual al parámetro poblacional. Cuando se dice que los estimadores de MCO son insesgados bajo los supuestos RLM.1 a RLM.4, en realidad se quiere decir que el *procedimiento* mediante el cual se obtienen las estimaciones de MCO es insesgado cuando se le considera aplicado a todas las muestras aleatorias posibles. Se espera haber obtenido una muestra que dé una estimación cercana al valor poblacional pero, por desgracia, esto no puede asegurarse. Lo que se asegura es que no hay razón para creer ni que sea probablemente muy grande ni que sea probablemente muy pequeño.

Inclusión de variables irrelevantes en un modelo de regresión

Un problema que se puede despachar con rapidez es el de la **inclusión de una variable irrelevante** o de la **sobreespecificación del modelo** en el análisis de regresión múltiple. Esto significa que una (o más) de las variables independientes está incluida en el modelo aun cuando en la población no tiene ningún efecto parcial sobre y. (Es decir, su coeficiente poblacional es cero.)

Para ilustrar esto, suponga que se especifica un modelo como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{3.38}$$

y que éste satisface los supuestos RLM.1 a RLM.4. Pero que x_3 no tiene ningún efecto sobre y una vez que x_1 y x_2 se han controlado, lo que significa que $\beta_3 = 0$. La variable x_3 puede estar o no correlacionada con x_1 y x_2; lo único que importa es que, una vez que x_1 y x_2 han sido controladas, x_3 no tiene efecto sobre y. En términos de la esperanza condicional, $\mathrm{E}(y|x_1,x_2,x_3) = \mathrm{E}(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Como no se sabe que $\beta_3 = 0$, nos inclinamos a estimar esta ecuación incluyendo x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3. \tag{3.39}$$

En nuestra regresión se ha incluido la variable irrelevante, x_3. ¿Qué efecto tiene incluir x_3 en (3.39) cuando su coeficiente en el modelo poblacional (3.38) es cero? En términos del insesgamiento de $\hat{\beta}_1$ y $\hat{\beta}_2$, esto no tiene *ningún efecto*. Llegar a esta conclusión no requiere deducción especial, ya que se sigue de inmediato del teorema 3.1. Recuerde, insesgamiento significa $\mathrm{E}(\hat{\beta}_j) = \beta_j$ para *cualquier* valor de β_j, incluyendo $\beta_j = 0$. Por tanto, se puede concluir que $\mathrm{E}(\hat{\beta}_0) = \beta_0$, $\mathrm{E}(\hat{\beta}_1) = \beta_1$, $\mathrm{E}(\hat{\beta}_2) = \beta_2$, y $\mathrm{E}(\hat{\beta}_3) = 0$ (para todos los valores de β_0, β_1 y β_2). Aun cuando $\hat{\beta}_3$ nunca será exactamente cero, su valor promedio sobre todas las muestras aleatorias será cero.

La conclusión que se obtiene del ejemplo anterior es mucho más general: incluir una o más variables irrelevantes en un modelo de regresión múltiple, o sobreespecificar el modelo, no afecta el insesgamiento de los estimadores de MCO. ¿Significa esto que no causa ningún daño incluir variables irrelevantes? No. Como se verá en la sección 3.4, incluir variables irrelevantes puede tener efectos indeseables en la *varianza* de los estimadores de MCO.

Sesgo de variable omitida: caso sencillo

Ahora suponga que, en lugar de incluir una variable irrelevante, se omite una variable que, en realidad, pertenece al verdadero modelo (o al modelo poblacional). Esto se suele conocer como el problema de la **exclusión de una variable relevante** o de **subespecificación del modelo**. Como se dijo en el capítulo 2 y antes en este capítulo, este problema en general hace que los estimadores sean sesgados. Ahora es el momento de mostrar esto de manera explícita y, de igual importancia, de determinar la dirección y la magnitud de este sesgo.

Determinar el sesgo ocasionado por la omisión de una variable importante es un ejemplo de **análisis de error de especificación**. Para empezar se verá el caso en el que el verdadero modelo poblacional tiene dos variables explicativas y un término del error:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.40}$$

y se supondrá que este modelo satisface los supuestos RLM.1 a RLM.4.

Suponga que lo que más interesa es β_1, el efecto parcial de x_1 sobre y. Por ejemplo, y es el salario por hora (o el logaritmo del salario por hora), x_1 es educación y x_2 es una medida de las habilidades innatas del individuo. Para obtener un estimador insesgado de β_1, se *debe* correr la regresión de y sobre x_1 y x_2 (con lo que se obtendrán estimadores insesgados de β_0, β_1 y β_2). Sin embargo, por ignorancia o por falta de datos, el modelo se estima *excluyendo* x_2. En otras palabras, se realiza una regresión simple de y sólo sobre x_1 con lo que se obtiene la ecuación

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.41}$$

Se usa el símbolo "~" en lugar de "^" para hacer hincapié en que $\tilde{\beta}_1$ proviene de un modelo mal especificado.

Cuando se empieza a estudiar el problema de la variable omitida, puede ser difícil distinguir entre el verdadero modelo, en este caso (3.40), y el que en realidad se estima, representado por la ecuación (3.41). Puede parecer descabellado omitir la variable x_2 cuando ésta pertenece al modelo, pero en algunas ocasiones no queda otra posibilidad. Suponga, por ejemplo, que *wage* está determinado por

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \tag{3.42}$$

Como la habilidad innata (*abil*) es no observada, en lugar de este modelo se estima

$$wage = \beta_0 + \beta_1 educ + v,$$

donde $v = \beta_2 abil + u$. Al estimador de β_1 en la regresión simple de *wage* sobre *educ* se le llama $\tilde{\beta}_1$.

A continuación se obtiene el valor esperado de $\tilde{\beta}_1$ condicional en los valores muestrales de x_1 y x_2. No es difícil obtener esta esperanza ya que $\tilde{\beta}_1$ es el estimador de pendiente de MCO de una regresión simple y este estimador ya ha sido estudiado de manera amplia en el capítulo 2. Aquí, la diferencia es que hay que analizar sus propiedades cuando el modelo de regresión simple está mal especificado debido a una variable omitida.

Como se ve, ya casi todo el trabajo para obtener el sesgo del estimador de $\tilde{\beta}_1$ está realizado. De acuerdo con la ecuación (3.23) se tiene la relación algebraica $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$, donde $\hat{\beta}_1$ y $\hat{\beta}_2$ son los estimadores de pendiente (si se tuvieran) de la regresión múltiple

$$y_i \text{ sobre } x_{i1}, x_{i2}, i = 1, \ldots, n \tag{3.43}$$

y $\tilde{\delta}_1$ es la pendiente de la regresión simple

$$x_{i2} \text{ sobre } x_{i1}, i = 1, \ldots, n. \tag{3.44}$$

Como $\tilde{\delta}_1$ sólo depende de las variables independientes de la muestra, al calcular $\mathrm{E}(\tilde{\beta}_1)$ se considera fija (como no aleatoria). Además, puesto que el modelo dado en (3.40) satisface los supuestos RLM.1 a RLM.4, se sabe que $\hat{\beta}_1$ y $\hat{\beta}_2$ son estimadores insesgados de β_1 y β_2, respectivamente. Por tanto,

$$\begin{aligned}\mathrm{E}(\tilde{\beta}_1) &= \mathrm{E}(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \mathrm{E}(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1,\end{aligned} \tag{3.45}$$

lo que implica que el sesgo en $\tilde{\beta}_1$ es

$$\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1. \tag{3.46}$$

Dado que el sesgo en este caso surge por la omisión de la variable explicativa x_2, al término del lado derecho de la ecuación (3.46) suele llamársele **sesgo de variable omitida**.

De acuerdo con la ecuación (3.46), se observa que hay dos casos en los que $\tilde{\beta}_1$ es insesgado. El primero es bastante obvio: cuando $\beta_2 = 0$ —de manera que x_2 no aparece en el verdadero modelo (3.40)— entonces $\tilde{\beta}_1$ es insesgado. Esto ya se sabe por el análisis de regresión simple del capítulo 2. El segundo caso es más interesante. Si $\tilde{\delta}_1 = 0$, entonces $\tilde{\beta}_1$ es un estimador insesgado de β_1, aun cuando $\beta_2 \neq 0$.

Como $\tilde{\delta}_1$ es la covarianza muestral entre x_1 y x_2 sobre la varianza muestral de x_1, $\tilde{\delta}_1 = 0$ si y sólo si x_1 y x_2 no están correlacionadas en la muestra. De esta manera se tiene la importante conclusión de que, si x_1 y x_2 no están correlacionadas en la muestra, entonces $\tilde{\beta}_1$ es insesgado. Esto parece razonable: en la sección 3.2 se mostró que el estimador $\tilde{\beta}_1$ de la regresión simple y el estimador $\hat{\beta}_1$ de la regresión múltiple son iguales cuando x_1 y x_2 no están correlacionadas en la muestra. [También se puede demostrar que $\tilde{\beta}_1$ es insesgado sin condicionar sobre x_{i2} si $\text{E}(x_2|x_1) = \text{E}(x_2)$; entonces estimar β_1, dejando x_2 en el término del error no viola el supuesto de media condicional cero para el error, una vez que se ajuste el intercepto.]

Si x_1 y x_2 están correlacionadas, $\tilde{\delta}_1$ tiene el mismo signo que la correlación entre x_1 y x_2: $\tilde{\delta}_1 > 0$ si x_1 y x_2 están positivamente correlacionadas y $\tilde{\delta}_1 < 0$ si x_1 y x_2 están correlacionadas de manera negativa. El signo del sesgo en $\tilde{\beta}_1$ depende de los signos de β_2 y $\tilde{\delta}_1$ y en la tabla 3.2 se resumen los cuatro posibles casos en los que hay sesgo. La tabla 3.2 se merece un cuidadoso estudio. Por ejemplo, el sesgo en $\tilde{\beta}_1$ es positivo si $\beta_2 > 0$ (x_2 tiene efecto positivo sobre y) y x_1 y x_2 están positivamente correlacionadas, el sesgo es negativo si $\beta_2 > 0$ y x_1 y x_2 están negativamente correlacionadas, etcétera.

En la tabla 3.2 se resumen las direcciones de los sesgos, pero la magnitud del sesgo es también muy importante. Un sesgo pequeño, del signo que sea, no por fuerza es causa de preocupación. Por ejemplo, si en la población el rendimiento de la educación es 8.6% y el sesgo en el estimador de MCO es 0.1% (una décima de un punto porcentual), entonces no hace falta preocuparse mucho. Por otro lado, un sesgo del orden de tres puntos porcentuales sería mucho más serio. La magnitud del sesgo está determinada por las magnitudes de β_2 y $\tilde{\delta}_1$.

En la práctica, como β_2 es un parámetro poblacional desconocido, no se puede saber con certeza si es positivo o negativo. No obstante, en general se tiene una idea bastante buena acerca de la dirección del efecto parcial de x_2 sobre y. Además, aun cuando el signo de la correlación entre

TABLA 3.2

Resumen de los sesgos en $\tilde{\beta}_1$ cuando x_2 es omitida en la ecuación de estimación (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	Sesgo positivo	Sesgo negativo
$\beta_2 < 0$	Sesgo negativo	Sesgo positivo

x_1 y x_2 no se pueda conocer si x_2 no es observada, en muchos casos, se puede hacer una conjetura acerca de si x_1 y x_2 están positiva o negativamente correlacionadas.

En la ecuación del salario (3.42), por definición, más habilidad innata conduce a más productividad y, por tanto, a salarios más altos: $\beta_2 > 0$. Además, hay razones para pensar que *educ* y *abil* estén correlacionadas de forma positiva: en promedio, las personas con más habilidad innata eligen niveles de educación más altos. Por tanto, las estimaciones de MCO de la ecuación de regresión simple $wage = \beta_0 + \beta_1 educ + v$ son *en promedio* demasiado grandes. Esto no significa que la estimación obtenida a partir de la muestra presente sea demasiado grande. Lo único que se puede decir es que si se recolectan muestras aleatorias y de cada una se obtienen las estimaciones de regresión simple, el promedio de estas estimaciones será mayor a β_1.

Ejemplo 3.6

[Ecuación del salario por hora]

Suponga que el modelo $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ ssatisface los supuestos RLM.1 a RLM.4. La base de datos WAGE1.RAW no contiene datos sobre la habilidad, de manera que β_1 se estima mediante la regresión simple

$$\widetilde{\log(wage)} = .584 + .083\ educ$$
$$n = 526,\ R^2 = .186. \tag{3.47}$$

Este es el resultado obtenido con una sola muestra, de manera que no se puede decir que .083 sea mayor que β_1; el verdadero rendimiento de la educación puede ser menor o mayor a 8.3% (esto jamás se sabrá con seguridad). Sin embargo, se sabe que el promedio de las estimaciones de todas las muestras aleatorias será demasiado grande.

Como segundo ejemplo, suponga que, en la escuela primaria, la calificación promedio de los estudiantes en un examen estandarizado se determina mediante

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u, \tag{3.48}$$

donde *expend* es el gasto que se hace por estudiante y *povrate* es el índice de pobreza de los niños de la escuela. Usando datos de distritos escolares, sólo se tienen observaciones sobre el porcentaje de estudiantes con calificación aprobatoria y sobre el gasto que se hace por estudiante; no se tiene información sobre el índice de pobreza. Por lo tanto β_1 se estima mediante la regresión simple de *avgscore* sobre *expend*.

Aquí, también, es posible obtener el posible sesgo de $\tilde{\beta}_1$. Primero, es probable que β_2 sea negativa: existen amplias evidencias de que niños que viven en la pobreza obtienen, en promedio, calificaciones bajas en los exámenes estandarizados. Segundo, el gasto promedio por estudiante es probable que esté correlacionado de manera negativa con el índice de pobreza. Cuanto más alto sea el índice de pobreza, menor será el gasto promedio por estudiante, de manera que $\text{Corr}(x_1, x_2) < 0$. De acuerdo con la tabla 3.2, $\tilde{\beta}_1$ tendrá un sesgo positivo. Esta observación tiene consecuencias importantes. Puede ser que el verdadero efecto del gasto sea cero; es decir, $\beta_1 = 0$. Sin embargo, el estimado mediante regresión simple de β_1 será, por lo general, mayor a cero y esto puede llevar a concluir que el gasto es importante cuando no lo sea.

Cuando se lee o cuando se realizan trabajos empíricos en economía es importante dominar la terminología relacionada con los estimadores sesgados. En el contexto de la omisión de una variable del modelo (3.40), si $E(\tilde{\beta}_1) > \beta_1$,se dice que $\tilde{\beta}_1$ tiene un **sesgo hacia arriba**. Si $E(\tilde{\beta}_1) < \beta_1$, se dice que $\tilde{\beta}_1$ tiene un **sesgo hacia abajo**. En estas definiciones no tiene importancia que β_1 sea positivo o negativo. La frase **sesgado hacia cero** se refiere al caso en el que $E(\tilde{\beta}_1)$ se encuentra más cerca de cero que β_1. Por tanto, si β_1 es positivo, $\tilde{\beta}_1$ es sesgado hacia cero si tiene sesgo hacia abajo. Por otro lado, si $\beta_1 < 0$, entonces $\tilde{\beta}_1$ es sesgado hacia cero si tiene sesgo hacia arriba.

Sesgo de la variable omitida: casos más generales

Obtener el signo del sesgo de la variable omitida cuando hay múltiples regresores en el modelo estimado es más difícil. Hay que recordar que la correlación entre una sola variable explicativa y el error, por lo general, da como resultado que *todos* los estimadores de MCO sean sesgados. Suponga, por ejemplo, que el modelo poblacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{3.49}$$

satisface los supuestos RLM.1 a RLM.4, pero se omite x_3 y que el modelo se estima como

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2. \tag{3.50}$$

Ahora, suponga que x_2 y x_3 no están correlacionadas, pero que x_1 sí está correlacionada con x_3. En otras palabras, x_1 está correlacionada con la variable omitida, pero x_2 no. Se pensará que, mientras es posible que $\tilde{\beta}_1$ esté sesgado, de acuerdo con lo visto en la sección anterior, $\tilde{\beta}_2$ no esté sesgado ya que x_2 no está correlacionada con x_3. Por desgracia, por lo general este *no* es el caso: normalmente tanto $\tilde{\beta}_1$ como $\tilde{\beta}_2$ son sesgados. La única excepción a esto es cuando x_1 y x_2 tampoco están correlacionadas.

Aun en el modelo anterior, que es bastante sencillo, puede ser difícil obtener la dirección de los sesgos de $\tilde{\beta}_1$ y $\tilde{\beta}_2$. Esto se debe a que x_1, x_2 y x_3 pueden estar correlacionadas por pares. De cualquier manera, una aproximación suele ser de utilidad práctica. Si se supone que x_1 y x_2 no están correlacionadas, entonces se puede estudiar el sesgo de $\tilde{\beta}_1$ como si x_2 estuviera ausente tanto del modelo poblacional como del modelo estimado. De hecho si x_1 y x_2 no están correlacionadas, se puede demostrar que

$$E(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i3}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}.$$

Esta ecuación es como la (3.45), pero sustituyendo β_3 por β_2 y x_3 por x_2 en la regresión (3.44). Por tanto, el sesgo en $\tilde{\beta}_1$ se obtiene al sustituir β_2 por β_3 y x_2 por x_3 en la tabla 3.2. Si $\beta_3 > 0$ y $\text{Corr}(x_1, x_3) > 0$, el sesgo en $\tilde{\beta}_1$ es positivo, y así de manera sucesiva.

Para dar un ejemplo, suponga que se agrega *exper* al modelo del salario:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u.$$

Si *abil* es omitida del modelo, los estimadores tanto de β_1 como de β_2 son sesgados, aun cuando se suponga que *exper* no está correlacionada con *abil*. En lo que se está interesado sobre todo es en el rendimiento de la educación, de manera que sería bueno que se pudiera concluir que $\tilde{\beta}_1$ tiene un sesgo hacia abajo o un sesgo hacia arriba debido a la omisión de la habilidad (*abil*).

Esta conclusión no es posible sin otros supuestos. Como una *aproximación*, suponga que, además de que *exper* y *abil* no están correlacionadas, *educ* y *exper* tampoco lo están. (En realidad, están correlacionadas de manera negativa.) Como $\beta_3 > 0$ y *educ* y *abil* están correlacionadas de forma positiva, $\tilde{\beta}_1$ tendrá un sesgo hacia arriba, tal como si *exper* no estuviera en el modelo.

El razonamiento empleado en el ejemplo anterior suele seguirse como una guía burda para obtener el posible sesgo en los estimadores de modelos más complicados. En general, la atención se centra en la relación entre una determinada variable explicativa, por ejemplo x_1 y el factor clave omitido. Hablando de manera estricta, ignorar todas las demás variables explicativas es válido sólo cuando cada una de ellas no esté correlacionada con x_1, pero de cualquier manera es una guía útil. En el apéndice 3A se encuentra un análisis más cuidadoso del sesgo de variable omitida con múltiples variables explicativas.

3.4 Varianza de los estimadores de MCO

Ahora se obtendrá la varianza de los estimadores de MCO de manera que, además de conocer la tendencia central de los $\hat{\beta}_j$ también se tendrá una medida de dispersión en su distribución de muestreo. Antes de hallar la varianza, se agregará un supuesto de homocedasticidad como en el capítulo 2. Esto se hace por dos razones. Primero, imponiendo el supuesto de varianza constante del error, se simplifican las fórmulas. Segundo, en la sección 3.5 se verá que si se agrega el supuesto de homocedasticidad, el método de MCO tiene una importante propiedad de eficiencia.

En el marco de la regresión múltiple, la homocedasticidad se expresa como sigue:

Supuesto RLM.5 (Homocedasticidad)

Dado cualquier valor de las variables explicativas, el error u tiene la misma varianza. En otras palabras, $\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2$.

El supuesto RLM.5 significa que la varianza en el término del error, u, condicional en las variables explicativas, es la *misma* para todas las combinaciones de valores de las variables explicativas. Si este supuesto no se satisface, entonces el modelo muestra heterocedasticidad, como ocurre en el caso de dos variables.

En la ecuación

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u,$$

la homocedasticidad requiere que la varianza del error no observado u no dependa de la educación, la experiencia o la antigüedad. Es decir,

$$\mathrm{Var}(u|educ, exper, tenure) = \sigma^2.$$

Si esta varianza cambia de acuerdo con alguna de las tres variables explicativas, se tiene heterocedasticidad.

A los supuestos RLM.1 a RLM.5 se les conoce como **supuestos de Gauss-Markov** (para regresiones de corte transversal). Estos supuestos, como se han dado hasta ahora, sólo son adecuados para el análisis de corte transversal con muestreo aleatorio. Como se verá, los supuestos de Gauss-Markov para el análisis de series de tiempo y para otras situaciones como el análisis de datos de panel, son más difíciles de expresar, aunque hay muchas semejanzas.

En el análisis que sigue, se empleará el símbolo **x** para denotar el conjunto de todas las variables independientes $(x_1, \ldots, x_k)$. Así, en la regresión con *educ*, *exper* y *tenure* como variables independientes, $\mathbf{x} = (educ, exper, tenure)$. Entonces, los supuestos RLM.1 a RLM.4 pueden expresarse como

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

y el supuesto RLM.5 corresponde a $\mathrm{Var}(y|\mathbf{x}) = \sigma^2$. Expresando los supuestos de esta manera, se muestra con claridad que el supuesto RLM.5 difiere enormemente del supuesto RLM.4. Este último dice que el valor esperado de y, dado **x**, es lineal en los parámetros, pero este valor esperado depende de $x_1, x_2, \ldots, x_k$. El supuesto RLM.5 dice que la varianza de y, dado **x**, *no* depende de los valores de las variables independientes.

Ahora se pueden obtener las varianzas de los $\hat{\beta}_j$, para lo que, una vez más, se condiciona sobre los valores muestrales de las variables independientes. La demostración se encuentra en el apéndice de este capítulo.

Teorema 3.2 (Varianza de muestreo de los estimadores de pendiente de MCO)

Bajo los supuestos RLM.1 a RLM.5, condicionales en los valores muestrales de las variables independientes,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{STC}_j(1 - \mathrm{R}_j^2)}, \qquad \boxed{3.51}$$

para $j = 1, 2, \ldots, k$, donde $\mathrm{STC}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ es la variación muestral total en x_j y R_j^2 es la R cuadrada de regresión de x_j sobre todas las otras variables independientes (e incluyendo un intercepto).

Antes de estudiar con más detalle la ecuación (3.51), es importante saber que para obtener esta fórmula se usan todos los supuestos de Gauss-Markov. Mientras que el supuesto de homocedasticidad no se necesitó para concluir que los estimadores de MCO son insesgados, sí se necesita para demostrar la ecuación (3.51).

La magnitud de $\mathrm{Var}(\hat{\beta}_j)$ tiene importancia práctica. Una varianza grande significa un estimador menos preciso y esto se traduce en intervalos de confianza grandes y pruebas de hipótesis menos exactas (como se verá en el capítulo 4). En la sección siguiente se analizan los elementos que comprende (3.51).

Los componentes de las varianzas de los estimadores de MCO: multicolinealidad

La ecuación 3.51 muestra que la varianza de $\hat{\beta}_j$ depende de tres factores: σ^2, STC_j y R_j^2. Recuerde que el subíndice j denota una de las variables independientes (por ejemplo, educación o tasa de pobreza). A continuación se considerarán cada uno de los factores que afectan $\mathrm{Var}(\hat{\beta}_j)$.

La varianza del error, σ^2. De acuerdo con la ecuación (3.51), una σ^2 más grande significa varianzas más grandes para los estimadores de MCO. Esto no es nada sorprendente: más "ruido" en la ecuación (una σ^2 más grande) dificulta más estimar el efecto parcial de cualquier variable independiente sobre y, y esto se refleja en varianzas más grandes para los estimadores de pendiente de MCO. Como σ^2 es una característica de la población, no tiene nada que ver con el tamaño de la muestra. El único componente de (3.51) que es desconocido es σ^2. Más adelante se verá cómo obtener un estimador insesgado de σ^2.

Dada una variable dependiente y sólo hay, en realidad, una manera de reducir la varianza del error: agregar más variables explicativas a la ecuación (extraer algunos factores del término del error). Por desgracia, no siempre es posible hallar factores adicionales justificados que afecten a y.

La variación muestral total en x_j, STC$_j$. De acuerdo con la ecuación (3.51) se observa que cuanto mayor sea la variación total en x_j, menor será Var($\hat{\beta}_j$). Por tanto, manteniendo constante todo lo demás, para estimar β_j se prefiere tener tanta variación muestral en x_j como sea posible. Esto ya se descubrió en el capítulo 2 en el caso de la regresión simple. Aunque es difícil que se puedan elegir los valores muestrales de las variables independientes, *hay* una manera de aumentar la variación muestral en cada una de las variables independientes: aumentar el tamaño de la muestra. En efecto, al muestrear de manera aleatoria una población, STC$_j$ aumenta sin límite a medida que la muestra se hace más grande. Este es el componente de la varianza que depende sistemáticamente del tamaño de la muestra.

Si STC$_j$ es pequeño, Var($\hat{\beta}_j$) puede volverse muy grande, pero una STC$_j$ pequeña no viola el supuesto RLM.3. Técnicamente, a medida que STC$_j$ se aproxima a cero, Var($\hat{\beta}_j$) se aproxima a infinito. El caso extremo en el que no hay variación muestral en x_j, STC$_j$ $= 0$ no es permitdo por el supuesto RLM.3.

Relaciones lineales entre las variables independientes, R_j^2. En la ecuación (3.51), el término R_j^2 es el más difícil de entender de los tres. Este término no aparece en el análisis de regresión simple porque en tales casos sólo hay una variable independiente. Es importante ver que esta R-cuadrada es distinta de la R-cuadrada de la regresión de y sobre $x_1, x_2, \dots, x_k$: esta R_j^2 se obtiene de una regresión en la que sólo intervienen las variables independientes del modelo original y donde x_j actúa como una variable dependiente.

Considere primero el caso $k = 2$: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Entonces, Var($\hat{\beta}_1$) $= \sigma^2/$[STC$_1(1 - R_1^2)$], donde R_1^2 es la R-cuadrada de la regresión simple de x_1 sobre x_2 (y, como siempre, un intercepto). Como la R-cuadrada mide la bondad de ajuste, un valor de R_1^2 cercano a uno indica que x_2 explica gran parte de la variación de x_1 en la muestra. Esto significa que x_1 y x_2 están fuertemente correlacionadas.

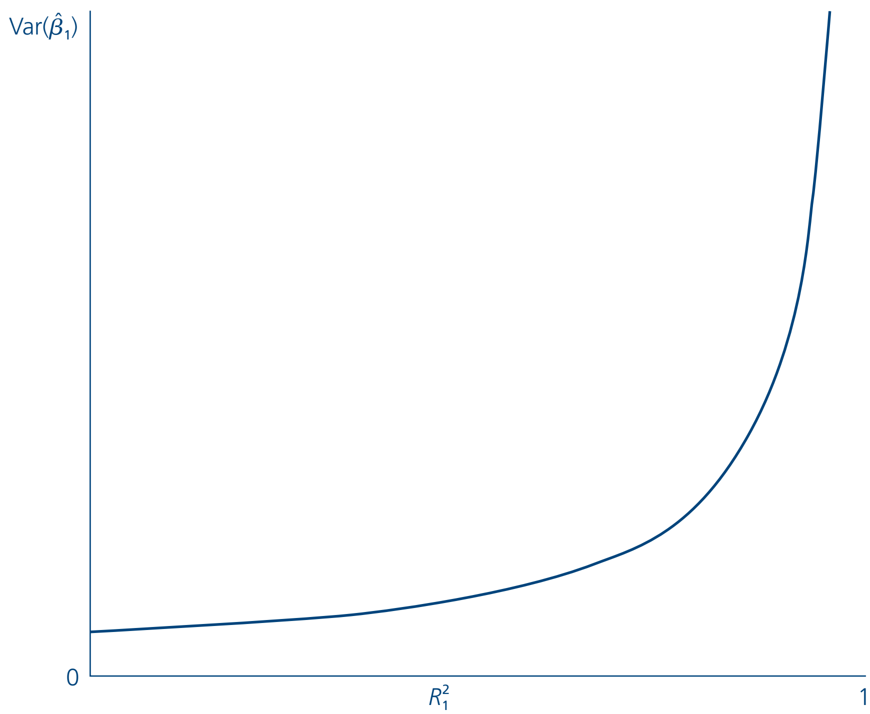
A medida que R_1^2 se aproxima a uno, Var($\hat{\beta}_1$) se hace cada vez más grande. Por tanto, un alto grado de relación lineal entre x_1 y x_2 puede conducir a varianzas grandes en los estimadores de pendiente de MCO. (Un argumento similar aplica a $\hat{\beta}_2$.) En la figura 3.1 se muestra la relación entre Var($\hat{\beta}_1$) y la R-cuadrada de la regresión de x_1 sobre x_2.

En el caso general, R_j^2 es la proporción de la variación total en x_j que puede ser explicada por las *otras* variables independientes que aparecen en la ecuación. Para σ^2 y STC$_j$ dadas, la menor Var($\hat{\beta}_j$) se obtiene cuando $R_j^2 = 0$, lo cual ocurre si y sólo si, x_j tiene correlación muestral cero con *cada una de las otras* variables independientes. Este es el mejor caso para estimar β_j, pero rara vez sucede.

El otro caso extremo, $R_j^2 = 1$, queda excluido por el supuesto RLM.3, porque $R_j^2 = 1$ significa que, en la muestra, x_j es una combinación lineal *perfecta* de algunas de las demás variables independientes de la regresión. Un caso más interesante es cuando el valor de R_j^2 es "cercano" a uno. De acuerdo con la ecuación (3.51) y con la figura 3.1, se ve que esto puede ocasionar

FIGURA 3.1

$\text{Var}(\hat{\beta}_1)$ como función de R_1^2.



que $\text{Var}(\hat{\beta}_j)$ sea grande: $\text{Var}(\hat{\beta}_j) \to \infty$ a medida que $R_j^2 \to 1$. A una correlación fuerte (pero no perfecta) entre dos o más variables independientes se le llama **multicolinealidad**.

Antes de analizar de manera más amplia el problema de la multicolinealidad es muy importante tener clara una cosa: el caso en que R_j^2 es cercana a uno *no* viola el supuesto RLM.3.

Como la multicolinealidad no viola ninguno de los supuestos, el "problema" de la multicolinealidad no está, en realidad, bien definido. Cuando se dice que la multicolinealidad surge al estimar β_j cuando R_j^2 es "cercana" a uno, "cercana" se pone entre comillas porque no hay un número absoluto que se pueda citar para concluir que la multicolinealidad es un problema. Por ejemplo, $R_j^2 = .9$ significa que 90% de la variación muestral en x_j puede ser explicada por las demás variables independientes del modelo de regresión. Sin duda, esto significa que x_j tiene una fuerte relación lineal con las demás variables independientes. Pero que esto se traduzca en que $\text{Var}(\hat{\beta}_j)$ sea demasiado grande para ser útil depende de las magnitudes de σ^2 y de STC_j. Como se verá en el capítulo 4 sobre inferencia estadística, lo que al final importa es qué tan grande es $\hat{\beta}_j$ en relación con su desviación estándar.

Así como valores grandes de R_j^2 pueden ocasionar que $\text{Var}(\hat{\beta}_j)$, sea grande, esto también puede ser ocasionado por valores pequeños de STC_j. Por tanto, también muestras pequeñas pueden conducir a varianzas de muestreo grandes. Preocuparse por altos grados de correlación entre las variables independientes en realidad es lo mismo que preocuparse por un tamaño de muestra pequeño: ambas cosas hacen que $\text{Var}(\hat{\beta}_j)$ aumente. Arthur Goldberger, el famoso econometrista de la Universidad de Wisconsin, ante la obsesión de los econometristas con la multicolinealidad, acuñó (en tono irónico) el término **micronumerosidad**, que se define como el problema del "tamaño de muestra pequeño". [Un interesante análisis sobre multicolinealidad y micronumerosidad se puede encontrar en Goldberger (1991).]

Aunque el problema de la multicolinealidad no está bien definido, una cosa está clara: permaneciendo todo lo demás constante, para estimar β_j, lo mejor es tener poca correlación entre x_j y las demás variables independientes. Esta observación suele conducir a la discusión de cómo "resolver" el problema de multicolinealidad. En las ciencias sociales, donde por lo común se es recolector pasivo de los datos, no hay otra manera de reducir la varianza de los estimadores insesgados que recolectar más datos. Dado un conjunto de datos, uno puede tratar de eliminar otras variables independientes del modelo con objeto de reducir la multicolinealidad. Por desgracia, eliminar una variable que pertenece al modelo poblacional puede llevar a sesgo, como se vio en la sección 3.3.

Quizás un ejemplo pueda ayudar a aclarar algunos de los problemas que surgen en relación con la multicolinealidad. Suponga que se desea estimar el efecto de diversas categorías de gastos en la educación sobre el desempeño de los estudiantes. Es posible que los gastos en sueldos para los profesores, material didáctico, deporte, etc., estén fuertemente correlacionados: las escuelas con mejor situación económica gastan más en todo y las escuelas pobres gastan menos en todo. Es claro que es difícil estimar el efecto de una determinada categoría de gastos sobre el desempeño de los estudiantes cuando es poca la variación en una categoría que no puede ser explicada por la variación en las otras categorías (esto conduce a una R_j^2 alta para todas las variables de gastos). Tales problemas de multicolinealidad pueden atenuarse recolectando más datos, pero en cierto sentido el problema ha sido autoimpuesto: se están haciendo preguntas que pueden ser demasiado difíciles de responder con precisión mediante los datos con que se cuenta. Tal vez sea mejor modificar el enfoque del análisis y colocar juntas todas las categorías de gastos, dado que ya no se trataría de estimar el efecto de cada una.

Otro punto importante es que un alto grado de correlación entre ciertas variables independientes puede ser irrelevante respecto a qué tan bien pueden estimarse otros parámetros del modelo. Considere, por ejemplo, un modelo con tres variables independientes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

en el que x_2 y x_3 están fuertemente correlacionadas. Entonces $\mathrm{Var}(\hat{\beta}_2)$ y $\mathrm{Var}(\hat{\beta}_3)$ pueden ser grandes. Pero la cantidad de correlación entre x_2 y x_3 no tiene efecto directo sobre $\mathrm{Var}(\hat{\beta}_1)$. De hecho, si x_1 no está correlacionada con x_2 y x_3, entonces $R_1^2 = 0$ y $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/\mathrm{STC}_1$, sin importar qué tanta correlación exista entre x_2 y x_3. Si β_1 es el parámetro de interés, no hay por qué preocuparse por la cantidad de correlación entre x_2 y x_3.

La observación anterior es importante porque los economistas suelen incluir muchas variables de control con objeto de aislar el efecto causal de una determinada variable. Por ejemplo, para estudiar la relación entre tasa de aprobación de créditos y el porcentaje de minorías en una vecindad pueden incluirse variables como ingreso promedio, valor promedio de la vivienda, medidas de capacidad crediticia, etc., ya que estos factores deben ser tomados en consideración con objeto de obtener conclusiones causales acerca de la discriminación. Ingreso, valor promedio de la vivienda y capacidad de crédito por lo general están fuertemente correlacionados

Pregunta 3.4

Suponga que plantea un modelo que explica la puntuación en el examen final en términos de la asistencia a clases. De manera que la variable independiente es la calificación en el examen final y la variable explicativa clave es la cantidad de clases a las que se asistió. Para controlar las habilidades de los estudiantes y sus esfuerzos fuera del aula, se incluye entre las variables explicativas el promedio general acumulado (GPA), la puntuación en el examen de admisión (SAT) y algunos datos sobre el desempeño en el bachillerato. Alguien le dice, "No puedes esperar obtener nada de esto porque es posible que el promedio general acumulado (GPA), la puntuación en el examen de admisión (SAT) y los datos sobre el desempeño en el bachillerato están fuertemente correlacionados". ¿Qué respondería usted a esto?

unos con otros. Pero la fuerte correlación entre estos controles no dificulta determinar los efectos de la discriminación

A algunos investigadores les parece útil calcular estadísticos con los que pretenden determinar la severidad de la multicolinealidad en una aplicación determinada. Por desgracia, es fácil usar mal estos estadísticos, ya que, como se ha dicho, no se puede especificar cuánta correlación entre las variables explicativas es "demasiada". Algunos "diagnósticos" de la multicolinealidad son estadísticos generales en el sentido de que detectan relaciones lineales fuertes entre cualquier conjunto de variables explicativas. Por las razones recién vistas, tales estadísticos tienen un valor cuestionable porque pueden indicar que hay un "problema" tan sólo porque dos variables control, cuyos coeficientes no interesan, están fuertemente correlacionadas. [Tal vez el estadístico general para multicolinealidad más común sea el llamado *número de condición*, que se define en términos de la matriz de datos completos y que está fuera del alcance de este libro. Vea, por ejemplo, Belsley, Kuh y Welsh (1980).]

Algo más útil, pero que también suele ser mal empleado, son los estadísticos para coeficientes individuales. El más común de éstos es el **factor inflacionario de la varianza (FIV)**, el cual se obtiene directamente de la ecuación (3.51). El FIV para el coeficiente de pendiente j es $\text{FIV}_j = 1/(1 - R_j^2)$, precisamente el término de $\text{Var}(\hat{\beta}_j)$ determinado por la correlación entre x_j y las demás variables explicativas. FIV_j es el factor por el cual $\text{Var}(\hat{\beta}_j)$ es más alta debido a que x_j no está no-correlacionada con todas las otras variables explicativas. Como FIV_j es función de R_j^2 —en realidad, la figura 3.1 es en esencia una gráfica de FIV_1— el análisis previo puede repetirse en términos del FIV. Por ejemplo, si se pudiera elegir, se desearía que FIV_j fuera más pequeño (permaneciendo todo lo demás constante). Pero rara vez se puede elegir. Si se cree que ciertas variables explicativas deben ser incluidas en la regresión para inferir la causalidad de x_j, entonces no se pensaría en omitirlas, y aun la creencia de que el FIV_j; es "demasiado alto" no puede afectar esta decisión. Si, por ejemplo, lo que se desea conocer es el efecto causal de x_1 especialmente sobre y, entonces deben ignorarse por completo los FIV de los demás coeficientes. Por último, establecer un valor límite para los FIV por encima del cual se concluya que la multicolinealidad es un "problema" es arbitrario y no es especialmente útil. Algunas veces se elige el valor 10: si FIV_j es mayor que 10 (en forma equivalente, R_j^2 es mayor a .9), entonces se concluye que la multicolinealidad es un "problema" para estimar β_j. Pero un FIV_j arriba de 10 no significa que la desviación estándar de $\hat{\beta}_j$ sea demasiado grande para ser útil porque la desviación estándar depende también de σ y de STC_j, y este último puede incrementarse aumentando el tamaño de la muestra. Por tanto, como ocurre al observar directamente la magnitud de R_j^2, observar la magnitud de FIV_j tiene una utilidad limitada, aunque puede interesar por curiosidad

Varianzas en modelos mal especificados

La decisión de incluir una determinada variable en un modelo de regresión puede tomarse analizando la disyuntiva entre sesgo y varianza. En la sección 3.3 se describió el sesgo que se induce al dejar fuera una variable relevante cuando el verdadero modelo contiene dos variables explicativas. Ahora se continúa el análisis de este modelo comparando las varianzas de los estimadores de MCO.

Se expresa el modelo poblacional verdadero, que satisface los supuestos de Gauss-Markov, como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Se considerarán dos estimadores de β_1. El estimador $\hat{\beta}_1$ proviene de la regresión múltiple

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.52}$$

Es decir, en el modelo de regresión se incluyen x_2 y x_1, El estimador $\tilde{\beta}_1$ se obtiene omitiendo x_2 del modelo y corriendo la regresión simple de y sobre x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.53}$$

Cuando $\beta_2 \neq 0$, en la ecuación (3.53) se excluye del modelo una variable relevante y, como se vio en la sección 3.3, esto induce un sesgo en $\tilde{\beta}_1$ a menos que x_1 y x_2 no estén correlacionadas. Por otro lado, $\hat{\beta}_1$ es un estimador insesgado de β_1 para cualquier valor de β_2, incluso para $\beta_2 = 0$. Se concluye que si el sesgo se usa como el único criterio, se prefiere $\hat{\beta}_1$ a $\tilde{\beta}_1$.

La conclusión anterior ya no es válida cuando entra en juego la varianza. Condicionando sobre los valores de x_1 y x_2 en la muestra, se tiene, de acuerdo con la ecuación (3.51),

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2/[\mathrm{STC}_1(1 - R_1^2)], \tag{3.54}$$

donde STC_1 es la variación total en x_1 y R_1^2 es la R-cuadrada de la regresión de x_1 sobre x_2. Además, una sencilla modificación de la demostración del capítulo 2 para la regresión en dos variables muestra que

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2/\mathrm{STC}_1. \tag{3.55}$$

Comparando (3.55) con (3.54) se observa que $\mathrm{Var}(\tilde{\beta}_1)$ es siempre *menor* que $\mathrm{Var}(\hat{\beta}_1)$, a menos que en la muestra x_1 y x_2 no estén correlacionadas, en cuyo caso los dos estimadores $\tilde{\beta}_1$ y $\hat{\beta}_1$ son iguales. Suponiendo que x_1 y x_2 no estén no-correlacionadas, se pueden formular las conclusiones siguientes:

1. Si $\beta_2 \neq 0$, $\tilde{\beta}_1$ es sesgado, $\hat{\beta}_1$ es insesgado y $\mathrm{Var}(\tilde{\beta}_1) < \mathrm{Var}(\hat{\beta}_1)$.
2. Si $\beta_2 = 0$, $\tilde{\beta}_1$ y $\hat{\beta}_1$ son insesgados y $\mathrm{Var}(\tilde{\beta}_1) < \mathrm{Var}(\hat{\beta}_1)$.

De acuerdo con la segunda conclusión, es claro que $\tilde{\beta}_1$ se prefiere si $\beta_2 = 0$. De manera intuitiva, si x_2 no tiene efecto parcial sobre y, entonces incluirla en el modelo sólo puede exacerbar el problema de multicolinealidad, lo que lleva a un estimador menos eficiente de β_1. El costo de incluir una variable irrelevante en el modelo es una varianza mayor del estimador de β_1.

El caso en el que $\beta_2 \neq 0$ es más difícil. Dejar x_2 fuera del modelo da como resultado un estimador sesgado de β_1. Los econometristas han sugerido comparar la posible magnitud del sesgo por omitir x_2 con la disminución de la varianza —resumida en la magnitud de R_1^2— para decidir si se incluye x_2. Sin embargo, si $\beta_2 \neq 0$, hay dos razones favorables para incluir x_2 en el modelo. La más importante de ellas es que cualquier sesgo en $\tilde{\beta}_1$ no se reduce al aumentar el tamaño de la muestra; de hecho, el sesgo no sigue por fuerza algún patrón. Por lo tanto, se puede pensar que el sesgo es más o menos el mismo cualquiera que sea el tamaño de la muestra. Por otro lado, tanto $\mathrm{Var}(\tilde{\beta}_1)$ como $\mathrm{Var}(\hat{\beta}_1)$ se reducen a cero a medida que n aumenta, lo que significa que la colinealidad inducida por agregar x_2 se vuelve menos importante a medida que aumenta el tamaño de la muestra. Con muestras grandes se prefiere $\hat{\beta}_1$.

La otra razón a favor de $\hat{\beta}_1$ es más sutil. La fórmula de la varianza en (3.55) está condicionada a los valores de x_{i1} y x_{i2} en la muestra, lo que proporciona el mejor escenario para $\tilde{\beta}_1$. Si $\beta_2 \neq 0$, la varianza de $\tilde{\beta}_1$ condicionada sólo a x_1 es mayor que la presentada en (3.55). De manera intuitiva, si $\beta_2 \neq 0$ y x_2 se excluye del modelo, la varianza del error aumenta debido a que el error en realidad contiene parte de x_2. Pero (3.55) ignora que la varianza del error aumenta porque considera a ambos regresores como no aleatorios. Analizar sobre cuáles variables independientes condicionar nos llevaría demasiado lejos. Baste decir que (3.55) es demasiado generosa cuando se trata de medir la precisión en $\tilde{\beta}_1$.

Estimación de σ^2: errores estándar de los estimadores de MCO

Ahora se mostrará cómo elegir un estimador insesgado de σ^2, el cual permitirá después obtener estimadores insesgados de $\text{Var}(\hat{\beta}_j)$.

Como $\sigma^2 = \text{E}(u^2)$, un estimador "insesgado" de σ^2 es el promedio muestral de los errores cuadrados: $n^{-1}\sum_{i=1}^{n} u_i^2$. Por desgracia, este no es un verdadero estimador porque los u_i no se pueden observar. Sin embargo, recuerde que es posible expresar los errores como $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}$, y la razón por la que no se pueden observar los u_i es que no se conocen los β_j. Cuando se sustituyen los β_j por sus estimadores de MCO, se obtienen los residuales de MCO:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}.$$

Parece natural estimar σ^2 sustituyendo las u_i por las $\hat{u}_i$. En el caso de la regresión simple, se vio que esto conduce a un estimador sesgado. El estimador insesgado de σ^2 en el caso general de la regresión múltiple es

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) \Big/ (n - k - 1) = \text{SRC}/(n - k - 1). \tag{3.56}$$

Este estimador ya se encontró en el caso $k = 1$ de la regresión simple.

El término $n - k - 1$ en (3.56) son los **grados de libertad** (***gl***) para el problema general de MCO con n observaciones y k variables independientes. Como en un modelo de regresión con k variables independientes y un intercepto hay $k + 1$ parámetros, se puede escribir

$$\begin{aligned} gl &= n - (k + 1) \\ &= (\text{número de observaciones}) - (\text{cantidad de parámetros estimados}). \end{aligned} \tag{3.57}$$

La manera más sencilla de calcular los grados de libertad en una determinada aplicación es contar la cantidad de parámetros, incluyendo al intercepto, y restar esta cantidad del número de observaciones. (En el raro caso de que no se estime la intersección, la cantidad de parámetros disminuyen uno.)

Técnicamente, la división entre $n - k - 1$ en (3.56) se debe a que el valor esperado de la suma de los residuales cuadrados es $\text{E(SRC)} = (n - k - 1)\sigma^2$. De manera intuitiva, uno puede imaginar por qué es necesario el ajuste de los grados de libertad volviendo a las condiciones de primer orden de los estimadores de MCO. Éstas pueden escribirse como $\sum_{i=1}^{n} \hat{u}_i = 0$ y $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, donde $j = 1, 2, \ldots, k$. Por tanto, en la obtención de las estimaciones de MCO, a los residuales de MCO se les impone $k + 1$ restricciones. Esto significa que, dados $n - (k + 1)$ residuales, los restantes $k + 1$ residuales se conocen: sólo hay $n - (k + 1)$ grados de libertad en los residuales. (Esto se puede comparar con los *errores* u_i, que tienen n grados de libertad en la muestra.)

Para referencia, se resume este análisis en el teorema 3.3. En el capítulo 2 se probó este teorema para el caso de la regresión simple (vea el teorema 2.3) (En el apéndice E se encuentra una demostración general que requiere álgebra matricial.)

Teorema 3.3 (Estimación insesgada de σ^2)

Bajo los supuestos RLM.1 a RLM.5 de Gauss-Markov, $E(\hat{\sigma}^2) = \sigma^2$.

A la raíz cuadrada positiva de $\hat{\sigma}^2$, que se denota $\hat{\sigma}$, se le llama **error estándar de la regresión** (**EER**). El EER es un estimador de la desviación estándar del término de error. Los paquetes que corren regresiones suelen dar esta estimación, aunque le dan distintos nombres. (Además de EER, a $\hat{\sigma}$ también se le llama *error estándar de la estimación* y *raíz cuadrática medio del error*).

Observe que cuando se agrega otra variable independiente a la regresión (para una muestra dada) $\hat{\sigma}$ puede aumentar o disminuir. Esto se debe a que, aunque SRC debe disminuir cuando se agrega otra variable explicativa, los grados de libertad también disminuyen en uno. Como SRC está en el numerador y *gl* en el denominador, no se puede decir de antemano cuál será el efecto que domine.

En el capítulo 4, para construir intervalos de confianza y realizar pruebas, se necesitará estimar la **desviación estándar de $\hat{\beta}_j$**, que es la raíz cuadrada de la varianza:

$$\text{de}(\hat{\beta}_j) = \sigma/[\text{STC}_j(1 - R_j^2)]^{1/2}.$$

Como σ no se conoce, se sustituye por su estimador, $\hat{\sigma}$. Esto da el **error estándar de $\hat{\beta}_j$**:

$$\text{ee}(\hat{\beta}_j) = \hat{\sigma}/[\text{STC}_j(1 - R_j^2)]^{1/2}. \tag{3.58}$$

Al igual que las estimaciones de MCO, los errores estándar pueden obtenerse de cualquier muestra. Como $\text{ee}(\hat{\beta}_j)$ depende de $\hat{\sigma}$, el error estándar tiene una distribución de muestreo, que será importante en el capítulo 4.

Hay que resaltar un punto acerca de los errores estándar. Como (3.58) se obtiene directamente de la fórmula de la varianza en (3.51) y dado que esta última se apoya en el supuesto de homocedasticidad RLM.5, se sigue que la fórmula del error estándar en (3.58) *no* es un estimador válido de $\text{de}(\hat{\beta}_j)$ si los errores muestran heterocedasticidad. Por tanto, mientras que la presencia de heterocedasticidad no causa sesgo en las $\hat{\beta}_j$, sí conduce a un sesgo en la fórmula usual para $\text{Var}(\hat{\beta}_j)$, lo cual invalida los errores estándar. Esto es importante porque los paquetes para regresión, si no se les indica otra cosa, calculan (3.58) como el error estándar predeterminado de cada coeficiente (con una representación un poco diferente para el intercepto). Si se sospecha de heterocedasticidad, entonces los errores estándar "usuales" de MCO no son válidos y se deberán tomar medidas para corregir el problema. En el capítulo 8 se verán los métodos para el problema de la heterocedasticidad.

3.5 Eficiencia de MCO: el teorema de Gauss-Markov

En esta sección se enuncia y analiza el importante **teorema de Gauss-Markov**, el cual justifica el uso del método de MCO en lugar de otros diversos estimadores. Ya se conoce una justificación para el método de MCO: bajo los supuestos RLM.1 a RLM.4, el método de MCO es insesgado. Sin embargo, bajo estos supuestos hay *muchos* estimadores insesgados de las β_j (véase, por ejemplo, el problema 3.13). ¿Hay otros estimadores cuyas varianzas sean menores que las de los estimadores de MCO?

Si se limita la clase de los posibles estimadores apropiados, entonces se puede demostrar que el método de MCO *es* el mejor dentro de su clase. En concreto, se argumentará que, bajo los supuestos RLM.1 a RLM.5, el estimador de MCO $\hat{\beta}_j$ es el **mejor estimador lineal insesgado** (**MELI**) para β_j. Para enunciar el teorema es necesario entender cada componente del acrónimo "MELI". Primero, ya se sabe qué es un estimador: es una regla que puede aplicarse a cualquier muestra de datos para obtener una estimación. Ya se sabe qué es un estimador insesgado: en el contexto presente, un estimador de β_j, por ejemplo, $\tilde{\beta}_j$, es un estimador insesgado de β_j si $E(\tilde{\beta}_j) = \beta_j$ para toda $\beta_0, \beta_1, \ldots, \beta_k$.

¿Qué significa el término "lineal"? En el presente contexto, un estimador $\tilde{\beta}_j$ de β_j es lineal si, y sólo si, se puede expresar como una función lineal de los datos de la variable dependiente:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \tag{3.59}$$

donde cada w_{ij} puede ser una función de los valores muestrales de todas las variables independientes. Como se puede ver de acuerdo con la ecuación (3.22), los estimadores de MCO son lineales.

Por último, ¿cómo se define "mejor"? En el presente teorema, mejor se define como *menor varianza*. Dados dos estimadores insesgados, es lógico preferir el que tenga menor varianza (véase el apéndice C).

Ahora, sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ los estimadores de MCO del modelo (3.31) bajo los supuestos RLM.1 a RLM.5. El teorema de Gauss-Markov dice que: dado cualquier estimador $\tilde{\beta}_j$ que sea *lineal* e *insesgado*, $\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\tilde{\beta}_j)$ y esta desigualdad es, por lo general, estricta. Es decir, en la clase de los estimadores lineales insesgados, los estimadores de MCO tienen la mínima varianza (bajo los cinco supuestos de Gauss-Markov). En realidad, el teorema dice más aún. Si se quiere estimar una función lineal de los β_j, entonces la correspondiente combinación lineal de los estimadores de MCO proporciona la menor varianza entre todos los estimadores lineales insesgados. Se concluye con un teorema que se demuestra en el apéndice 3A.

Teorema 3.4 (Teorema de Gauss-Markov)

Bajo los supuestos RLM.1 a RLM.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ son los mejores estimadores lineales insesgados (MELI) de $\beta_0, \beta_1, \ldots, \beta_k$, respectivamente.

A este teorema se debe que los supuestos RLM.1 a RLM.5 se conozcan como los supuestos de Gauss-Markov (en el análisis de corte transversal).

La importancia del teorema de Gauss-Markov es que, si el conjunto estándar de supuestos se satisface, no es necesario buscar otros estimadores insesgados de la forma (3.59): ninguno será mejor que los estimadores de MCO. Esto es equivalente a decir que para cualquier otro estimador que sea lineal e insesgado, su varianza será por lo menos tan grande como la varianza de los estimadores de MCO; no es necesario hacer ningún cálculo para saber esto.

Para los propósitos presentes, el teorema 3.4 justifica el uso del método de MCO para estimar los modelos de regresión múltiple. Si no se satisface alguno de los supuestos de Gauss-Markov, entonces este teorema no es válido. Se sabe ya que si el supuesto de media condicional cero no se satisface (supuesto RLM.4), esto ocasiona que los estimadores de MCO sean sesgados, con lo que el teorema 3.4 ya no es válido. También se sabe ya que la heterocedasticidad (insatisfacción del supuesto RLM.5) no ocasiona sesgo. Sin embargo, en presencia de heterocedasticidad, los estimadores de MCO ya no son los de menor varianza entre los estimadores lineales insesgados. En el capítulo 8 se analiza un estimador que perfecciona los estimadores de MCO en presencia de heterocedasticidad.

RESUMEN

1. El modelo de regresión múltiple permite mantener de manera efectiva otros factores constantes mientras se examina el efecto de una determinada variable independiente sobre la variable dependiente. Esto permite de forma explícita que las variables independientes estén correlacionadas.
2. Aunque el modelo es lineal en sus *parámetros*, puede usarse para modelar relaciones no lineales mediante la elección adecuada de las variables dependientes e independientes.
3. El método de mínimos cuadrados ordinarios es fácil de emplear para estimar el modelo de regresión múltiple. Cada estimación de pendiente mide el efecto parcial de la variable independiente correspondiente sobre la variable dependiente, manteniendo constantes todas las demás variables independientes.
4. R^2 es la proporción de la variación muestral en la variable dependiente explicada por las variables independientes y se usa como una medida de la bondad de ajuste. Al evaluar modelos econométricos es importante no darle demasiada importancia a R^2.
5. Bajo los primeros cuatro supuestos de Gauss-Markov (RLM.1 a RLM.4), los estimadores de MCO son insesgados. Esto implica que la inclusión de una variable irrelevante en el modelo no tiene ningún efecto sobre el insesgamiento del intercepto ni de los otros estimadores de pendiente. Por otro lado, la omisión de una variable relevante ocasiona que los estimadores de MCO sean sesgados. En muchas circunstancias se puede determinar la dirección del sesgo.
6. Bajo los cinco supuestos de Gauss-Markov, la varianza de un estimador de pendiente de MCO está dada por $\mathrm{Var}(\hat{\beta}_j) = \sigma^2/[\mathrm{STC}_j(1 - R_j^2)]$. A medida que la varianza del error σ^2 aumenta, también aumenta $\mathrm{Var}(\hat{\beta}_j)$, mientras que $\mathrm{Var}(\hat{\beta}_j)$ disminuye a medida que aumenta la variación en x_j, STC_j. El término R_j^2 mide la cantidad de colinealidad entre x_j y las demás variables explicativas. A medida que R_j^2 se aproxima a uno, $\mathrm{Var}(\hat{\beta}_j)$ crece ilimitadamente.
7. Agregar una variable irrelevante a una ecuación por lo general hace que aumente la varianza del resto de los estimadores de MCO debido a la multicolinealidad.
8. Bajo los supuestos de Gauss-Markov (RLM.1 a RLM.5), los estimadores de MCO son los mejores estimadores lineales insesgados (MELI).

Los supuestos de Gauss-Markov

El siguiente es un resumen de los cinco supuestos de Gauss-Markov empleados en este capítulo. Recuerde que los cuatro primeros se usaron para demostrar el insesgamiento de los estimadores de MCO, mientras que el último se agregó para obtener las fórmulas usuales para la varianza y para concluir que los estimadores de MCO son los mejores estimadores lineales insesgados.

Supuesto RLM.1 (Lineal en los parámetros)

El modelo poblacional puede expresarse como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros (constantes) desconocidos de interés y u es el error aleatorio o término de perturbación no observable.

Supuesto RLM.2 (Muestreo aleatorio)

Se tiene un muestreo aleatorio con n observaciones, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, de acuerdo con el modelo poblacional del supuesto RLM.1.

Supuesto RLM.3 (No hay colinealidad perfecta)

En la muestra (y por tanto en la población), ninguna de las variables independientes es constante y no hay relaciones *lineales exactas* entre las variables independientes.

Supuesto RLM.4 (Media condicional cero)

El error u tiene un valor esperado de cero dados cualesquiera valores de las variables independientes. En otras palabras,

$$E(u|x_1, x_2, \ldots, x_k) = 0.$$

Supuesto RLM.5 (Homocedasticidad)

El error u tiene la misma varianza dado cualquier valor de las variables explicativas. En otras palabras,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

TÉRMINOS CLAVE

Análisis de error de especificación
Análisis de regresión múltiple
Ceteris paribus
Colinealidad perfecta
Condiciones de primer orden
Desviación estándar de $\hat{\beta}_j$
Efecto parcial
Error estándar de $\hat{\beta}_j$
Error estándar de la regresión (EER)
Estimación del intercepto de MCO
Estimación de pendiente de MCO
Exclusión de una variable relevante
Factor inflacionario de la varianza (FIV)
Función de regresión muestral (FRM)
Grados de libertad (*gl*)
Inclusión de una variable irrelevante
Intercepto
Línea de regresión de MCO
Mejor estimador lineal insesgado (MELI)
Micronumerosidad
Mínimos cuadrados ordinarios (MCO)
Modelo de regresión lineal múltiple
Modelo poblacional
Modelo verdadero
Multicolinealidad
Parámetro de pendiente
Perturbación
Residual
Sesgado hacia cero
Sesgo de la variable omitida
Sesgo hacia abajo
Sesgo hacia arriba
Sobreespecificación del modelo
Subespecificación de modelo
Suma de residuales cuadrados (SRC)
Suma explicada de cuadrados (SEC)
Suma total de cuadrados (STC)
Supuestos de Gauss-Markov
Teorema de Gauss-Markov
Término de error
Variable explicativa endógena
Variable explicativa exógena

PROBLEMAS

3.1 Utilizando la base de datos GPA2.RAW de 4,137 alumnos universitarios, se estimó la ecuación siguiente de MCO:

$$\widehat{colgpa} = 1.392 - .0135\ hsperc + .00148\ sat$$
$$n = 4{,}137,\ R^2 = .273,$$

donde *colgpa* es el promedio de calificaciones que se mide sobre una escala de cuatro puntos, *hsperc* es el percentil en la clase de bachillerato que se gradúa (definida de manera que, por ejemplo, $hsperc = 5$ significa el 5% *superior* de la clase), y *sat* son los puntajes combinados en matemáticas y habilidades verbales en la prueba de logro de los alumnos.

i) ¿Por qué es lógico que el coeficiente en *hsperc* sea negativo?

ii) ¿Cual es el promedio de calificaciones universitario predicho cuando $hsperc = 20$ y $sat = 1{,}050$?

iii) Suponga que dos graduados de bachillerato, A y B, se gradúan en el mismo percentil de bachillerato, pero el puntaje SAT del estudiante A es 140 puntos más alto (aproximadamente una desviación estándar en la muestra). ¿Cuál es la diferencia predicha en el promedio de calificaciones universitario para estos dos alumnos? ¿Es grande la diferencia?

iv) Manteniendo *hsperc* constante, ¿qué diferencia en las puntuaciones SAT conduce a una diferencia estimada *colgpa* de .50, o medio punto de puntuación? Comente su respuesta.

3.2 Los datos en el archivo WAGE2.RAW sobre trabajadores hombres se utilizan para estimar la ecuación siguiente:

$$\widehat{educ} = 10.36 - .094\ sibs + .131\ meduc + .210\ feduc$$
$$n = 722,\ R^2 = .214,$$

donde *educ* es años de escolaridad, *sibs* es número de hermanos, *meduc* es años de escolaridad de la madre y *feduc* años de escolaridad del padre.

i) ¿Tiene *sibs* el efecto esperado? Explique. Manteniendo constantes *meduc* y *feduc* ¿cuánto tiene que aumentar *sibs* para tener una reducción de un año en los años de educación predichos? (Aquí es aceptable una respuesta en números no enteros.)

ii) Explique la interpretación del coeficiente de *meduc*.

iii) Suponga que el hombre A no tiene hermanos, y que su madre y su padre tienen cada uno 12 años de escolaridad. El hombre B no tiene hermanos y su madre y su padre tienen cada uno 16 años de escolaridad. ¿Cuál es la diferencia entre B y A en años predichos de escolaridad?

3.3 El siguiente modelo es una versión simplificada del modelo de regresión múltiple utilizado por Biddle y Hamermesh (1990) para estudiar el intercambio entre tiempo dedicado al sueño y dedicado al trabajo, así como ver otros factores que afectan el sueño:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

donde *sleep* y *totwrk* (trabajo total) se miden en minutos por semana y *educ* y *age* se miden en años. (Vea también el ejercicio de computadora C2.3.)

i) Si los adultos intercambian sueño por trabajo, ¿cuál es el signo de β_1?

ii) ¿Qué signos cree que tendrán β_2 y β_3?

iii) Utilizando los datos del archivo SLEEP75.RAW, la ecuación estimada es

$$\widehat{sleep} = 3{,}638.25 - .148\ totwrk - 11.13\ educ + 2.20\ age$$
$$n = 706,\ R^2 = .113.$$

Si una persona trabaja cinco horas más a la semana, ¿cuántos minutos se predice que disminuya *sleep*? ¿Es este un intercambio grande?

iv) Analice el signo y la magnitud del coeficiente estimado para *educ*.

v) ¿Diría que *totwrk*, *educ* y *age* explican gran parte de la variación en *sleep*? ¿Qué otros factores podrían afectar el tiempo dedicado al sueño? ¿Es probable que estén correlacionados con *totwrk*?

3.4 El sueldo inicial medio para los recién graduados de la Facultad de Derecho se determina mediante

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

donde *LSAT* es la media del puntaje LSAT del grupo de graduados, *GPA* es la media del GPA (promedio general) del grupo, *libvol* es el número de volúmenes en la biblioteca de la Facultad de Derecho, *cost* es el costo anual por asistir a dicha facultad y *rank* es una clasificación de las escuelas de derecho (siendo $rank = 1$ la mejor).

i) Explique por qué se espera $\beta_5 \leq 0$.

ii) ¿Qué signos espera para los otros parámetros de pendiente? Justifique sus respuestas.

iii) Utilizando los datos del archivo LAWSCH85.RAW, la ecuación estimada es

$$\widehat{\log(salary)} = 8.34 + .0047\, LSAT + .248\, GPA + .095 \log(libvol) + .038 \log(cost) - .0033\, rank$$

$$n = 136,\; R^2 = .842.$$

¿Cuál es la diferencia *ceteris paribus* predicha en sueldo entre escuelas cuya media en el GPA difiera por un punto? (Responda en términos porcentuales.)

iv) Interprete el coeficiente de la variable log(*libvol*).

v) ¿Diría que es preferible asistir a una Facultad de Derecho con ranking más alto? En términos de sueldo inicial predicho, ¿cuánto vale una diferencia de 20 en el ranking?

3.5 En un estudio que relaciona el promedio de puntaje en las calificaciones universitarias con el tiempo utilizado en diversas actividades, usted distribuye una encuesta entre varios alumnos. A los alumnos se les pregunta cuántas horas utilizan a la semana en cuatro actividades: estudiar, dormir, trabajar y diversión. Toda actividad que realicen se ubica en una de las cuatro categorías, de modo que para cada alumno la suma de horas en las cuatro actividades debe ser 168.

i) En el modelo

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

¿tiene sentido mantener constantes *sleep*, *work* y *leisure*, y variar *study*?

ii) Explique por qué este modelo viola el supuesto RLM.3.

iii) ¿Cómo podría reformular el modelo para que sus parámetros tuvieran una interpretación útil y se satisfaga el supuesto RLM.3?

3.6 Considere un modelo de regresión múltiple que contiene tres variables independientes, bajo los Supuestos RLM.1 a RLM.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Le interesa calcular la suma de los parámetros de x_1 y x_2; llame a esto $\theta_1 = \beta_1 + \beta_2$.

i) Muestre que $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ es un estimador insesgado de θ_1.
ii) Determine $\text{Var}(\hat{\theta}_1)$ en términos de $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, y de $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

3.7 ¿Qué de lo siguiente puede causar que los estimadores MCO sean sesgados?
i) La heterocedasticidad.
ii) La omisión de una variable importante.
iii) Un coeficiente de correlación muestral de .95 entre dos variables independientes incluidas en el modelo.

3.8 Suponga que la productividad promedio de los trabajadores de las empresas manufactureras (*avgprod*) depende de dos factores, el promedio de horas de capacitación (*avgtrain*) y la habilidad promedio de los trabajadores (*avgabil*):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u.$$

Asuma que esta ecuación satisface los supuestos de Gauss-Markov. Si se han otorgado subvenciones a las empresas cuyos trabajadores tienen habilidades inferiores al promedio, de manera que *avgtrain* y *avgabil* están correlacionadas de manera negativa, ¿cuál es el sesgo probable en $\tilde{\beta}_1$ que se obtiene de la regresión simple de *avgprod* sobre *avgtrain*?

3.9 La siguiente ecuación describe la media del precio de la vivienda en una comunidad en términos de cantidad de contaminación (*nox* por óxido nitroso) y del número promedio de habitaciones en las casas de la comunidad (*rooms*):

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u.$$

i) ¿Cuáles son los signos probables de β_1 y β_2? ¿Cuál es la interpretación de β_1? Explique.
ii) ¿Por qué podría *nox* [o con más precisión, log(*nox*)] y *rooms* estar correlacionados de manera negativa? Si es este el caso, ¿produce la regresión simple de log(*price*) sobre log(*nox*) un estimador de β_1 con sesgo hacia arriba o hacia abajo?
iii) Utilizando los datos del archivo HPRICE2.RAW, se estimaron las siguientes ecuaciones:

$$\widehat{\log(price)} = 11.71 - 1.043 \log(nox), n = 506, R^2 = .264.$$

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306\, rooms, n = 506, R^2 = .514.$$

¿Las estimaciones de la elasticidad de *price* con respecto a *nox* de las regresiones simple y múltiple tienen la relación de lo que usted hubiera predicho, dada su respuesta en el inciso ii)? ¿Quiere esto decir que $-.718$ está en definitiva más cerca de la verdadera elasticidad que -1.043?

3.10 Suponga que está interesado en calcular la relación *ceteris paribus* entre y y x_1. Para tal propósito, puede allegarse datos sobre dos variables de control, x_2 y x_3. (Para mayor concreción, puede pensar en y como el puntaje de examen final, en x_1 como la asistencia a clases, en x_2 como el GPA acumulado hasta el semestre previo, y en x_3 como el puntaje de los exámenes de admisión SAT o ACT.) Sea $\tilde{\beta}_1$ la estimación de la regresión simple de y sobre x_1 y sea $\hat{\beta}_1$ el estimado de la regresión múltiple de y sobre x_1, x_2, x_3.
i) Si x_1 está altamente correlacionada con x_2 y x_3 en la muestra, y x_2 y x_3 tienen efectos parciales grandes sobre y, ¿esperaría que $\tilde{\beta}_1$ y $\hat{\beta}_1$ sean similares o muy diferentes? Explique.

ii) Si x_1 casi no está correlacionada con x_2 y x_3, pero x_2 y x_3 están fuertemente correlacionadas ¿$\tilde{\beta}_1$ y $\hat{\beta}_1$ tenderán a ser similares o muy diferentes? Explique.

iii) Si x_1 está fuertemente correlacionada con x_2 y x_3, y x_2 y x_3 tienen efectos parciales pequeños sobre y, ¿esperaría que ee($\tilde{\beta}_1$) o ee($\hat{\beta}_1$) fueran más pequeños? Explique.

iv) Si x_1 casi no está correlacionada con x_2 y x_3, x_2 y x_3 tienen efectos parciales grandes sobre y, y x_2 y x_3 están fuertemente correlacionadas, ¿esperaría que ee($\tilde{\beta}_1$) o ee($\hat{\beta}_1$) fueran más pequeños? Explique.

3.11 Suponga que el modelo poblacional para determinar y es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

y que este modelo satisface los supuestos RLM.1 a RLM.4. Sin embargo, se estima el modelo en el que se omite x_3. Sean $\tilde{\beta}_0$, $\tilde{\beta}_1$ y $\tilde{\beta}_2$ los estimadores de MCO de la regresión de y sobre x_1 y x_2. Muestre que el valor esperado de $\tilde{\beta}_1$ (dados los valores de las variables independientes en la muestra) es

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

donde los $\hat{r}_{i1}$ son los residuales de MCO de la regresión de x_1 sobre x_2. [*Sugerencia:* la fórmula para $\tilde{\beta}_1$ viene de la ecuación (3.22). Sustituya $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ en esta ecuación. Después de algunas manipulaciones algebraicas, tome la esperanza considerando a x_{i3} y a $\hat{r}_{i1}$ como no aleatorias.]

3.12 La siguiente ecuación representa los efectos de la composición del ingreso fiscal sobre el crecimiento del empleo subsecuente en la población de los condados de Estados Unidos:

$$growth = \beta_0 + \beta_1 share_{\mathrm{P}} + \beta_2 share_{\mathrm{I}} + \beta_3 share_{\mathrm{S}} + other\ factors,$$

donde *growth* es el cambio porcentual en el empleo de 1980 a 1990, $share_{\mathrm{P}}$ es la partición de los impuestos a la propiedad en el total de ingresos fiscales, $share_{\mathrm{I}}$ es la participación de impuestos sobre el ingreso, y $share_{\mathrm{S}}$ es la participación del impuesto sobre las ventas en el ingreso fiscal. Todas estas variables están medidas en 1980. La participación omitida, $share_{\mathrm{F}}$, incluye tarifas e impuestos misceláneos. Por definición, las cuatro porciones suman uno. Otros factores incluirían gastos en educación, infraestructura, etc. (todos medidos en 1980).

i) ¿Por qué se debe omitir de la ecuación una de las variables de participación en los impuestos?

ii) Dé una interpretación detallada de β_1.

3.13 i) Considere el modelo de regresión simple $y = \beta_0 + \beta_1 x + u$ bajo los primeros cuatro supuestos de Gauss-Markov. Para alguna función $g(x)$, por ejemplo $g(x) = x^2$ o $g(x) = \log(1 + x^2)$, defina $z_i = g(x_i)$. Defina un estimador de pendiente como

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z}) y_i\right) \Bigg/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right).$$

Muestre que $\tilde{\beta}_1$ es lineal e insesgado. Recuerde que, debido a $\mathrm{E}(u|x) = 0$, puede tratar a x_i y a z_i como no aleatorios en sus cálculos.

ii) Agregue el supuesto de homocedasticidad RLM.5. Muestre que

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} (z_i - \bar{z})^2\right) \Bigg/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right)^2.$$

iii) Muestre de manera directa que, bajo los supuestos de Gauss-Markov, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$, donde $\hat{\beta}_1$ es el estimador de MCO. [*Sugerencia:* la desigualdad de Cauchy-Schwartz en el apéndice B implica que

$$\left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right)\left(n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right);$$

note que podemos eliminar $\bar{x}$ de la covarianza muestral.]

EJERCICIOS EN COMPUTADORA

C3.1 Un problema de interés para los funcionarios de salud (y para otros) es determinar los efectos que el fumar durante el embarazo tiene sobre la salud infantil. Una medida de la salud infantil es el peso al nacer; un peso demasiado bajo puede ubicar al niño en riesgo de contraer varias enfermedades. Ya que es probable que otros factores que afectan el peso al nacer estén correlacionados con fumar, deben considerarse. Por ejemplo, un nivel de ingresos más alto en general da como resultado el acceso a mejores cuidados prenatales y a una mejor nutrición de la madre. Una ecuación que reconoce estos factores es

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

i) ¿Cual es el signo más probable para β_2?

ii) ¿Cree que *cigs* y *faminc* estén correlacionados? Explique por qué la correlación puede ser positiva o negativa.

iii) Ahora, calcule la ecuación con y sin *faminc* utilizando los datos del archivo BWGHT.RAW. Dé los resultados en forma de ecuación incluyendo el tamaño de la muestra y la R-cuadrada. Explique sus resultados enfocándose en si el añadir *faminc* modifica de manera sustancial el efecto esperado de *cigs* sobre *bwght*.

C3.2 Utilice los datos del archivo HPRICE1.RAW para estimar el modelo

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

donde *price* es el precio de casas dado en miles de dólares.

i) Escriba los resultados en forma de ecuación.

ii) ¿Cual es el incremento en precio estimado para una casa con una habitación (*bdrms*) más, manteniendo constante la superficie en pies cuadrados (*sqrft*)?

iii) ¿Cual es el incremento en precio estimado para una casa con una habitación adicional de 140 pies cuadrados? Compare esto con su respuesta al inciso (ii).

iv) ¿Qué porcentaje de la variación en el precio se explica por la extensión en pies cuadrados y el número de habitaciones?

v) La primera casa en la muestra tiene *sqrft* = 2,438 y *bdrms* = 4. Determine el precio de venta estimado para esta casa con la línea de regresión de MCO.

vi) El precio de venta de la primera casa en la muestra fue $300,000 (así que *price* = 300). Determine el residual para esta casa. ¿Sugiere esto que el comprador pagó de más o de menos por la casa?

C3.3 El archivo CEOSAL2.RAW contiene datos de 177 CEO (directores generales) y puede utilizarse para examinar los efectos del desempeño de la empresa sobre el sueldo de los CEO.

i) Estime un modelo que relacione el sueldo anual (*salary*) con las ventas de la empresa (*sales*) y el precio de mercado (*mktval*). Use el tipo de modelo que tiene elasticidad constante para ambas variables independientes. Escriba los resultados en forma de ecuación.

ii) Añada *profits* (utilidades de la empresa) al modelo del inciso (i). ¿Por qué esta variable no puede incluirse en forma logarítmica? ¿Diría usted que estas variables de desempeño de la empresa explican la mayor parte de la variación en sueldos de los CEO?

iii) Añada la variable *ceoten* (antigüedad del CEO en el puesto) al modelo del inciso (ii). ¿Cuál es el rendimiento porcentual estimado por un año más de permanencia del CEO en la empresa, manteniendo constantes los otros factores?

iv) Encuentre el coeficiente de correlación muestral entre las variables log(*mktval*) y *profits*. ¿Estas variables están fuertemente correlacionadas? ¿Qué indica esto sobre los estimadores de MCO?

C3.4 Para este ejercicio, utilice los datos del archivo ATTEND.RAW.

i) Obtenga los valores mínimo, máximo y promedio para las variables *atndrte*, *priGPA*, y *ACT* (porcentaje de asistencia a clases, calificación promedio general acumulada, calificación en el examen de admisión a la universidad, respectivamente).

ii) Estime el modelo

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u,$$

y escriba los resultados en forma de ecuación. Interprete el intercepto. ¿Tiene un significado útil?

iii) Analice los coeficientes de pendiente estimados. ¿Hay alguna sorpresa?

iv) ¿Cuál es el *atndrte* si $priGPA = 3.65$ y $ACT = 20$? ¿Qué piensa de este resultado? ¿Hay alumnos en la muestra con estos valores de las variables explicativas?

v) Si el alumno A tiene $priGPA = 3.1$ y $ACT = 21$ y el alumno B tiene $priGPA = 2.1$ y $ACT = 26$, ¿cuál es la diferencia predicha en sus tasas de asistencia?

C3.5 Confirme la interpretación de descuento de los efectos parciales de las estimaciones de MCO, haciendo de manera explícita tal descuento para el ejemplo 3.2. Esto requiere primero regresar *educ* sobre *exper* y *tenure* y guardando los residuales, $\hat{r}_1$. Después, regrese log(*wage*) sobre $\hat{r}_1$. Compare el coeficiente de $\hat{r}_1$ con el coeficiente de *educ* en la regresión de log(*wage*) sobre *educ*, *exper* y *tenure*.

C3.6 Para este problema, utilice los datos del archivo WAGE2.RAW. Como de costumbre, asegúrese de que todas las regresiones que siguen contengan un intercepto.

i) Corra una regresión simple de *IQ* sobre *educ* para obtener el coeficiente de pendiente, por ejemplo, $\tilde{\delta}_1$.

ii) Corra la regresión simple de log(*wage*) sobre *educ* y obtenga el coeficiente de pendiente, $\tilde{\beta}_1$.

iii) Corra la regresión múltiple de log(*wage*) sobre *educ* e *IQ* y obtenga los coeficientes de pendiente, $\hat{\beta}_1$ y $\hat{\beta}_2$, respectivamente.

iv) Verifique que $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

C3.7 Utilice los datos del archivo MEAP93.RAW para responder esta pregunta.

i) Estime el modelo (visto en el ejemplo 2.12):

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u,$$

y dé los resultados en la forma usual, incluyendo el tamaño de la muestra y la *R*-cuadrada. ¿Son los signos de los coeficientes de pendiente los que usted esperaba? Explique.

ii) ¿Qué piensa del intercepto estimado en el inciso i)? en particular, ¿tiene sentido igualar a cero las dos variables explicativas? [*Sugerencia*: recuerde que log(1)=0.]

iii) Ahora corra la regresión simple de *math10* sobre log(*expend*) y compare el coeficiente de pendiente con la estimación obtenida en el inciso i). ¿Es ahora el efecto estimado de los gastos por estudiante mayor o menor que en el inciso i)?

iv) Determine la correlación entre *lexpend* = log(*expend*) y *lnchprg*. ¿Le parece razonable el signo?

v) Use el inciso iv) para explicar sus hallazgos del inciso iii).

C3.8 Utilice la base de datos DISCRIM.RAW para responder esta pregunta. Hay datos sobre los precios de diversos artículos de restaurantes de comida rápida situados en zonas con distinto código postal, así como características de la población residente en dicho código, pertenecientes a Nueva Jersey y Pennsylvania. La idea es ver si los restaurantes de comida rápida tienen precios más altos en áreas con mayor concentración de población afroestadounidense.

i) Determine los valores promedio, en la muestra, de *prpblck* (proporción de afroestadounidenses) y de *income* (ingreso familiar medio), junto con sus desviaciones estándar. ¿Cuáles son las unidades de *prpblck* y de *income*?

ii) Considere un modelo para explicar el precio de las bebidas refrescantes, *psoda*, en términos de la proporción de población afroestadounidense y de ingreso medio:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

Estime este modelo mediante MCO y dé los resultados en forma de ecuación incluyendo el tamaño de la muestra y la *R*-cuadrada (No use notación científica al dar las estimaciones). Interprete el coeficiente de *prpblck*. ¿Considera que es grande desde el punto de vista económico?

iii) Compare la estimación del inciso ii) con la estimación mediante regresión simple de *psoda* sobre *prpblck*. ¿Es el efecto de la discriminación mayor o menor cuando se controla el ingreso?

iv) Un modelo con una elasticidad constante del precio respecto al ingreso puede ser más apropiado. Proporcione las estimaciones del modelo

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + u.$$

Si *prpblck* aumenta en .20 (20 puntos porcentuales), ¿cuál es el cambio porcentual estimado para *psoda*? (*Sugerencia*: la respuesta es 2.*xx*, usted tiene que dar las "*xx*".)

v) Agregue ahora la variable *prppov* (proporción de personas en pobreza) a la regresión del inciso iv). ¿Qué pasa con $\hat{\beta}_{prpblck}$?

vi) Encuentre la correlación entre log(*income*) y *prppov*. ¿Es aproximadamente lo que esperaba?

vii) Evalúe la afirmación siguiente: "como log(*income*) y *prppov* están fuertemente correlacionadas, no tiene caso que estén en la misma regresión".

C3.9 Use los datos del archivo CHARITY.RAW para responder a las preguntas siguientes:

i) Estime la ecuación

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

mediante MCO y dé el resultado en la forma usual, incluyendo el tamaño de la muestra y la *R*-cuadrada. Compare la *R*-cuadrada con la de la regresión simple en la que se omite *giftlast* (monto de la donación más reciente) y *propresp* (tasa de respuesta). (Vea el ejercicio (2.7).

ii) Interprete el coeficiente de *mailsyear*. ¿Es mayor o menor que el coeficiente correspondiente en la regresión simple?

iii) Interprete el coeficiente de *propresp*. Tenga cuidado con las unidades de medición de *propresp*.

iv) Ahora agregue a la ecuación la variable *avggift*. ¿Qué pasa con el efecto estimado de *mailsyear*?

v) ¿Qué ha pasado con el coeficiente de *giftlast* en la ecuación del inciso iv)? ¿Qué cree que esté pasando?

Apéndice 3A

3A.1 Obtención de las condiciones de primer orden de la ecuación (3.13)

Este análisis es muy semejante al análisis en el caso de la regresión simple. Se necesita caracterizar las soluciones del problema

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1x_{i1} - \ldots - b_kx_{ik})^2.$$

Obteniendo las derivadas parciales con respecto a cada una de las b_j (vea el apéndice A), evaluándolas en las soluciones e igualándolas a cero se obtiene

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik}) = 0$$

$$-2\sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik}) = 0, \quad \text{para toda } j = 1, \ldots, k.$$

Eliminando el -2 se obtienen las condiciones de primer orden dadas en (3.13).

3A.2 Obtención de la ecuación (3.22)

Para obtener (3.22), se expresa x_{i1} en términos de su valor ajustado y de su residual en la regresión de x_1 sobre $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, para toda $i = 1, \ldots, n$. Ahora, esto se sustituye en la segunda ecuación dada en (3.13):

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik}) = 0. \tag{3.60}$$

De acuerdo con la definición de los residuales $\hat{u}_i$, de MCO, como $\hat{x}_{i1}$ es una función lineal de las variables explicativas $x_{i2}, \ldots, x_{ik}$, se sigue que $\sum_{i=1}^{n}\hat{x}_{i1}\hat{u}_i = 0$. Por tanto, la ecuación (3.60) se puede expresar como

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \ldots - \hat{\beta}_kx_{ik}) = 0. \tag{3.61}$$

Como las $\hat{r}_{i1}$ son los residuales de la regresión de x_1 sobre $x_2, \ldots, x_k$, $\sum_{i=1}^{n}x_{ij}\hat{r}_{i1} = 0$, para toda $j = 2, \ldots, k$. Por lo tanto, (3.61) es equivalente a $\sum_{i=1}^{n}\hat{r}_{i1}(y_i - \hat{\beta}_1x_{i1}) = 0$. Por último, se usa el hecho de que $\sum_{i=1}^{n}\hat{x}_{i1}\hat{r}_{i1} = 0$, lo que significa que $\hat{\beta}_1$ es solución de

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1\hat{r}_{i1}) = 0.$$

Ahora, mediante manipulaciones algebraicas sencillas se obtiene (3.22), siempre que, por supuesto $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; esto queda asegurado por el supuesto RLM.3.

3A.3 Demostración del teorema 3.1

El teorema 3.1 se demuestra para $\hat{\beta}_1$; la demostración para el resto de los parámetros de pendiente es casi idéntica (Vea el apéndice E para una demostración más concreta empleando matrices). Bajo el supuesto RLM.3 los estimadores de MCO existen y $\hat{\beta}_1$ se puede escribir como en (3.22). Bajo el supuesto RLM.1 y_i puede expresarse como en (3.32); sustituya esta última ecuación en la y_i de (3.22). Después, usando $\sum_{i=1}^{n} \hat{r}_{i1} = 0$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$, para toda $j = 2, \ldots, k$ y $\sum_{i=1}^{n} x_{i1}\hat{r}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}^2$, se obtiene

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}u_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \tag{3.62}$$

Ahora, bajo los supuestos RLM.2 y RLM.4, el valor esperado de cada u_i, dadas todas las variables independientes de la muestra, es cero. Como las $\hat{r}_{i1}$ son funciones sólo de las variables independientes, se sigue que

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1|\mathbf{X}) &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\mathrm{E}(u_i|\mathbf{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) \\ &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} \cdot 0\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) = \beta_1, \end{aligned}$$

donde $\mathbf{X}$ denota los datos sobre todas las variables independientes y $\mathrm{E}(\hat{\beta}_1|\mathbf{X})$ es el valor esperado de $\hat{\beta}_1$, dados $x_{i1}, \ldots, x_{ik}$, para toda $i = 1, \ldots, n$. Esto completa la demostración.

3A.4 Sesgo general por variable omitida

El sesgo por variable omitida en el modelo general, ecuación (3.31), puede obtenerse a partir de los cuatro primeros supuestos de Gauss-Markov. En particular, sean $\hat{\beta}_j$, $j = 0, 1, \ldots, k$ los estimadores de MCO de la regresión empleando el conjunto completo de variables explicativas. Sean los $\tilde{\beta}_j$, $j = 0, 1, \ldots, k-1$ los estimadores de MCO de la regresión que deja fuera x_k. Sea $\tilde{\delta}_j$, $j = 1, \ldots, k-1$ el coeficiente de pendiente de x_j en la regresión auxiliar de x_{ik} sobre x_{i1}, $x_{i2}, \ldots x_{i,k-1}$, $i = 1, \ldots, n$. Un hecho útil es que

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j. \tag{3.63}$$

Esto muestra de manera explícita que, cuando no se controla x_k en la regresión, el efecto parcial estimado de x_j es igual al efecto parcial cuando se incluye x_k más el efecto parcial de x_k sobre $\hat{y}$ multiplicado por la relación parcial entre la variable omitida, x_k y x_j, $j < k$. Condicional sobre todo el conjunto de las variables explicativas, $\mathbf{X}$, se sabe que todas las $\hat{\beta}_j$ son insesgadas para su correspondiente β_j, $j = 1, \ldots, k$. Además, como $\tilde{\delta}_j$ sólo es función de $\mathbf{X}$, se tiene

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_j|\mathbf{X}) &= \mathrm{E}(\hat{\beta}_j|\mathbf{X}) + \mathrm{E}(\hat{\beta}_k|\mathbf{X})\tilde{\delta}_j \\ &= \beta_j + \beta_k\tilde{\delta}_j. \end{aligned} \tag{3.64}$$

La ecuación (3.64) muestra que $\tilde{\beta}_j$ es sesgada respecto a β_j a menos que $\beta_k = 0$ —en cuyo caso x_k no tiene efecto parcial en la población— o $\tilde{\delta}_j$ sea igual a cero, lo que significa que x_{ik} y x_{ij} no están parcialmente correlacionadas en la muestra. La clave para obtener la ecuación (3.64) es la ecuación (3.63). Para mostrar la ecuación (3.63), se puede usar la ecuación (3.22) un par de veces. Para simplificar se considerará $j = 1$. Ahora, $\tilde{\beta}_1$ es el coeficiente de pendiente en la regresión simple de y_i sobre $\tilde{r}_{i1}$, $i = 1, \ldots, n$, donde las $\tilde{r}_{i1}$ son los residuales de MCO de la regresión de x_{i1} sobre $x_{i2}, x_{i3}, \ldots, x_{i,k-1}$. Considérese el numerador de la expresión para $\tilde{\beta}_1$: $\sum_{i=1}^{n} \tilde{r}_{i1} y_i$. Pero para cada i se puede escribir $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$ y sustituir y_i en la expresión anterior. Entonces, de acuerdo con las propiedades de los residuales de MCO, los $\tilde{r}_{i1}$ tienen media muestral cero y no están correlacionados en la muestra con $x_{i2}, x_{i3}, \ldots, x_{i,k-1}$. De manera similar, los $\hat{u}_i$ tienen media muestral cero y correlación muestral cero con $x_{i1}, x_{i2}, \ldots, x_{ik}$. Se sigue que los $\tilde{r}_{i1}$ y los $\hat{u}_i$ no están correlacionadas en la muestra (ya que los $\tilde{r}_{i1}$ son combinaciones lineales sólo de $x_{i1}, x_{i2}, \ldots, x_{i,k-1}$). De manera que

$$\sum_{i=1}^{n} \tilde{r}_{i1} y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} \right) + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right). \tag{3.65}$$

Ahora, $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$, que es también el denominador de $\tilde{\beta}_1$. Por tanto, se ha mostrado que

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right) \Bigg/ \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2 \right), \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1. \end{aligned}$$

Esta es la relación que se quería demostrar.

3A.5 Demostración del teorema 3.2

Una vez más, se demuestra para $j = 1$. Se expresa $\hat{\beta}_1$ como en la ecuación (3.62). Ahora, bajo RLM.5, $\mathrm{Var}(u_i|\mathbf{X}) = \sigma^2$, para toda $i = 1, \ldots, n$. Bajo muestreo aleatorio, las u_i son independientes, incluso condicionales sobre $\mathbf{X}$, y las $\hat{r}_{i1}$ son no aleatorias condicionales en $\mathbf{X}$. Por tanto,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1|\mathbf{X}) &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \ \mathrm{Var}(u_i|\mathbf{X}) \right) \Bigg/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 \\ &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \sigma^2 \right) \Bigg/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 = \sigma^2 \Bigg/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \end{aligned}$$

Ahora, como $\sum_{i=1}^{n} \hat{r}_{i1}^2$ es la suma de los residuales cuadrados de la regresión de x_1 sobre $x_2, \ldots, x_k$, $\sum_{i=1}^{n} \hat{r}_{i1}^2 = \mathrm{STC}_1(1 - R_1^2)$. Con lo que queda terminada la demostración.

3A.6 Demostración del teorema 3.4

Se muestra que, para cualquier otro estimador lineal insesgado $\tilde{\beta}_1$ de β_1, $\mathrm{Var}(\tilde{\beta}_1) \geq \mathrm{Var}(\hat{\beta}_1)$, donde $\hat{\beta}_1$ es el estimador de MCO. La demostración se hace para $j = 1$ sin pérdida de generalidad.

Dado $\tilde{\beta}_1$ como en la ecuación (3.59), se puede sustituir y_i y obtener

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} u_i.$$

Ahora, como las w_{i1} son funciones de x_{ij},

$$\mathrm{E}(\tilde{\beta}_1|\mathbf{X}) = \beta_0\sum_{i=1}^{n} w_{i1} + \beta_1\sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2\sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k\sum_{i=1}^{n} w_{i1}x_{ik} + \sum_{i=1}^{n} w_{i1}\mathrm{E}(u_i|\mathbf{X})$$

$$= \beta_0\sum_{i=1}^{n} w_{i1} + \beta_1\sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2\sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k\sum_{i=1}^{n} w_{i1}x_{ik}$$

dado que $\mathrm{E}(u_i|\mathbf{X}) = 0$, para todo $i = 1, \ldots, n$ bajo RLM.2 y RLM.4. Por lo tanto, para que $\mathrm{E}(\tilde{\beta}_1|\mathbf{X})$ sea igual a β_1 para cualquier valor de los parámetros, se necesita tener

$$\sum_{i=1}^{n} w_{i1} = 0, \sum_{i=1}^{n} w_{i1}x_{i1} = 1, \sum_{i=1}^{n} w_{i1}x_{ij} = 0, \quad j = 2, \ldots, k. \tag{3.66}$$

Ahora, sean $\hat{r}_{i1}$ los residuales de la regresión de x_{i1} sobre $x_{i2}, \ldots, x_{ik}$. Entonces, de acuerdo con (3.66), se sigue que

$$\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} = 1 \tag{3.67}$$

ya que $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ y $\sum_{i=1}^{n} w_{i1}\hat{x}_{i1} = 0$. Ahora, considérese la diferencia entre $\mathrm{Var}(\tilde{\beta}_1|\mathbf{X})$ y $\mathrm{Var}(\hat{\beta}_1|\mathbf{X})$ bajo RLM.1 a RLM.5:

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \tag{3.68}$$

Dado (3.67), la diferencia en (3.68) puede escribirse sin σ^2, como

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1}\right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \tag{3.69}$$

Pero (3.69) es simplemente

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1\hat{r}_{i1})^2, \tag{3.70}$$

donde $\hat{\gamma}_1 = \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1}\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)$, como puede verse al elevar al cuadrado cada término de (3.70), sumar y después cancelar términos. Como (3.70) es la suma de los residuales cuadrados de la regresión simple de w_{i1} sobre $\hat{r}_{i1}$ —recuerde que el promedio muestral de los $\hat{r}_{i1}$ es cero— (3.70) puede ser no negativa. Esto completa la demostración.

CAPÍTULO 4

Análisis de regresión múltiple: inferencia

En este capítulo se continúa con el estudio del análisis de regresión múltiple. Ahora la atención se enfoca hacia las pruebas de hipótesis acerca de los parámetros del modelo de regresión poblacional. Se empieza por determinar la distribución de los estimadores de MCO bajo el supuesto adicional de que el error poblacional está distribuido normalmente. En las secciones 4.2 y 4.3 se ven las pruebas de hipótesis acerca de parámetros individuales y en la sección 4.4 se analiza cómo probar una sola hipótesis respecto a más de un parámetro. En la sección 4.5 se ve la manera de probar restricciones múltiples y en especial cómo determinar si un grupo de variables independientes puede ser omitido del modelo.

4.1 Distribución de muestreo de los estimadores de MCO

Hasta ahora se ha formado un conjunto de supuestos bajo los cuales MCO es insesgado; también se han obtenido y discutido los sesgos ocasionados por las variables omitidas. En la sección 3.4 se obtuvieron las varianzas de los estimadores de MCO empleando los supuestos de Gauss-Markov. En la sección 3.5 se mostró que esta varianza es la menor entre los estimadores lineales insesgados.

Conocer el valor esperado y la varianza de los estimadores de MCO sirve para describir la precisión de los estimadores de MCO. Sin embargo, para realizar una inferencia estadística no sólo se necesita conocer los dos primeros momentos de $\hat{\beta}_j$; se requiere conocer toda la distribución muestral de los $\hat{\beta}_j$. Aun bajo los supuestos de Gauss-Markov, la distribución de $\hat{\beta}_j$ puede tener casi cualquier forma.

Cuando se condiciona sobre los valores de las variables independientes en la muestra, es claro que la distribución muestral de los estimadores de MCO depende de la distribución subyacente de los errores. Para hacer manejables las distribuciones de muestreo de los $\hat{\beta}_j$, ahora se supondrá que en la población el error no observado está *distribuido normalmente*. Esto se conoce como **supuesto de normalidad**.

Supuesto RLM.6 (Normalidad)

El error poblacional u es *independiente* de las variables explicativas $x_1, x_2, \ldots, x_k$ y está distribuido normalmente, con media cero y varianza σ^2: $u \sim \text{Normal}(0,\sigma^2)$.

El supuesto RLM.6 es mucho más fuerte que cualesquiera de los supuestos anteriores. En realidad, como bajo RLM.6 u es independiente de las x_j, $\text{E}(u|x_1, \ldots, x_k) = \text{E}(u) = 0$ y $\text{Var}(u|x_1, \ldots, x_k) = \text{Var}(u) = \sigma^2$. De manera que con RLM.6 se están suponiendo necesariamente RLM.1 a RLM.5. Para subrayar que se está suponiendo más que antes, nos referiremos al conjunto completo de los supuestos RLM.1 a RLM.6.

En aplicaciones de la regresión a cortes transversales, a los supuestos RLM.1 a RLM.6 se les conoce como **supuestos del modelo lineal clásico (MLC)**. De manera que al modelo con estos seis supuestos se le llama **modelo lineal clásico**. Lo mejor es considerar que los supuestos del MLC contienen todos los supuestos de Gauss-Markov *más* el supuesto de un término del error distribuido normalmente.

Bajo los supuestos del MLC, los estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ de MCO tienen una propiedad de eficiencia más fuerte que bajo los supuestos de Gauss-Markov. Puede demostrarse que los estimadores de MCO son los **estimadores insesgados de varianza mínima**, lo que significa que MCO tiene la menor varianza entre los estimadores insesgados; no es ya necesario restringir la comparación a estimadores lineales en las y_i. En el apéndice E se estudia esta propiedad de los MCO bajo los supuestos del MLC.

Una manera sucinta de resumir los supuestos poblacionales del MLC es

$$y|\mathbf{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2),$$

donde $\mathbf{x}$ es, una vez más, una manera abreviada de escribir $(x_1, \ldots, x_k)$. De manera que, condiciona en $\mathbf{x}$, y tiene una distribución normal cuya media es lineal en $x_1, \ldots, x_k$ y cuya varianza es constante. En la figura 4.1 se muestra esta situación en el caso de una sola variable independiente.

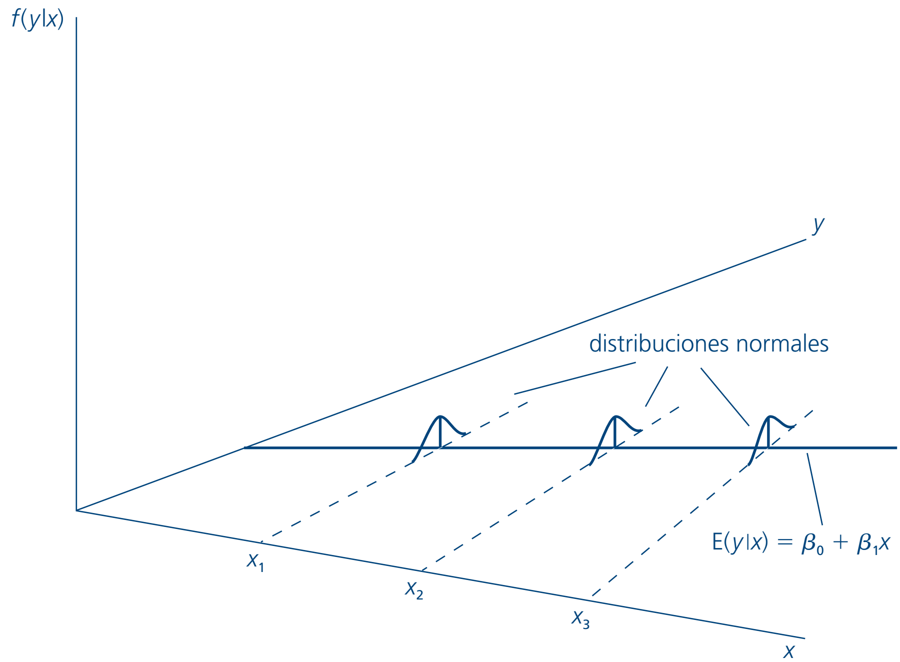
El argumento que justifica la distribución normal de los errores es más o menos como esto: como u es la suma de muchos factores distintos no observados que afectan a y, se puede apelar al teorema del límite central (vea el apéndice C) para concluir que u tiene una distribución aproximadamente normal. Este argumento tiene cierto mérito, pero no carece de debilidades. Primero, los factores en u pueden tener distribuciones muy diferentes en la población (por ejemplo capacidad innata y calidad de la escolaridad en el error de una ecuación de salario). Aunque el teorema del límite central (TLC) puede satisfacerse en tales casos, la aproximación normal puede ser mala dependiendo de cuántos factores aparezcan en u y de qué tan diferentes sean sus distribuciones.

Otro problema serio con el argumento del TLC es que éste supone que todos los factores no observados afectan a y por separado en forma aditiva. Nada garantiza que esto sea así. Si u es una complicada función de los factores no observados, entonces el argumento del TLC no puede emplearse.

Que en una aplicación pueda suponerse normalidad de u es en realidad una cuestión empírica. Por ejemplo, no hay un teorema que diga que *wage* (salario) condicional en *educ*, *exper* y *tenure* (antigüedad) esté distribuido normalmente. En todo caso, un razonamiento sencillo

FIGURA 4.1

Distribución normal homocedástica con una sola variable explicativa



sugiere que en realidad ocurre lo contrario: como *wage* (salario) nunca puede ser menor a cero no puede tener, estrictamente hablando, una distribución normal. Además, como existen leyes para el salario mínimo, una fracción de la población gana exactamente el salario mínimo, lo cual viola también el supuesto de normalidad. Sin embargo, en la práctica, puede uno preguntarse si la distribución condicional del salario se "aproxima" a la normalidad. Evidencias empíricas anteriores indican que la normalidad *no* es un buen supuesto en el caso del salario.

Con frecuencia, usando una transformación, en especial empleando el logaritmo, se obtiene una distribución que es cercana a la normal. Por ejemplo, log(*price*) tiende a tener una distribución que parece más normal que la de *price* (precio). Una vez más, esta es una cuestión empírica. En el capítulo 5 se discutirán las consecuencias de la falta de normalidad en la inferencia estadística.

Existen algunos ejemplos en los que RLM.6 es falso. Siempre que *y* tome sólo unos cuantos valores no podrá tener ninguna distribución que se acerque a una normal. La variable dependiente del ejemplo 3.5 proporciona un buen ejemplo. La variable *narr86*, cantidad de veces que un hombre joven fue arrestado en 1986, toma un rango pequeño de valores enteros y para la mayoría de los hombres toma el valor cero. De manera que, *narr86* está muy lejos de tener una distribución normal. ¿Qué puede hacerse en estos casos? Como se verá en el capítulo 5 —y esto es importante— el que no haya normalidad en los errores no es un problema serio cuando los tamaños de muestra son grandes. Por ahora, sólo se hace el supuesto de normalidad.

La normalidad del término de error se traduce en una distribución de muestreo normal de los estimadores de MCO:

Teorema 4.1 (Distribuciones de muestreo normales)

Bajo los supuestos RLM.1 a RLM.6 del MLC, condicionalmente en los valores muestrales de las variables independientes,

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)], \tag{4.1}$$

donde $\text{Var}(\hat{\beta}_j)$ fue dada en el capítulo 3 [ecuación (3.51)]. Por lo tanto,

$$(\hat{\beta}_j - \beta_j)/\text{de}(\hat{\beta}_j) \sim \text{Normal}(0,1).$$

La demostración de (4.1) no es difícil, dadas las propiedades de las variables aleatorias normalmente distribuidas del apéndice B. Cada $\hat{\beta}_j$ puede escribirse como $\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij}u_i$, donde $w_{ij} = \hat{r}_{ij}/\text{SRC}_j$, $\hat{r}_{ij}$ es el residual i-ésimo de la regresión de x_j sobre todas de las demás variables independientes y SRC_j es la suma de los residuales cuadrados de esta regresión [vea la ecuación (3.62)]. Dado que los w_{ij} dependen sólo de las variables independientes, pueden tratarse como no aleatorios. Entonces, $\hat{\beta}_j$ es sólo una combinación lineal de los errores en la muestra $\{u_i: i = 1, 2, \ldots, n\}$. Bajo el supuesto RLM.6 (y el RLM.2 sobre muestreo aleatorio), los errores son variables aleatorias independientes e idénticamente distribuidas como Normal$(0,\sigma^2)$. Un hecho importante acerca de las variables aleatorias normales independientes es que una combinación lineal de éstas tiene una distribución normal (vea el apéndice B). Con esto termina básicamente la demostración. En la sección 3.3 se mostró que $\text{E}(\hat{\beta}_j) = \beta_j$ y en la sección 3.4 se obtuvo $\text{Var}(\hat{\beta}_j)$; no es necesario volver a deducir estos hechos.

Pregunta 4.1

Suponga que u es independiente de las variables explicativas y que toma los valores -2, -1, 0, 1 y 2 todos con probabilidad de 1/5. ¿Viola esto los supuestos de Gauss-Markov? ¿Viola esto los supuestos del MLC?

La segunda parte de este teorema sigue de inmediato del hecho de que cuando se estandariza una variable aleatoria normal restándole su media y dividiendo el resultado entre su desviación estándar, se obtiene una variable aleatoria normal estándar.

La conclusión del teorema 4.1 puede fortalecerse. Además de (4.1), cualquier combinación lineal de los $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ tiene también una distribución normal y cualquier subconjunto de los $\hat{\beta}_j$ tiene una distribución *conjunta* normal. Estos hechos subyacen a los resultados de las pruebas en el resto del capítulo. En el capítulo 5 se mostrará que la normalidad de los estimadores de MCO es *aproximadamente* verdadera en muestras grandes aun sin normalidad de los errores.

4.2 Prueba de hipótesis sobre un solo parámetro poblacional: la prueba *t*

Esta sección cubre el importante tema de las pruebas de hipótesis para un solo parámetro de la función de regresión poblacional. El modelo poblacional puede escribirse como

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \tag{4.2}$$

y se supone que este modelo satisface los supuestos del MLC. Se sabe que MCO produce estimadores insesgados de las β_j. En esta sección se estudia cómo probar hipótesis acerca de una β_j en particular. Para una comprensión completa de las pruebas de hipótesis, hay que recordar que las β_j son características desconocidas de la población, que nunca se conocerán con certeza. Aun así, pueden hacerse *hipótesis* acerca del valor de β_j y después usar la inferencia estadística para probarlas.

Para construir las pruebas de hipótesis, se necesita el resultado siguiente:

Teorema 4.2 (Distribución *t* para estimadores estandarizados)

Bajo los supuestos RLM.1 a RLM.6 del MLC,

$$(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j) \sim t_{n-k-1}, \quad \textbf{4.3}$$

donde $k + 1$ es la cantidad de parámetros desconocidos en el modelo poblacional $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ (k parámetros de pendiente y el intercepto β_0).

Este resultado difiere del teorema 4.1 en algunos aspectos importantes. El teorema 4.1 muestra que, bajo los supuestos del MLC, $(\hat{\beta}_j - \beta_j)/\text{de}(\hat{\beta}_j) \sim \text{Normal}(0,1)$. La distribución t en (4.3) proviene del hecho de que la constante σ en $\text{de}(\hat{\beta}_j)$ ha sido sustituida por la variable aleatoria $\hat{\sigma}$. La demostración de que esto lleva a una distribución t con $n - k - 1$ grados de libertad, no es en especial intuitiva. En esencia, la demostración muestra que (4.3) puede escribirse como el cociente de la variable aleatoria normal estándar $(\hat{\beta}_j - \beta_j)/\text{de}(\hat{\beta}_j)$ entre la raíz cuadrada de $\hat{\sigma}^2/\sigma^2$. Se puede demostrar que estas variables aleatorias son independientes y que $(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. El resultado se sigue de la definición de una variable aleatoria t (ver sección B.5).

El teorema 4.2 es importante porque permite probar hipótesis en las que intervienen las β_j. En la mayoría de las aplicaciones, el interés principal reside en probar la **hipótesis nula**

$$\text{H}_0\text{: } \beta_j = 0, \quad \textbf{4.4}$$

donde j corresponde a cualquiera de las k variables independientes. Cuando se trata de una aplicación específica, es importante entender lo que (4.4) significa y poder describir esta hipótesis en un lenguaje sencillo. Dado que β_j mide el efecto parcial de x_j sobre (el valor esperado de) y, una vez que se han controlado todas las demás variables independientes, (4.4) significa que, una vez que $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ han sido tomadas en cuenta, x_j no tiene *ningún efecto* sobre el valor esperado de y. La hipótesis nula no puede enunciarse como "x_j sí tiene un efecto parcial sobre y" porque esto es verdad para cualquier valor de β_j distinto de cero. El método clásico de prueba es adecuado para probar *hipótesis sencillas* como (4.4).

Como ejemplo, considere la ecuación de salario (*wage*)

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

La hipótesis nula $\text{H}_0\text{: } \beta_2 = 0$ significa que una vez que la educación y la antigüedad (*tenure*) han sido tomadas en cuenta, la cantidad de años en la fuerza de trabajo (*exper*) no tiene efecto sobre el salario por hora. Esta es una hipótesis económicamente interesante. Si es así, esto implica que la historia laboral de una persona, antes del empleo actual, no afecta al salario. Si $\beta_2 > 0$, la experiencia laboral anterior contribuye a la productividad y por tanto al salario.

Tal vez usted recuerde, de sus cursos de estadística, los principios básicos de la prueba de hipótesis para la media de una población normal. (En el apéndice C se encuentra un repaso de esto.) La mecánica para probar (4.4) en el contexto de la regresión múltiple es muy parecida. La parte difícil es obtener las estimaciones de los coeficientes, los errores estándar y los valores críticos, pero la mayor parte de esta es realizada de manera automática por el software para econometría. Nuestra tarea es saber cómo puede usarse el resultado de la regresión para probar la hipótesis de interés.

El estadístico que se emplea para probar (4.4) (contra cualquier alternativa se llama "el" **estadístico *t*** o "el" **coeficiente *t*** de $\hat{\beta}_j$ y se define como

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j/\text{ee}(\hat{\beta}_j). \tag{4.5}$$

Se ha puesto "el" entre comillas porque, como se verá en breve, para probar otras hipótesis acerca de β_j se necesita una forma más general del estadístico t. Por ahora, es importante saber que (4.5) sólo es adecuada para probar (4.4). Para aplicaciones específicas, es útil usar el nombre de la variable independiente como subíndice del estadístico t; por ejemplo, t_{educ} será el estadístico t correspondiente a $\hat{\beta}_{educ}$.

El estadístico t correspondiente a $\hat{\beta}_j$ es fácil de calcular dados $\hat{\beta}_j$ y su error estándar. En realidad, la mayoría de los paquetes para regresión realiza esta división y dan el estadístico t junto con cada coeficiente y su error estándar.

Antes de analizar cómo usar de manera formal (4.5) para probar H$_0$: $\beta_j = 0$, es útil ver por qué $t_{\hat{\beta}_j}$ tiene características que lo hacen un estadístico razonable de prueba para detectar $\beta_j \neq 0$. Primero, como ee$(\hat{\beta}_j)$ es siempre positivo, $t_{\hat{\beta}_j}$ tiene el mismo signo que $\hat{\beta}_j$: si $\hat{\beta}_j$ es positivo, también lo es $t_{\hat{\beta}_j}$ y si $\hat{\beta}_j$ es negativo, también lo es $t_{\hat{\beta}_j}$. Segundo, para un valor dado de ee$(\hat{\beta}_j)$, un valor mayor de $\hat{\beta}_j$ lleva a un valor mayor de $t_{\hat{\beta}_j}$. Si $\hat{\beta}_j$ se vuelve más negativo, lo mismo ocurre con $t_{\hat{\beta}_j}$.

Como se está probando H$_0$: $\beta_j = 0$, resulta natural fijarse en el estimador insesgado de β_j, $\hat{\beta}_j$. En cualquier aplicación interesante, el estimador puntual $\hat{\beta}_j$ *nunca* será exactamente cero, sea o no verdadera H$_0$. La pregunta es: ¿qué tan lejos está $\hat{\beta}_j$ de cero? Si el valor muestral de $\hat{\beta}_j$ está muy lejos de cero, esto proporciona una evidencia contra H$_0$: $\beta_j = 0$. Sin embargo, hay que reconocer que existe un error de muestreo en la estimación $\hat{\beta}_j$, de manera que la magnitud de $\hat{\beta}_j$ debe ser ponderada contra su error de muestreo. Como el error estándar de $\hat{\beta}_j$ es una estimación de la desviación estándar de $\hat{\beta}_j$, $t_{\hat{\beta}_j}$ mide a cuántas desviaciones estándar estimadas se encuentra $\hat{\beta}_j$ de cero. Esto es lo que se hace cuando se prueba si la media de una población es cero, usando el estadístico t estándar de la introducción a la estadística. Si el valor de $t_{\hat{\beta}_j}$ se encuentra suficientemente lejos de cero se rechazará H$_0$. La regla exacta de rechazo depende de la hipótesis alternativa y del nivel de significancia elegido para la prueba.

Determinar una regla para rechazar (4.4) a un determinado nivel de significancia —es decir, la probabilidad de rechazar H$_0$ cuando es verdadera—, requiere que se conozca la distribución de muestreo de $t_{\hat{\beta}_j}$ cuando H$_0$ es verdadera. De acuerdo con el teorema 4.2, se sabe que ésta es t_{n-k-1}. Este es el resultado teórico clave para probar (4.4)

Antes de proceder, es importante recordar que se están probando hipótesis acerca de parámetros *poblacionales, no* hipótesis acerca de estimaciones a partir de una muestra determinada. Por tanto, no tiene sentido enunciar la hipótesis nula como "H$_0$: $\hat{\beta}_1 = 0$" o, peor aún, como "H$_0$: $.237 = 0$", si la estimación de un parámetro es .237 en la muestra. Lo que se está probando es si el valor poblacional desconocido, β_1, es cero.

En algunos tratamientos del análisis de regresión se define el estadístico t como el *valor absoluto* de (4.5), de manera que el estadístico t sea siempre positivo. Esto tiene la desventaja de hacer que la prueba contra una alternativa unilateral sea torpe. En este libro el estadístico t tiene siempre el mismo signo que la estimación correspondiente del coeficiente de MCO.

Pruebas contra alternativas de una cola

Para determinar una regla para rechazar H_0, hay que decidir sobre la **hipótesis alternativa** relevante. Primero, considere una **alternativa de una cola** de la forma

$$H_1\colon \beta_j > 0. \tag{4.6}$$

Esto significa que no interesan alternativas a H_0 que sean de la forma $H_1\colon \beta_j < 0$; por alguna razón, quizá con base en la introspección de una teoría económica, se excluyen valores poblacionales de β_j menores a cero. (Otra manera de ver esto es que la hipótesis nula sea en realidad $H_0\colon \beta_j \leq 0$; en cualquier caso, el estadístico $t_{\hat{\beta}_j}$ se usa como estadístico de prueba).

¿Cómo debe elegirse la regla de rechazo? Primero hay que decidirse por un **nivel de significancia** o probabilidad de rechazar H_0 cuando en realidad es verdadera. Para concretar, suponga que se ha decidido usar 5% como nivel de significancia, ya que este es el nivel más empleado. Por tanto, se estará dispuesto a rechazar de manera equivocada H_0 siendo ésta verdadera 5% de las veces. Ahora, mientras que $t_{\hat{\beta}_j}$ tiene una distribución t bajo H_0 —de manera que tiene media cero— bajo la alternativa $\beta_j > 0$, el valor esperado de $t_{\hat{\beta}_j}$ es positivo. Por tanto, se está buscando un valor positivo "suficientemente grande" de $t_{\hat{\beta}_j}$ que permita rechazar $H_0\colon \beta_j = 0$ a favor de $H_1\colon \beta_j > 0$. Los valores negativos de $t_{\hat{\beta}_j}$ no proporcionan evidencia a favor de H_1.

La definición de "suficientemente grande", con un nivel de significancia de 5%, es el percentil 95 de una distribución t con $n - k - 1$ grados de libertad; denótese c. En otras palabras, la **regla de rechazo** es que al nivel de significancia de 5% se rechaza H_0 a favor de H_1 si

$$t_{\hat{\beta}_j} > c. \tag{4.7}$$

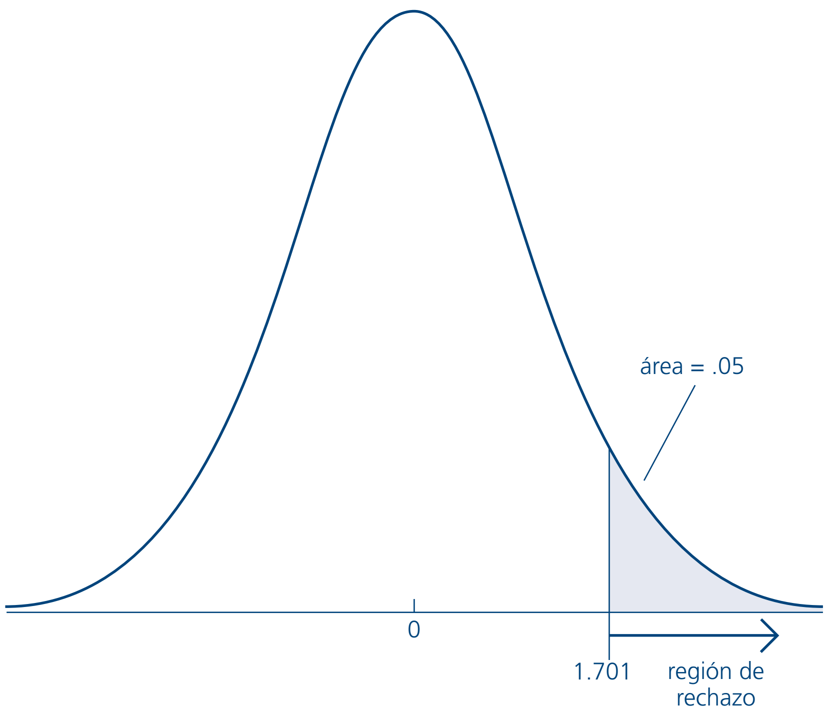
De acuerdo con la elección del **valor crítico** c, el rechazo de H_0 ocurrirá en 5% de todas las muestras aleatorias cuando H_0 sea verdadera.

La regla de rechazo en (4.7) es un ejemplo de una **prueba de una cola**. Para obtener c, sólo se necesitan el nivel de significancia y los grados de libertad. Por ejemplo, para una prueba al nivel de 5% con $n - k - 1 = 28$ grados de libertad, el valor crítico es $c = 1.701$. Si $t_{\hat{\beta}_j} < 1.701$, no se rechaza H_0 a favor de (4.6) al nivel de 5%. Observe que si $t_{\hat{\beta}_j}$ tiene un valor negativo, no importa qué tan grande en valor absoluto, esto lleva a no rechazar H_0 a favor de (4.6). (Ver figura 4.2.)

Este mismo procedimiento puede emplearse con otros niveles de significancia. En una prueba al nivel de 10% y si $gl = 21$, el valor crítico es $c = 1.323$. Si el nivel de significancia es de 1% y $gl = 21$, $c = 2.518$. Todos estos valores críticos fueron obtenidos directamente de la tabla G.2. Debe observarse el patrón que siguen los valores críticos: a medida que el nivel de significancia disminuye, el valor crítico aumenta, de manera que se requiere un valor cada vez más grande de $t_{\hat{\beta}_j}$ para rechazar H_0. Por lo tanto, si se rechaza H_0 por ejemplo al nivel de 5%, estará automáticamente rechazada también al nivel de 10%. No tiene sentido rechazar la hipó-

FIGURA 4.2

Regla de rechazo de 5% para la alternativa H_1: $\beta_j > 0$ con 28 *gl*.



tesis nula, por ejemplo, al nivel de 5%, y después repetir la prueba para determinar el resultado al nivel de 10%.

A medida que aumentan los grados de libertad de la distribución t, ésta se aproxima a la distribución normal estándar. Por ejemplo, cuando $n - k - 1 = 120$, el valor crítico de 5% para la alternativa de una cola (4.7) es 1.658, y el valor normal estándar es 1.645. Estos valores están suficientemente cercanos para fines prácticos; para grados de libertad mayor a 120, pueden emplearse los valores críticos de la normal estándar.

Ejemplo 4.1

[Ecuación del salario por hora]

Con ayuda de los datos del archivo WAGE1.RAW se obtiene la ecuación estimada

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(.104)}{.284} + \underset{(.007)}{.092}\ educ + \underset{(.0017)}{.0041}\ exper + \underset{(.003)}{.022}\ tenure \\ n &= 526,\ R^2 = .316,\end{aligned}$$

donde los errores estándar aparecen entre paréntesis debajo de los coeficientes estimados. Tal convención se seguirá en todo el libro. Esta ecuación puede emplearse para probar si en la población el rendimiento de *exper*, controlando *educ* y *tenure* es cero contra la alternativa de que este rendimiento sea positivo. Esto se escribe como H_0: $\beta_{exper} = 0$ versus H_1: $\beta_{exper} > 0$. (En las aplicaciones, emplear como índices de los parámetros los nombres de las variables correspondientes es una buena manera de identificar los parámetros, pues los índices numéricos usados en el modelo general son arbitrarios y pueden causar confusión.) Recuerde que β_{exper} denota el parámetro poblacional desconocido. Escribir "H_0: $.0041 = 0$" o "H_0: $\hat{\beta}_{exper} = 0$" no tiene sentido.

Como se tienen 522 grados de libertad, se pueden emplear los valores críticos de la distribución normal estándar. El valor crítico para 5% es 1.645 y para 1% es 2.326. El estadístico t para $\hat{\beta}_{exper}$ es

$$t_{exper} = .0041/.0017 \approx 2.41,$$

y de esta manera $\hat{\beta}_{exper}$, o *exper*, es estadísticamente significativo aun al nivel 1%. Se dice también que "$\hat{\beta}_{exper}$ es estadísticamente mayor que cero al nivel de significancia 1%".

El rendimiento de un año más de experiencia, manteniendo antigüedad y educación constantes, no es en especial grande. Por ejemplo, con tres años más log(*wage*) aumenta a 3(.0041) = .0123, de manera que el salario es sólo 1.2% más alto. De cualquier manera, se ha mostrado convincentemente que, en la población, el efecto parcial de la experiencia *es* positivo.

La alternativa de una cola de que el parámetro sea menor que cero,

$$H_1: \beta_j < 0, \tag{4.8}$$

también suele surgir en las aplicaciones. La regla de rechazo para la alternativa (4.8) es con exactitud la imagen en el espejo del caso anterior. Ahora, el valor crítico se encuentra en la cola izquierda de la distribución t. En la práctica es más fácil considerar la regla de rechazo como

$$t_{\hat{\beta}_j} < -c, \tag{4.9}$$

donde c es el valor crítico para la alternativa H_1: $\beta_j > 0$. Para simplificar, se supone siempre que c es positivo, dado que así es como están dados los valores críticos en las tablas de la distribución t y de esta manera el valor crítico $-c$ es un número negativo.

Por ejemplo, si el nivel de significancia es 5% y los grados de libertad son 18, entonces $c = 1.734$ y al nivel de significancia 5% se rechaza H_0: $\beta_j = 0$ a favor de H_1: $\beta_j < 0$ si $t_{\hat{\beta}_j} < -1.734$. Es importante recordar que para rechazar H_0 contra la alternativa negativa (4.8), debe obtenerse un estadístico t negativo. Un cociente t positivo, no importa qué tan grande, no proporciona ninguna evidencia a favor de (4.8). En la figura 4.3 se ilustra esta regla de rechazo.

Pregunta 4.2

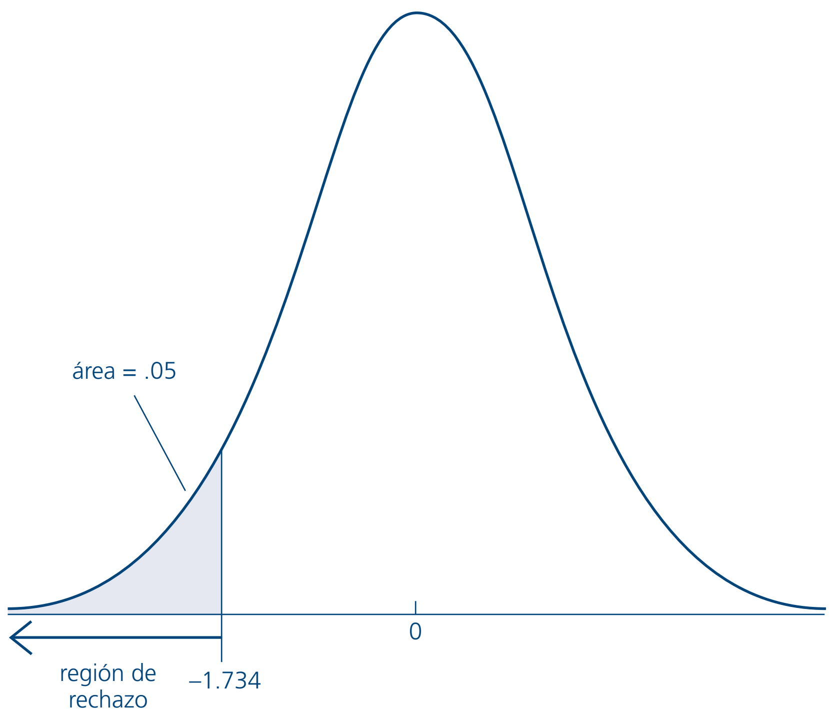
Suponga que las tasas de aprobación de préstamos a la comunidad (*apprate*) están determinadas por

$$apprate = \beta_0 + \beta_1 percmin + \beta_2 avginc + \beta_3 avgwlth + \beta_4 avgdebt + u,$$

donde *percmin* es el porcentaje de minorías en la comunidad, *avginc* es el ingreso promedio, *avgwlth* es la riqueza promedio y *avgdebt* es una medida del promedio de obligaciones de deuda. ¿Cómo establece usted la hipótesis nula de que entre barrios de distinta composición racial y étnica no hay diferencia en las tasas de préstamo, una vez controlados el ingreso promedio, la riqueza promedio y el promedio de obligaciones de deuda? ¿Cómo establece la hipótesis alternativa de que en las tasas de aprobación de préstamos sí hay discriminación contra las minorías?

FIGURA 4.3

Regla de rechazo de 5% para la alternativa H_1: $\beta_j < 0$ con 18 *gl*.



Ejemplo 4.2

[Desempeño de los estudiantes y tamaño de la escuela]

Existe gran interés acerca del efecto del tamaño de las escuelas en el desempeño de los estudiantes. (Vea, por ejemplo, *The New York Times Magazine*, 5/28/95.) Se cree que, siendo todo lo demás igual, los estudiantes de escuelas pequeñas obtienen mejores resultados que los de escuelas grandes. Esta hipótesis se supone cierta aun tomando en cuenta las diferencias de los tamaños de los grupos en las escuelas.

El archivo MEAP93.RAW contiene datos de 1993 sobre 408 escuelas de nivel bachillerato en Michigan. Estos datos pueden emplearse para probar la hipótesis nula de que el tamaño de la escuela no tiene efecto sobre las puntuaciones en los exámenes estandarizados contra la alternativa de que el tamaño tiene un efecto negativo. El desempeño se mide por el porcentaje de estudiantes que obtienen una puntuación aprobatoria en la prueba estandarizada de matemáticas del décimo grado (*math*10) en el programa de evaluación educativa de Michigan (MEAP, Michigan Educational Assessment Program). El tamaño de la escuela se mide por la cantidad de estudiantes inscritos (*enroll*). La hipótesis nula es H_0: $\beta_{enroll} = 0$ y la alternativa es H_1: $\beta_{enroll} < 0$. Por ahora, se controlarán dos factores, compensación promedio anual a los profesores (*totcomp*) y cantidad de personal por cada mil estudiantes (*staff*). La compensación a los profesores es una medida de la calidad de los mismos y *staff* es una medida aproximada de la atención que se les da a los estudiantes.

La ecuación estimada, con los errores estándar entre paréntesis, es

$$\begin{aligned}\widehat{math10} &= \underset{(6.113)}{2.274} + \underset{(.00010)}{.00046}\ totcomp + \underset{(.040)}{.048}\ staff - \underset{(.00022)}{.00020}\ enroll \\ n &= 408,\ R^2 = .0541.\end{aligned}$$

El coeficiente de *enroll*, $-.00020$, es acorde con la conjetura de que las escuelas más grandes dificultan el desempeño: mayor número de alumnos inscritos conduce a un porcentaje menor de estudiantes con una puntuación aprobatoria en el décimo grado de matemáticas. (También los coeficientes de *totcomp* y *staff* tienen el signo esperado.) El hecho de que *enroll* tenga un coeficiente distinto de cero podría deberse sólo a un error muestral; para convencerse de que existe algún efecto, se necesita realizar una prueba *t*.

Como $n - k - 1 = 408 - 4 = 404$, se emplea el valor crítico normal estándar. Al nivel de 5%, el valor crítico es -1.65; para rechazar H_0 al nivel de 5%, el estadístico *t* para *enroll* debe ser *menor* que -1.65.

El estadístico *t* para *enroll* es $-.00020/.00022 \approx -.91$, que es mayor que -1.65: al nivel de significancia de 5% *no se puede* rechazar H_0 a favor de H_1. En realidad, el valor crítico de 15% es -1.04 y como $-.91 > -1.04$, no se puede rechazar H_0 ni al nivel de 15%. Se concluye que *enroll* no es estadísticamente significativa al nivel de 15%.

La variable *totcomp* es estadísticamente significativa aun al nivel de significancia de 1% ya que su estadístico *t* es 4.6. Por otro lado, el estadístico *t* para *staff* es 1.2 y, por tanto, no se puede rechazar H_0: $\beta_{staff} = 0$ contra H_1: $\beta_{staff} > 0$ ni al nivel de significancia de 10% (El valor crítico es $c = 1.28$ en la distribución normal estándar).

Para mostrar cómo cambiando la forma funcional pueden afectarse las conclusiones, también se estimará el modelo con todas las variables independientes en forma logarítmica. Esto permite, por ejemplo, que el efecto del tamaño de la escuela disminuya a medida que el tamaño de la escuela aumenta. La ecuación estimada es

$$\begin{aligned}\widehat{math10} &= \underset{(48.70)}{-207.66} + \underset{(4.06)}{21.16}\log(totcomp) + \underset{(4.19)}{3.98}\log(staff) - \underset{(0.69)}{1.29}\log(enroll) \\ n &= 408,\ R^2 = .0654.\end{aligned}$$

El estadístico *t* para log(*enroll*) es aproximadamente -1.87; como este valor está debajo del valor crítico de 5%, -1.65, se rechaza H_0: $\beta_{\log(enroll)} = 0$ a favor de H_1: $\beta_{\log(enroll)} < 0$ al nivel de 5%.

En el capítulo 2 se encontró un modelo (llamado modelo *nivel-log*) en el que la variable dependiente aparecía en su forma original (llamada forma *nivel*), mientras que la variable independiente aparecía en forma logarítmica. En el contexto de la regresión múltiple, la interpretación de los parámetros es la misma, excepto que, por supuesto, a los parámetros se les puede dar una interpretación *ceteris paribus*. Manteniendo constantes *totcomp* y *staff*, se tiene $\Delta\widehat{math10} = -1.29[\Delta\log(enroll)]$, de manera que

$$\Delta\widehat{math10} \approx -(1.29/100)(\%\Delta enroll) \approx -.013(\%\Delta enroll).$$

Una vez más, se ha usado el hecho de que la variación en log(*enroll*), multiplicada por 100, es aproximadamente igual al cambio porcentual de *enroll*. De esta manera, si en una escuela la matrícula es 10% más alta, se pronostica que, $\widehat{math10}$ será $.013(10) = 0.13$ puntos porcentuales inferior (*math*10 se mide como porcentaje).

¿Qué modelo se prefiere: aquel en el que se emplea el nivel de *enroll o el que usa* log(*enroll*)? En el modelo nivel-nivel, la matrícula no tiene un efecto estadístico significativo, pero en el modelo nivel-log sí. Esto se traduce en una *R*-cuadrada mayor en el modelo nivel-log, lo que significa que usando *enroll* en forma logarítmica se explica más de la variación en *math*10 (6.5% a 5.4%). Se prefiere el modelo nivel-log porque capta mejor la relación entre *math*10 y *enroll*. En el capítulo 6 se dirá más del uso de *R*-cuadrada para elegir la forma funcional.

Alternativas de dos colas

En las aplicaciones, es usual probar la hipótesis nula H_0: $\beta_j = 0$ contra una **alternativa de dos colas**; es decir,

$$H_1: \beta_j \neq 0. \tag{4.10}$$

En esta alternativa, x_j tiene un efecto *ceteris paribus* sobre y sin especificar si es positivo o negativo. Esta es la alternativa relevante cuando el signo de β_j no queda determinado por la teoría (o por el sentido común). Aun cuando se sepa si bajo la alternativa β_j es positiva o negativa, suele ser prudente hacer una prueba de dos colas. Cuando menos, emplear una alternativa de dos colas evita mirar la ecuación estimada y basar la alternativa en si $\hat{\beta}_j$ es positiva o negativa. Usar las estimaciones de la regresión como ayuda para formular las hipótesis nula o alternativa no está permitido debido a que la inferencia estadística clásica presupone que las hipótesis nula y alternativa acerca de la población deben establecerse antes de mirar los datos. Por ejemplo, no se debe estimar primero la ecuación que relaciona el desempeño en matemáticas con la cantidad de alumnos inscritos, observar que el efecto estimado es negativo y después decidir que la alternativa relevante es H_1: $\beta_{enroll} < 0$.

Cuando la alternativa es de dos colas, lo que interesa es el *valor absoluto* del estadístico *t*. La regla de rechazo para H_0: $\beta_j = 0$ contra (4.10) es

$$|t_{\hat{\beta}_j}| > c, \tag{4.11}$$

donde $|\cdot|$ denota el valor absoluto y c es un valor crítico elegido de manera aproximada. Para determinar c, se especifica de nuevo un nivel de significancia, por ejemplo 5%. En una **prueba de dos colas**, c se elige de manera que el área en cada cola de la distribución t sea igual a 2.5%. En otras palabras, c es el percentil 97.5 en la distribución t con $n - k - 1$ grados de libertad. Cuando $n - k - 1 = 25$, el valor crítico de 5% para una prueba de dos colas es $c = 2.060$. En la figura 4.4 se proporciona una ilustración de esta distribución.

Cuando no se especifica una alternativa lo común es considerar que se trata de la de dos colas. En este libro, cuando no se especifique otra cosa, la alternativa será la de dos colas y el nivel de significancia será el de 5%. Cuando se realizan análisis econométricos empíricos, es bueno indicar de forma explícita la alternativa y el nivel de significancia. Cuando se rechaza H_0 en favor de (4.10) al nivel de 5%, se suele decir que "x_j es **estadísticamente significativa**, o estadísticamente distinta de cero, al nivel de 5%". Si no se rechaza H_0 se dice que "x_j es **estadísticamente no significativa** al nivel de 5%".

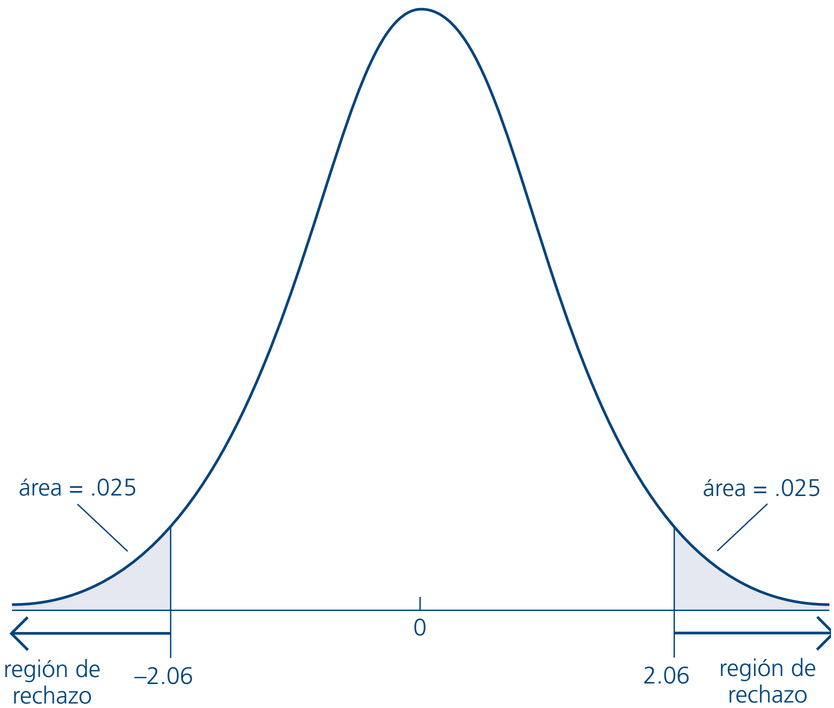
Ejemplo 4.3

[Determinantes del promedio general de calificaciones (GPA) en la universidad]

Se empleará el archivo GPA1.RAW para estimar el modelo que explica el promedio general de calificaciones, GPA, en la universidad (*colGPA*), empleando la cantidad promedio de faltas a clase por semana (*skipped*) como variable explicativa adicional. El modelo estimado es

FIGURA 4.4

Regla de rechazo de 5% para la alternativa H_1: $\beta_j \neq 0$ with 25 *gl*.



$$\widehat{colGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{.412}\, hsGPA + \underset{(.011)}{.015}\, ACT - \underset{(.026)}{.083}\, skipped$$
$$n = 141,\ R^2 = .234.$$

Para ver qué variables son estadísticamente significativas se puede calcular el estadístico *t*, usando una alternativa de dos colas en cada caso. El valor crítico correspondiente a 5% es más o menos 1.96, ya que el número de grados de libertad ($141 - 4 = 137$) es lo suficientemente grande para emplear la aproximación normal estándar. El valor crítico de 1% es más o menos 2.58.

El estadístico *t* para *hsGPA* es 4.38, el cual es significativo a niveles de significancia muy pequeños. Por tanto se dice que "*hsGPA* es estadísticamente significativa a cualquier nivel de significancia *convencional*". El estadístico *t* para *ACT* es 1.36, el cual no es estadísticamente significativa al nivel correspondiente a 10% en una alternativa de dos colas. El coeficiente de *ACT* es además pequeño en sentido práctico: se pronostica que un aumento de 10 puntos en *ACT*, que es grande, incremente *colGPA* sólo .15 puntos. Por tanto, la variable *ACT* es práctica y estadísticamente insignificante.

El coeficiente de *skipped* tiene un estadístico *t* de $-.083/.026 = -3.19$, de manera que *skipped* es estadísticamente significativo al nivel de significancia correspondiente a 1% ($3.19 > 2.58$). Este coeficiente significa que cada clase más que se pierda por semana hará que *colGPA* disminuya más o menos .083. De manera que, manteniendo *hsGPA* y *ACT* constantes, la diferencia que se pronostica en *colGPA* entre un estudiante que no falta a ninguna clase por semana y uno que falta a cinco clases por semana es de manera aproximada .42. Recuerde que esto no dice nada acerca de un determinado estudiante; sino que, .42 es el promedio estimado en una subpoblación de estudiantes.

En este ejemplo podría argumentarse que para cada variable del modelo es adecuada una prueba de una cola. Las variables *hsGPA* y *skipped* son muy significativas usando una prueba de dos colas y tienen los signos esperados, de manera que no hay razón para realizar una prueba de una cola. Por otro lado, contra una alternativa de una cola ($\beta_3 > 0$), *ACT* es significativa al nivel de 10% pero no al de 5%. Esto no cambia el hecho de que el coeficiente de *ACT* sea bastante pequeño.

Otras pruebas de hipótesis acerca de β_j

Aunque H_0: $\beta_j = 0$ es la hipótesis más común, algunas veces se desea probar si β_j es igual a alguna otra constante dada. Dos ejemplos usuales son $\beta_j = 1$ y $\beta_j = -1$. En general, si la hipótesis nula se establece como

$$H_0: \beta_j = a_j, \tag{4.12}$$

donde a_j es el valor hipotético de β_j, entonces el estadístico t apropiado es

$$t = (\hat{\beta}_j - a_j)/\text{ee}(\hat{\beta}_j).$$

Como antes, t mide cuántas desviaciones estándar estimadas se alejaba $\hat{\beta}_j$ del valor hipotético de β_j. Es útil escribir el estadístico t general como

$$t = \frac{(\textit{estimación} - \textit{valor hipotético})}{\text{error estándar}}. \tag{4.13}$$

Bajo (4.12), este estadístico t está distribuido como el estadístico t_{n-k-1} del teorema 4.2. El estadístico t usual se obtiene cuando $a_j = 0$.

El estadístico t general puede emplearse tanto para alternativas de una como de dos colas. Por ejemplo, si las hipótesis nula y alternativa son H_0: $\beta_j = 1$ y H_1: $\beta_j > 1$, respectivamente, el valor crítico para una alternativa de una cola se encuentra *tal* como antes: la diferencia está en cómo se calcula el estadístico t no en cómo se obtiene la c adecuada. H_0 se rechaza a favor de H_1 si $t > c$. En este caso se puede decir que "$\hat{\beta}_j$ es estadísticamente mayor que uno" al nivel de significancia adecuado.

Ejemplo 4.4

[Delincuencia y matrícula en un campus universitario]

Considere un modelo sencillo que relacione la cantidad anual de actos delictivos en el campus de una universidad (*crime*) con su matrícula (*enroll*):

$$\log(crime) = \beta_0 + \beta_1 \log(enroll) + u.$$

Este es un modelo de elasticidad constante, donde β_1 es la elasticidad de actos delictivos (*crime*) en relación con la matrícula (*enroll*). No tiene mucho caso probar H_0: $\beta_1 = 0$, ya que se espera que la cantidad de actos delictivos se incremente a medida que aumenta el tamaño del campus. Una hipótesis que resulta más interesante probar es que la elasticidad de actos delictivos respecto a matrícula sea uno: H_0: $\beta_1 = 1$. Esto significa que un aumento de 1% en la matrícula conduce a un aumento, promedio, de 1% en la actividad delictiva. Una alternativa digna de tomar en cuenta es H_1: $\beta_1 > 1$, que implica que un aumento de 1% en la matrícula hace que la actividad delictiva en el campus aumente *más* de 1%. Si $\beta_1 > 1$, entonces en sentido relativo —no sólo en sentido absoluto— las actividades delictivas son un problema mayor en los campus grandes. Una manera de ver esto es aplicar a la ecuación la función exponencial:

$$crime = \exp(\beta_0) enroll^{\beta_1} \exp(u).$$

(Vea en el apéndice A las propiedades de los logaritmos naturales y de las funciones exponenciales). En la figura 4.5 se grafica esta ecuación con $\beta_0 = 0$ y $u = 0$, para $\beta_1 < 1$, $\beta_1 = 1$ y $\beta_1 > 1$.

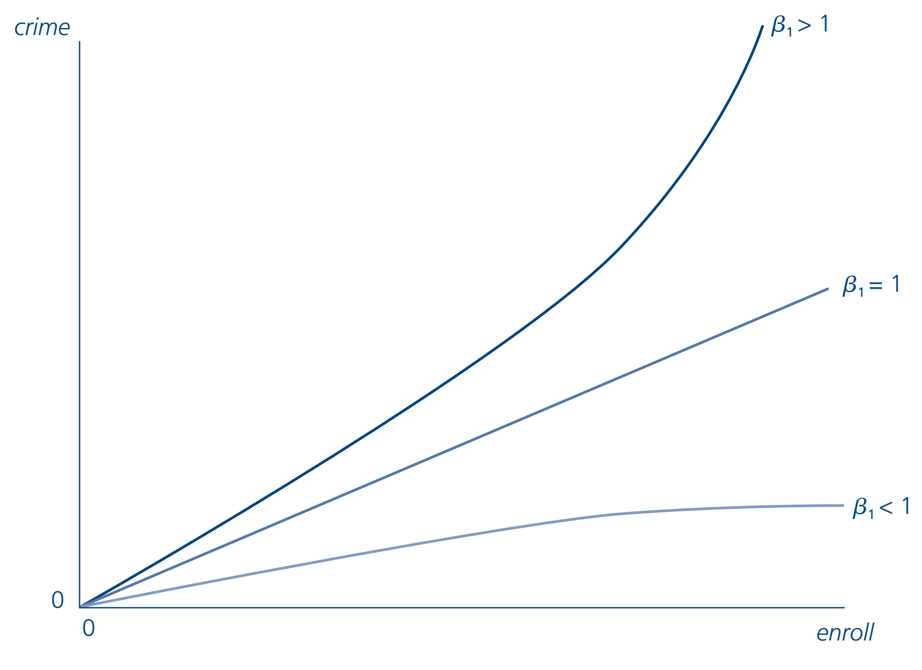
Empleando los datos de 1992 sobre 97 universidades de Estados Unidos contenidos en el archivo CAMPUS.RAW se probará $\beta_1 = 1$ contra $\beta_1 > 1$. Los datos provienen de los *Uniform Crime Reports* del FBI. En la muestra la cantidad promedio de actividades delictivas en un campus es aproximadamente 394 y la matrícula promedio es más o menos 16,076. La ecuación estimada (con las estimaciones y los errores estándar redondeados a dos cifras decimales) es

$$\begin{aligned} \widehat{\log(crime)} &= \underset{(1.03)}{-6.63} + \underset{(0.11)}{1.27} \log(enroll) \\ n &= 97,\ R^2 = .585. \end{aligned} \tag{4.14}$$

La elasticidad estimada para *crime* respecto a *enroll*, 1.27, está en la dirección de la alternativa $\beta_1 > 1$. Pero, ¿hay evidencias suficientes que permitan concluir que $\beta_1 > 1$? Al probar esta hipótesis, hay que ser cuidadosos, en especial debido a que los resultados estadísticos de los paquetes estándar de regresión son mucho más complicados que el sencillo resultado dado en la ecuación (4.14). La primera intención es

FIGURA 4.5

Gráfica de *crime* = *enroll*$^{\beta_1}$ para $\beta_1 < 1$, $\beta_1 = 1$ y $\beta_1 > 1$.



construir "el" estadístico t dividiendo el coeficiente de log(*enroll*) entre el error estándar, que es el estadístico t dado por los paquetes para regresión. Pero este *no* es el estadístico correcto para probar H_0: $\beta_1 = 1$. El estadístico t adecuado se obtiene empleando la ecuación (4.13): el valor hipotético, la unidad, se resta del valor estimado y el resultado se divide entre el error estándar de $\hat{\beta}_1$: $t = (1.27 - 1)/.11 = .27/.11 \approx 2.45$. El valor crítico de una cola correspondiente a 5% para una distribución t con $97 - 2 = 95$ *gl* es aproximadamente 1.66 (empleando $gl = 120$), de manera que al nivel de 5% con claridad se rechaza $\beta_1 = 1$ en favor de $\beta_1 > 1$. En realidad, el valor crítico correspondiente a 1% es más o menos 2.37 con lo que la hipótesis nula se rechaza a favor de la alternativa aun al nivel de 1%.

Hay que tener presente que en este análisis no se mantienen otros factores constantes, de manera que la elasticidad de 1.27 no necesariamente es una buena estimación del efecto *ceteris paribus*. Puede ser que una matrícula grande esté correlacionada con otros factores que ocasionen que la actividad delictiva sea alta: tal vez las escuelas más grandes se encuentran ubicadas en áreas de mayor criminalidad. Esto puede controlarse recolectando datos sobre las tasas de delincuencia en la ciudad de que se trate.

En el caso de una alternativa de dos colas, por ejemplo H_0: $\beta_j = -1$, H_1: $\beta_j \neq -1$, el estadístico t también se calcula como en la ecuación (4.13): $t = (\hat{\beta}_j + 1)/\text{ee}(\hat{\beta}_j)$ (observe que restar -1 significa sumar 1). La regla de rechazo es la usual para una prueba de dos colas: rechazar H_0 si $|t| > c$, donde c es el valor crítico de dos colas. Si se rechaza H_0 se dice que "$\hat{\beta}_j$ es estadísticamente diferente de menos uno" al nivel de significancia adecuado.

Ejemplo 4.5

[Precio de las viviendas y contaminación del aire]

Con ayuda de una muestra de 506 comunidades en la zona de Boston, se estima un modelo que relaciona el precio medio de las viviendas (*price*) en una comunidad con diversas características de la misma: *nox* es la cantidad de óxido de nitrógeno en el aire, dada en partes por millón; *dist* es la distancia ponderada de la comunidad a cinco centros de trabajo, dada en millas; *rooms* es la cantidad promedio de habitaciones en las viviendas de la comunidad, y *stratio* es el cociente promedio estudiantes-profesores en las escuelas de la comunidad. El modelo poblacional es

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2\log(dist) + \beta_3 rooms + \beta_4 stratio + u.$$

Por tanto, β_1 es la elasticidad de *price* respecto a *nox*. Se desea probar H_0: $\beta_1 = -1$ contra la alternativa H_1: $\beta_1 \neq -1$. El estadístico t para esta prueba es $t = (\hat{\beta}_1 + 1)/\text{ee}(\hat{\beta}_1)$.

Empleando los datos del archivo HPRICE2.RAW, el modelo estimado es

$$\begin{aligned}\widehat{\log(price)} &= \underset{(0.32)}{11.08} - \underset{(.117)}{.954}\log(nox) - \underset{(.043)}{.134}\log(dist) + \underset{(.019)}{.255}\,rooms - \underset{(.006)}{.052}\,stratio \\ n &= 506,\ R^2 = .581.\end{aligned}$$

Todas las estimaciones de pendiente tienen el signo previsto. Todos los coeficientes son estadísticamente distintos de cero a un nivel de significancia muy pequeño, incluyendo el coeficiente de log(*nox*). Pero no se desea probar que $\beta_1 = 0$. La hipótesis nula de interés es H_0: $\beta_1 = -1$, cuyo correspondiente estadístico t es $(-.954 + 1)/.117 = .393$. Cuando el estadístico t es tan pequeño, hace falta buscar un valor crítico en las tablas de t: la elasticidad estimada no es estadísticamente diferente a -1 aun a niveles de significancia muy grandes. Controlando los factores que se han incluido, hay poca evidencia de que la elasticidad sea diferente a -1.

Cálculo del valor-*p* en las pruebas *t*

Hasta ahora, se ha hablado de cómo probar hipótesis empleando un procedimiento clásico: una vez establecida la hipótesis alternativa, se elige un nivel de significancia, el cual determina un valor crítico. Ya identificado el valor crítico, el valor del estadístico t se compara con él y entonces la hipótesis nula se rechaza o no al nivel de significancia dado.

Después de elegir la alternativa apropiada, en el procedimiento clásico existe un componente arbitrario, el cual es el resultado de tener que elegir de antemano un nivel de significancia. Cada investigador prefiere un nivel de significancia distinto, dependiendo de la aplicación particular de que se trate. No existe un nivel de significancia "correcto".

Elegir de antemano un nivel de significancia hace que quede oculta información útil acerca del resultado de la prueba de hipótesis. Por ejemplo, suponga que se desea probar la hipótesis nula de que un parámetro es cero contra una alternativa de dos colas, y que con 40 grados de libertad se obtiene un estadístico t igual a 1.85. La hipótesis nula no se rechaza al nivel de 5% ya que el estadístico t es menor que el valor crítico de dos colas de $c = 2.021$. Un investigador cuya orden del día no sea rechazar la hipótesis nula puede que simplemente reporte este resultado junto con el estimado: la hipótesis nula no se rechaza al nivel de 5%. Es claro que si se da también el estadístico t, o el coeficiente y su error estándar, uno puede determinar que la hipótesis nula hubiera sido rechazada al nivel de 10%, ya que el valor crítico correspondiente a 10% es $c = 1.684$.

En lugar de probar a diferentes niveles de significancia, se da más información respondiendo a la pregunta siguiente: dado el valor observado del estadístico t, ¿cuál es el menor nivel de significancia al que se habría rechazado la hipótesis nula? Este nivel se conoce como el **valor-*p*** de la prueba (vea el apéndice C). En el ejemplo anterior, se sabe que el valor-p es mayor que .05, ya que la hipótesis nula no se rechazará al nivel de 5% y también que es menor que .10, ya que la hipótesis nula se rechazará al nivel de 10%. El verdadero valor-p se obtiene calculando la probabilidad de que una variable aleatoria t, con 40 *gl*, sea mayor que 1.85 en valor absoluto. Es decir, el valor-p es el nivel de significancia de la prueba cuando se usa el valor del estadístico de prueba, 1.85 en el ejemplo anterior, como valor crítico de la prueba. En la figura 4.6 se muestra este valor-p.

Como el valor-p es una probabilidad, su valor es siempre un número desde cero hasta uno. Para calcular el valor-p se necesita tener tablas de la distribución t extremadamente detalladas —lo que no es muy práctico— o tener un programa de computadora que calcule áreas bajo la función de densidad de probabilidad de la distribución t. La mayoría de los paquetes modernos de regresión cuentan con esta posibilidad. Si un paquete de regresión da un valor-p junto con el resultado estándar de MCO, casi seguro será el valor-p correspondiente a la prueba de hipótesis H_0: $\beta_j = 0$ contra la alternativa de dos colas. En este caso el valor-p es

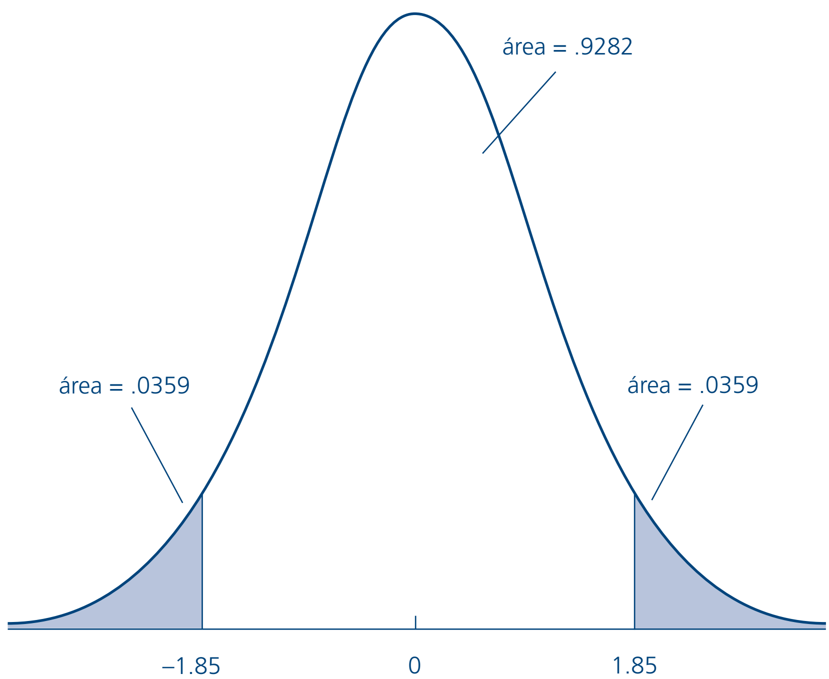
$$\mathrm{P}(|T| > |t|), \qquad \textbf{4.15}$$

donde, por claridad, con T se denota una variable aleatoria con distribución t y $n - k - 1$ grados de libertad y con t se denota el valor numérico del estadístico de prueba.

El valor-p resume la fortaleza o la debilidad de la evidencia empírica contra la hipótesis nula. Tal vez su interpretación más útil sea la siguiente: el valor-p es la probabilidad de observar un estadístico t tan extremo como el que se encontró *si la hipótesis nula es verdadera*. Esto significa que valores-p *pequeños* son evidencia *contra* la hipótesis nula; valores-p grandes proporcionan poca evidencia contra H_0. Por ejemplo, si el valor-$p = .50$ (se da siempre en forma decimal,

FIGURA 4.6

Obtención del valor-*p* en el caso de una alternativa de dos colas, siendo t = 1.85 y *gl* = 40.



no como porcentaje), entonces cuando la hipótesis nula sea verdadera se observará un valor del estadístico t tan extremo como el encontrado en 50% de todas las muestras aleatorias; esta es una evidencia bastante débil contra H_0.

En el ejemplo en el que $gl = 40$ y $t = 1.85$, el valor-p se calcula como sigue

$$\text{Valor-}p = P(|T| > 1.85) = 2P(T > 1.85) = 2(.0359) = .0718,$$

donde $P(T > 1.85)$ es el área a la derecha de 1.85 en una distribución t con 40 *gl*. (Este valor se calculó empleando el paquete econométrico Stata; no se encuentra en la tabla G.2.) Esto significa que, si la hipótesis nula es verdadera, un valor absoluto del estadístico t tan grande como 1.85 se observaría 7.2% de las veces. Esto proporciona cierta evidencia contra la hipótesis nula, pero al nivel de significancia de 5% no se rechazaría dicha hipótesis.

El ejemplo anterior ilustra que una vez calculado el valor-p, se puede realizar una prueba clásica a cualquier nivel deseado. Si α denota el nivel de significancia de la prueba (en forma decimal), entonces H_0 se rechaza si el valor-$p < \alpha$; de no ser así, H_0 no se rechaza al nivel $100{\cdot}\alpha\%$.

Calcular valores-p para alternativas de una cola es también bastante sencillo. Suponga, por ejemplo, que se prueba H_0: $\beta_j = 0$ contra H_1: $\beta_j > 0$. Si $\hat{\beta}_j < 0$, entonces no tiene caso calcular el valor-p: se sabe que es mayor que .50, lo cual nunca hará que se rechace H_0 en favor de H_1. Si $\hat{\beta}_j > 0$, entonces $t > 0$ y el valor-p es precisamente la probabilidad de que una variable aleatoria t con los *gl* apropiados tenga un valor mayor al valor t. Algunos paquetes para regresión sólo calculan valores-p para alternativas de dos colas. Pero obtener el valor-p unilateral es sencillo: simplemente se divide el valor-p de dos colas entre 2.

Si la alternativa es H_1: $\beta_j < 0$, calcular el valor-p tiene sentido si $\hat{\beta}_j < 0$ (y entonces $t < 0$): valor-$p = P(T < t) = P(T > |t|)$ ya que la distribución t es simétrica respecto a cero. De nuevo, esto puede obtenerse dividiendo entre 2 el valor-p de la prueba para dos colas.

Dado que con rapidez se familiarizará uno con las magnitudes de los estadísticos t que conducen a la significancia estadística, en especial cuando se trata de muestras grandes, no resulta siempre crucial dar los valores-p para los estadísticos t. Pero darlos no hace mal. Además, cuando en la sección 4.5 se discuta la prueba F, se verá que es importante calcular los valores-p, ya que no es fácil memorizar los valores críticos para las pruebas F.

Pregunta 4.3

Suponga que estima un modelo de regresión y obtiene $\hat{\beta}_1 = .56$ y valor-$p = .086$ en la prueba de H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$. ¿Cuál es el valor-p en la prueba de H_0: $\beta_1 = 0$ contra H_1: $\beta_1 > 0$?

Repaso del lenguaje empleado en las pruebas de hipótesis clásicas

Cuando no se rechaza H_0, se prefiere decir "no se puede rechazar H_0 al nivel de x%," en lugar de decir "se acepta H_0 al nivel x%". Para mostrar por qué se prefiere la primera manera, se usará el ejemplo 4.5; en él, la elasticidad estimada de *price* en relación a *nox* es $-.954$, y el estadístico t para probar H_0: $\beta_{nox} = -1$ es $t = .393$; por tanto, no se puede rechazar H_0. Pero existen muchos otros valores de β_{nox} (más de los que se pueden contar) que no se pueden rechazar. Por ejemplo, el estadístico t para H_0: $\beta_{nox} = -.9$ es $(-.954 + .9)/.117 = -.462$ y así esta hipótesis nula tampoco se rechaza. Es claro que $\beta_{nox} = -1$ y $\beta_{nox} = -.9$ no pueden ser simultáneamente verdaderos, de manera que no tiene sentido decir que se "acepta" alguna de estas dos hipótesis. Lo único que se puede decir es que los datos no permiten rechazar ninguna de estas hipótesis al nivel de significancia de 5%.

Significancia económica o práctica frente a significancia estadística

Como en esta sección se ha hecho hincapié en la *significancia estadística*, es momento de recordar que hay que poner atención, además de a la magnitud del estadístico t, a la magnitud de los *coeficientes* estimados. La significancia estadística de una variable x_j está por completo determinada por la magnitud de $t_{\hat{\beta}_j}$, mientras que la **significancia económica** o **significancia práctica** de esa variable está relacionada con la magnitud (y signo) de $\hat{\beta}_j$.

Recuerde que el estadístico t para probar H_0: $\beta_j = 0$ se obtiene al dividir la estimación entre su error estándar: $t_{\hat{\beta}_j} = \hat{\beta}_j / \text{ee}(\hat{\beta}_j)$. De manera que, $t_{\hat{\beta}_j}$ puede indicar significancia estadística, ya sea debido a que $\hat{\beta}_j$ sea "grande" o a que $\text{ee}(\hat{\beta}_j)$ sea "pequeño". En la práctica es importante distinguir entre estas razones para que un estadístico t sea significativo. Demasiada atención a la significancia estadística puede conducir a la falsa conclusión de que una variable sea "importante" para explicar y aun cuando su efecto estimado sea modesto.

Ejemplo 4.6

[Tasas de participación en los planes 401(k)]

En el ejemplo 3.3 se usaron los datos de los planes 401 (k) para estimar un modelo que describiera las tasas de participación en términos de la tasa de aportación de la empresa y la antigüedad del plan. Ahora se incluirá una medida del tamaño de la empresa, la cantidad total de empleados en la empresa (*totemp*). La ecuación estimada es

$$\widehat{prate} = \underset{(0.78)}{80.29} + \underset{(0.52)}{5.44}\,mrate + \underset{(.045)}{.269}\,age - \underset{(.00004)}{.00013}\,totemp$$
$$n = 1{,}534,\ R^2 = .100.$$

El menor estadístico t, en valor absoluto, es el de la variable de *totemp*: $t = -.00013/.00004 = -3.25$ y este es estadísticamente significativo a niveles de significancia muy pequeños. (El valor-p de dos colas para este estadístico t es aproximadamente .001). Por tanto, todas las variables son estadísticamente significativas a niveles de significancia bastante pequeños.

¿Qué tan grande, en sentido práctico, es el coeficiente de *totemp*? Manteniendo constantes *mrate* y *age*, si en la firma hay un aumento de 10,000 empleados, la tasa de participación disminuirá 10,000(.00013) = 1.3 puntos porcentuales. Este aumento en la cantidad de empleados es enorme y sólo tienen un efecto moderado sobre la tasa de participación. Por tanto, aunque el tamaño de la empresa afecta la tasa de participación, su efecto, en sentido práctico, no es muy grande.

El ejemplo anterior evidencia que cuando se trabaja con muestras grandes, además de mirar al estadístico t, es muy importante interpretar la magnitud de los coeficientes. Cuando se tienen tamaños de muestra grandes, los parámetros pueden estimarse de manera muy precisa: los errores estándar suelen ser bastante pequeños en relación con los coeficientes estimados, lo que en general da como resultado significancia estadística.

Algunos investigadores insisten en que se usen niveles de significancia cada vez más pequeños a medida que el tamaño de la muestra aumente; esto, en parte, para contrarrestar el hecho de que los errores estándar se hacen más pequeños. Por ejemplo, si para una n de algunos cientos, se emplea el nivel de 5%, cuando n sea de algunos miles, se emplee el nivel de 1%. Usar un nivel de significancia menor quiere decir que la significancia económica y la significancia estadística serán más coincidentes, aunque no existe ninguna garantía: en el ejemplo anterior, aun cuando se use un nivel de significancia tan pequeño como .1% (un décimo de 1%), se concluirá que *totemp* es estadísticamente significativa.

La mayoría de los investigadores están también dispuestos a emplear niveles de significancia mayores en aquellas aplicaciones en las que se tengan tamaños de muestra pequeños reflejando el hecho de que es más dificil encontrar significancia con tamaños de muestra pequeños (los valores críticos son mayores en magnitud y los estimadores son menos precisos). Por desgracia, si éste es o no el caso pueden depender de las intenciones del investigador.

Ejemplo 4.7

[Efecto de la capacitación laboral en la tasa de desperdicio en una empresa]

La tasa de desperdicio en una empresa es la cantidad de artículos defectuosos —productos que deben ser descartados— por cada 100 producidos. Por tanto, para una cantidad dada de artículos producidos, una disminución en la tasa de desperdicio refleja mayor productividad de los trabajadores.

La tasa de desperdicio puede emplearse para medir el efecto de la capacitación de los trabajadores en la productividad. Con ayuda de los datos del archivo JTRAIN.RAW sólo de 1987 y de empresas sin sindicato, se obtiene la ecuación estimada siguiente:

$$\widehat{\log(scrap)} = \underset{(5.69)}{12.46} - \underset{(.023)}{.029}\,hrsemp - \underset{(.453)}{.962}\log(sales) + \underset{(.407)}{.761}\log(employ)$$
$$n = 29,\ R^2 = .262.$$

La variable *hrsemp* representa las horas de capacitación por año por empleado, *sales* representa las ventas anuales de la empresa (en dólares) y *employ* es la cantidad de empleados en la empresa. En 1987 la tasa de desperdicio promedio en la muestra es más o menos 4.6 y el promedio de *hrsemp*, 8.9.

La variable que interesa sobre todo es *hrsemp*. Una hora más de capacitación por empleado hace que log(*scrap*) disminuya .029, lo que significa que la tasa de desperdicio es aproximadamente 2.9% menor. Por tanto, si *hrsemp* aumenta 5 —cada empleado recibe cinco horas más de capacitación por año— se estima que la tasa de desperdicio disminuya $5(2.9) = 14.5\%$. Esto parece ser un efecto razonablemente grande, pero si la capacitación adicional vale la pena para la empresa, depende de los costos de la misma y de los beneficios de una tasa de desperdicio menor. Aquí no se cuenta con los datos necesarios para hacer un análisis costo-beneficio, pero el efecto estimado parece no ser trivial.

¿Y qué se puede decir de la *significancia estadística* de la variable de capacitación? El estadístico t para *hrsemp* es $-.029/.023 = -1.26$, y ahora tal vez usted reconozca este valor como no suficientemente grande en magnitud para concluir que *hrsemp* sea estadísticamente significativa al nivel de 5%. En realidad, con $29 - 4 = 25$ grados de libertad para la alternativa de una cola, H_1, $\beta_{hrsemp} < 0$, el valor crítico correspondiente a 5% es aproximadamente -1.71. Por tanto, usando de manera estricta una prueba con nivel de 5%, debe concluirse que *hrsemp* no es estadísticamente significativa, aun usando una alternativa de una cola.

Dado que el tamaño de la muestra es bastante pequeño, se puede ser menos estricto con el nivel de significancia. El valor crítico correspondiente a 10% es -1.32, y de esta manera *hrsemp* casi es significativa contra una alternativa de una cola al nivel de 10%. El valor-p es $P(T_{25} < -1.26) = .110$. Este puede ser un valor suficientemente pequeño para concluir que el efecto estimado de la capacitación no se debe sólo al error muestral, pero las opiniones diferirán legítimamente sobre si un valor-p de una cola de .11 es suficientemente pequeño.

Recuerde que errores estándar grandes pueden ser también consecuencia de multicolinealidad (una fuerte correlación entre algunas de las variables independientes), aun cuando el tamaño de la muestra parezca bastante grande. Como se vio en la sección 3.4, cuando se tiene este problema no hay mucho que se pueda hacer salvo recolectar más datos o modificar el alcance del análisis eliminando o combinando ciertas variables independientes. Como en el caso de las muestras de tamaño pequeño, puede ser difícil estimar con precisión los efectos parciales cuando algunas de las variables explicativas están fuertemente correlacionadas (En la sección 4.5 se presenta un ejemplo).

Esta sección termina dando algunos lineamientos para analizar la significancia económica y estadística de una variable en un modelo de regresión múltiple:

1. Verificar la significancia estadística. Si una variable es estadísticamente significativa, analice la magnitud de su coeficiente para darse una idea de su importancia práctica o económica. En este último paso puede ser necesario tener cierto cuidado, dependiendo de cómo aparezcan las variables independiente y dependiente en la ecuación (En particular, ¿cuáles son las unidades de medición? ¿Aparecen las variables en forma logarítmica?).
2. Si una variable no es estadísticamente significativa a los niveles usuales (10%, 5% o 1%), de cualquier manera puede preguntarse si tiene el efecto esperado sobre y y si ese efecto es grande en sentido práctico. Si es grande, deberá calcularse el valor-p del estadístico t. En el caso de muestras de tamaño pequeño, algunas veces pueden justificarse valores-p hasta de .20 (pero no hay reglas establecidas). Con valores-p grandes, es decir, con estadísticos t pequeños, se está jugando con fuego, porque estimaciones que son grandes desde el punto de vista práctico pueden deberse al error muestral: puede ser que una muestra aleatoria diferente dé como resultado una estimación diferente.
3. Es común encontrar variables con estadísticos t pequeños que no tengan el signo "correcto". Para fines prácticos, éstas pueden ser ignoradas: se concluye que las variables son estadísticamente no significativas. Una variable significativa que tenga un signo inesperado

y un efecto práctico grande es mucho más problemática y difícil de resolver. Para resolver estos problemas suele ser necesario reflexionar más acerca del modelo y de la naturaleza de los datos. Con frecuencia, una estimación significativa, contraintuitiva, es resultado de la omisión de una variable clave o de uno de los importantes problemas que se analizan en los capítulos 9 y 15.

4.3 Intervalos de confianza

Bajo los supuestos del modelo lineal clásico, es fácil construir un **intervalo de confianza (IC)** para un parámetro poblacional β_j. A los intervalos de confianza se les llama también *estimaciones por intervalo*, porque proporcionan un rango de valores posibles para el parámetro poblacional y no sólo una estimación puntual.

Empleando el hecho de que $(\hat{\beta}_j - \beta_j)/\mathrm{ee}(\hat{\beta}_j)$ tiene una distribución t con $n - k - 1$ grados de libertad [ver (4.3)], mediante manipulaciones sencillas se puede obtener un IC para el parámetro desconocido β_j: un *intervalo de confianza* de 95% está dado por

$$\hat{\beta}_j \pm c \cdot \mathrm{ee}(\hat{\beta}_j), \tag{4.16}$$

donde la constante c es el percentil 97.5 en la distribución t_{n-k-1}. Con más exactitud, los límites inferior y superior de un intervalo de confianza están dados por

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot \mathrm{ee}(\hat{\beta}_j)$$

y

$$\bar{\beta}_j \equiv \hat{\beta}_j + c \cdot \mathrm{ee}(\hat{\beta}_j),$$

respectivamente.

En este punto es aconsejable revisar el significado de intervalo de confianza. Si se obtuvieran una y otra y otra muestra aleatoria, y cada vez se calcularán $\underline{\beta}_j$ y $\bar{\beta}_j$ entonces en 95% de las muestras el valor poblacional (desconocido) β_j estaría en el intervalo $(\underline{\beta}_j, \bar{\beta}_j)$. Por desgracia, en la única muestra que se emplea para construir el IC, no se sabe si β_j está o no contenida en el intervalo. Se espera haber obtenido una muestra que pertenezca a 95% de las muestras en las que la estimación por intervalo contiene a β_j, pero no se tiene ninguna garantía de esto.

Con la tecnología actual es fácil construir un intervalo de confianza. Se necesitan tres cantidades: $\hat{\beta}_j$, $\mathrm{ee}(\hat{\beta}_j)$ y c. El coeficiente estimado y su error estándar son proporcionados por los paquetes de regresión. Para obtener el valor de c, se necesitan los grados de libertad, $n - k - 1$, y el nivel de confianza –95% en este caso. Entonces, de la distribución t_{n-k-1} se obtiene el valor de c.

Por ejemplo, para $gl = n - k - 1 = 25$ un intervalo de confianza de 95% para cualquier β_j es el dado por $[\hat{\beta}_j - 2.06 \cdot \mathrm{ee}(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot \mathrm{ee}(\hat{\beta}_j)]$.

Cuando $n - k - 1 > 120$, la distribución t_{n-k-1} es lo suficientemente parecida a la distribución normal estándar como para usar el percentil 97.5 de la distribución normal estándar para construir un IC de 95%: $\hat{\beta}_j \pm 1.96 \cdot \mathrm{ee}(\hat{\beta}_j)$. En realidad, cuando $n - k - 1 > 50$, el valor de c es tan cercano a 2 que para un intervalo de confianza de 95% puede emplearse una *regla sencilla*: $\hat{\beta}_j$ más o menos dos veces su error estándar. Para grados de libertad pequeños, el percentil exacto debe obtenerse de las tablas t.

Para cualquier otro nivel de confianza también es fácil construir intervalos de confianza. Por ejemplo un IC de 90% se obtiene eligiendo c como el percentil 95 en la distribución t_{n-k-1}. Para $gl = n - k - 1 = 25$, $c = 1.71$, con lo que el IC de 90% es $\hat{\beta}_j \pm 1.71 \cdot \mathrm{ee}(\hat{\beta}_j)$, que necesariamente es más estrecho que el IC de 95%. Para un IC de 99%, c es el percentil 99.5 de la distribución

t_{25}. Si $gl = 25$, el IC de 99% es aproximadamente $\hat{\beta}_j \pm 2.79 \cdot \text{ee}(\hat{\beta}_j)$, que inevitablemente es más amplio que el IC de 95%.

Muchos paquetes de regresión modernos ahorran al usuario el trabajo de hacer cualquier cálculo dando junto con cada coeficiente y su error estándar un IC de 95%. Una vez que se ha construido un intervalo de confianza, realizar una prueba de hipótesis de dos colas es sencillo. Si la hipótesis nula es H_0: $\beta_j = a_j$, entonces se rechaza H_0 contra H_1: $\beta_j \neq a_j$ al nivel de significancia (por ejemplo) de 5% si y sólo si, a_j *no* está en el intervalo de confianza de 95%.

Ejemplo 4.8

[Modelo de los gastos en Investigación y Desarrollo]

A los economistas que se dedican al estudio de la organización industrial les interesa la relación entre el tamaño de la empresa —que suele medirse por las ventas anuales— y los gastos en investigación y desarrollo (I & D). En general se emplea un modelo de elasticidad constante. También puede ser interesante conocer el efecto *ceteris paribus* del margen de utilidad —es decir, las utilidades como porcentaje de las ventas— sobre los gastos en I & D. Con ayuda de los datos del archivo RDCHEM.RAW, sobre 32 empresas estadounidenses de la industria química, se estima la ecuación siguiente (donde los errores estándar aparecen entre paréntesis debajo de los coeficientes):

$$\begin{aligned}\widehat{\log(rd)} &= \underset{(.47)}{-4.38} + \underset{(.060)}{1.084} \log(sales) + \underset{(.0218)}{.0217}\, profmarg \\ n &= 32,\ R^2 = .918.\end{aligned}$$

La elasticidad estimada para los gastos en I & D en relación con las ventas de la empresa *sales*, es 1.084, de manera que manteniendo constante el margen de utilidad, *profmarg*, constante, a un aumento de 1% en las ventas le corresponde un aumento de 1.084% en los gastos en I & D (Por casualidad, tanto I & D como ventas están dados en millones de dólares, pero las unidades de medición no tienen ningún efecto en la elasticidad estimada). Una vez que se observa que el modelo estimado tiene $n - k - 1 = 32 - 2 - 1 = 29$ grados de libertad puede construirse un intervalo de confianza de 95% para la elasticidad de las ventas. En la tabla G.2 se encuentra el percentil 97.5 en la distribución t_{29}: $c = 2.045$. Por tanto, el intervalo de confianza de 95% para $\beta_{\log(sales)}$ es $1.084 \pm .060(2.045)$, es decir, aproximadamente (.961,1.21). Que el cero esté por completo fuera de este intervalo no debe sorprender: se espera que los gastos en I & D aumenten con el tamaño de la empresa. Más interesante resulta que la unidad se encuentre dentro del intervalo de confianza de 95% para $\beta_{\log(sales)}$, lo cual significa que no se puede rechazar H_0: $\beta_{\log(sales)} = 1$ contra H_1: $\beta_{\log(sales)} \neq 1$ al nivel de significancia de 5%. En otras palabras, la elasticidad estimada de I & D respecto a ventas no es estadísticamente distinta de 1 al nivel de 5% (La estimación tampoco es prácticamente distinta de 1).

El coeficiente estimado para *profmarg* también es positivo y el intervalo de confianza de 95% para el parámetro poblacional, $\beta_{profmarg}$, es $.0217 \pm .0218(2.045)$, es decir, aproximadamente $(-.0045,.0479)$. En este caso, el cero está dentro del intervalo de confianza de 95%, de manera que no se puede rechazar H_0: $\beta_{profmarg} = 0$ contra H_1: $\beta_{profmarg} \neq 0$ al nivel de 5%. De cualquier manera, el estadístico t es aproximadamente 1.70, con lo que se obtiene un valor-p de dos colas de más o menos .10, y se concluye que *profmarg* es estadísticamente significativo al nivel de 10% contra la alternativa de dos colas, o al nivel de 5% contra la alternativa de una cola H_1: $\beta_{profmarg} > 0$. Además, la magnitud económica del coeficiente del margen de utilidad no es trivial: manteniendo *sales* constante, se estima que un aumento de un punto porcentual en *profmarg* incremente los gastos en I & D en $100(.0217) \approx 2.2\%$. Un análisis completo de este ejemplo va más allá de establecer si un determinado valor, cero en este caso, está o no en el intervalo de confianza de 95%.

Hay que recordar que un intervalo de confianza sólo puede ser tan bueno como los supuestos subyacentes empleados para construirlo. Si se han omitido factores importantes que estén correlacionados con las variables explicativas, las estimaciones de los coeficientes no son confiables: MCO es sesgado. Si existe heterocedasticidad —como por ejemplo, si que la varianza de log(*rd*) depende de alguna de las variables explicativas— entonces el error estándar no es válido como estimación de($\hat{\beta}_j$) (como se analizó en la sección 3.4) y el intervalo de confianza calculado usando estos errores estándar no será en verdad un IC de 95%. Se ha usado también el supuesto de normalidad en los errores para obtener estos IC, pero, como se verá en el capítulo 5, esto no es tan importante en aplicaciones en las que se cuenta con cientos de observaciones.

4.4 Pruebas de hipótesis de una sola combinación lineal de los parámetros

En las dos secciones anteriores se vio cómo usar las pruebas de hipótesis o los intervalos de confianza clásicos para probar hipótesis acerca de un solo β_j. En las aplicaciones, con frecuencia se tienen que probar hipótesis en las que interviene más de un parámetro poblacional. En esta sección, se muestra cómo probar una sola hipótesis en la que interviene más de uno de los β_j. En la sección 4.5 se muestra cómo probar hipótesis múltiples.

Para ilustrar el método general, se considerará un modelo sencillo en el que se compara el rendimiento de la educación para carreras universitarias cortas (de dos años) y largas (cuatro años); para simplificar, a las últimas se les llamará "universidades". [Kane y Rouse (1995) proporcionan un análisis detallado de los rendimientos de estudios universitarios de dos y cuatro años.] La población consta de personas trabajadoras que tiene bachillerato terminado, y el modelo es

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u, \tag{4.17}$$

donde

jc = cantidad de años de asistencia a una carrera universitaria corta.
$univ$ = cantidad de años de asistencia a una carrera universitaria larga.
$exper$ = meses en la fuerza laboral.

Observe que se puede tener cualquier combinación de carrera corta y larga, incluyendo $jc = 0$ y $univ = 0$.

La hipótesis que interesa es si un año más de estudio en una carrera corta vale lo mismo que un año más de estudio en una universidad: esto se expresa como

$$H_0: \beta_1 = \beta_2. \tag{4.18}$$

De acuerdo con la H_0, un año más de estudio en una carrera corta y un año más en la universidad conducen al mismo aumento porcentual *ceteris paribus* en *wage* (salario). En la mayoría de los casos, la alternativa que interesa es la de una cola: un año más de una carrera corta vale menos que un año más de universidad. Esto se expresa como

$$H_1: \beta_1 < \beta_2. \tag{4.19}$$

En las hipótesis en (4.18) y en (4.19) intervienen *dos* parámetros, β_1 y β_2, situación que no había sido encontrada antes. Para probar H_0 no pueden usarse simplemente los estadísticos individuales correspondientes a $\hat{\beta}_1$ y $\hat{\beta}_2$. Sin embargo, de manera conceptual, no hay ninguna dificultad para construir un estadístico *t* para probar (4.18). Para esto, las hipótesis nula y alternativa se escriben de la manera siguiente $H_0: \beta_1 - \beta_2 = 0$ y $H_1: \beta_1 - \beta_2 < 0$, respectivamente. El estadístico *t* se basa en si la diferencia estimada $\hat{\beta}_1 - \hat{\beta}_2$ es suficientemente menor a cero para rechazar (4.18) a favor de (4.19). Para tomar en cuenta el error de muestreo en este estimador, se estandariza esta diferencia dividiendo entre el error estándar:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{ee}(\hat{\beta}_1 - \hat{\beta}_2)}. \tag{4.20}$$

Una vez que se tiene el estadístico t de (4.20), la prueba se realiza como antes. Se elige un nivel de significancia para la prueba y, de acuerdo con los *gl*, se obtiene un valor crítico. Debido a que la alternativa es de la forma dada en (4.19), la regla de rechazo es de la forma $t < -c$, donde c es un valor positivo elegido de la distribución t adecuada. También se puede calcular el estadístico t y calcular después el valor-p (vea la sección 4.2).

Lo único que dificulta más probar la igualdad de dos parámetros distintos, que hacer una prueba cerca de un solo β_j es obtener el error estándar que aparece en el denominador de (4.20). Obtener el numerador es trivial una vez realizada la regresión por MCO. Empleando los datos del archivo TWOYEAR.RAW, que provienen de Kane y Rouse (1995), se estima la ecuación (4.17):

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(.021)}{1.472} + \underset{(.0068)}{.0667}\, jc + \underset{(.0023)}{.0769}\, univ + \underset{(.0002)}{.0049}\, exper \\ n &= 6{,}763,\ R^2 = .222. \end{aligned} \tag{4.21}$$

De acuerdo con (4.21) es claro que tanto *jc* como *univ* tienen efectos económica y estadísticamente significativos sobre el salario. Esto es interesante, pero lo que importa ahora es probar si la *diferencia* estimada entre los coeficientes es estadísticamente significativa. La diferencia se estima como $\hat{\beta}_1 - \hat{\beta}_2 = -.0102$, de manera que el rendimiento de un año de carrera corta es aproximadamente un punto porcentual menor que un año en una universidad. En cuestión económica esta no es una diferencia trivial. La diferencia $-.0102$ es el numerador del estadístico t en (4.20).

Por desgracia, los resultados de la regresión dados en la ecuación (4.21) *no* contienen suficiente información para obtener el error estándar de $\hat{\beta}_1 - \hat{\beta}_2$. Puede pensarse que $\text{ee}(\hat{\beta}_1 - \hat{\beta}_2) = \text{ee}(\hat{\beta}_1) - \text{ee}(\hat{\beta}_2)$, pero esto no es verdad. De hecho, si se invierten los lugares de $\hat{\beta}_1$ y $\hat{\beta}_2$, usando la diferencia de los errores estándar se obtiene un error estándar negativo para la diferencia. Los errores estándar *siempre* deben ser positivos, porque son estimaciones de desviaciones estándar. Aunque el error estándar de la diferencia $\hat{\beta}_1 - \hat{\beta}_2$ sí depende de $\text{ee}(\hat{\beta}_1)$ y $\text{ee}(\hat{\beta}_2)$, lo hace de una forma un poco más complicada. Para determinar $\text{ee}(\hat{\beta}_1 - \hat{\beta}_2)$, primero se obtiene la varianza de la diferencia. Empleando los resultados sobre varianzas dados en el apéndice B, se tiene

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2). \tag{4.22}$$

Observe que las dos varianzas se *suman* y después se resta el doble de la covarianza. La desviación estándar de $\hat{\beta}_1 - \hat{\beta}_2$ es la raíz cuadrada de (4.22), y, como $[\text{ee}(\hat{\beta}_1)]^2$ es un estimador insesgado de $\text{Var}(\hat{\beta}_1)$ y lo mismo ocurre con $[\text{ee}(\hat{\beta}_2)]^2$, se tiene

$$\text{ee}(\hat{\beta}_1 - \hat{\beta}_2) = \left\{[\text{ee}(\hat{\beta}_1)]^2 + [\text{ee}(\hat{\beta}_2)]^2 - 2s_{12}\right\}^{1/2}, \tag{4.23}$$

donde s_{12} denota una estimación de $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Aquí no se ha dado la fórmula para $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Algunos paquetes para regresión permiten obtener s_{12}, en cuyo caso se puede calcular el error estándar de (4.23) y después el estadístico t de (4.20). En el apéndice E se muestra cómo usar el álgebra de matrices para obtener s_{12}.

Algunos de los programas econométricos más sofisticados tienen comandos especiales para probar hipótesis acerca de combinaciones lineales. Aquí se verá un método que puede emplearse con casi cualquier paquete estadístico. En lugar de tratar de calcular ee$(\hat{\beta}_1 - \hat{\beta}_2)$ de acuerdo con (4.23), es mucho más fácil estimar un modelo diferente que proporcione de forma directa el error estándar que interesa. Se define un parámetro nuevo que es la diferencia entre β_1 y β_2: $\theta_1 = \beta_1 - \beta_2$. Y lo que se quiere probar ahora es

$$\text{H}_0\text{: } \theta_1 = 0 \text{ contra } \text{H}_1\text{: } \theta_1 < 0. \tag{4.24}$$

En términos de $\hat{\theta}_1$, el estadístico t en (4.20) es $t = \hat{\theta}_1/\text{ee}(\hat{\theta}_1)$. El problema es hallar ee$(\hat{\theta}_1)$.

Para esto se reescribe el modelo de manera que θ_1 aparezca de manera directa en una de las variables independientes. Como $\theta_1 = \beta_1 - \beta_2$, se puede escribir $\beta_1 = \theta_1 + \beta_2$. Sustituyendo en (4.17) y reordenando se obtiene la ecuación

$$\begin{aligned} \log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + univ) + \beta_3 exper + u. \end{aligned} \tag{4.25}$$

La idea clave es que el parámetro acerca del cual se desea probar la hipótesis, θ_1, multiplica ahora la variable *jc*. El intercepto sigue siendo β_0 y *exper* sigue apareciendo multiplicada por β_3. Lo importante es que hay una variable nueva que multiplica a β_2, a saber $jc + univ$. Por tanto, si se desea estimar directamente θ_1 y obtener el error estándar de $\hat{\theta}_1$, se debe construir la variable nueva $jc + univ$ e incluirla en el modelo de regresión en lugar de *univ*. En este ejemplo, la variable nueva tiene una interpretación natural: es el *total* de los años de universidad, de manera que se define $totcoll = jc + univ$ y (4.25) se escribe como

$$\log(wage) = \beta_0 + \theta_1 jc + \beta_2 totcoll + \beta_3 exper + u. \tag{4.26}$$

El parámetro β_1 ha desaparecido del modelo, mientras que θ_1 aparece explícitamente. Este modelo es en realidad sólo una manera diferente de escribir el modelo original. La única razón por la que se ha definido este modelo nuevo es que, al estimarlo, el coeficiente en *jc* de $\hat{\theta}_1$, y, lo más importante, que ee$(\hat{\theta}_1)$ se obtiene junto con la estimación. El estadístico t que se desea conocer es el proporcionado por cualquier paquete de regresión para la variable *jc* (*no* para la variable *totcoll*).

Haciendo esto con las 6,763 observaciones antes empleadas el resultado es

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(.021)}{1.472} - \underset{(.0069)}{.0102}\, jc + \underset{(.0023)}{.0769}\, totcoll + \underset{(.0002)}{.0049}\, exper \\ n &= 6{,}763,\ R^2 = .222. \end{aligned} \tag{4.27}$$

El único número en esta ecuación que no se pudo obtener con (4.21) es el error estándar de la estimación $-.0102$, que es .0069. El estadístico t para probar (4.18) es $-.0102/.0069 = -1.48$. Contra la alternativa de una cola (4.19), el valor-p es aproximadamente .070, de manera que hay una cierta, aunque no una fuerte, evidencia en contra de (4.18).

Las estimaciones el intercepto y de la pendiente de *exper*, así como sus errores estándar, son los mismos que en (4.21). Esto *debe* ser así y es una manera de verificar si la ecuación transformada ha sido estimada de forma adecuada. El coeficiente de la variable nueva, *totcoll*, es el mismo que el de *univ* en (4.21) y su error estándar es también el mismo. Comparando (4.17) y (4.25) se sabe que esto debe ser así.

Calcular un intervalo de confianza de 95% para $\theta_1 = \beta_1 - \beta_2$. es sencillo. Usando la aproximación normal estándar, el IC se obtiene como de costumbre: $\hat{\theta}_1 \pm 1.96$ ee$(\hat{\theta}_1)$, lo que en este caso da $-.0102 \pm .0135$.

La estrategia de reescribir el modelo de manera que contenga el parámetro que interesa funciona en todos los casos y es fácil de realizar. (Vea los ejercicios para computadora C4.1 y C4.3 para más ejemplos.)

4.5 Pruebas para restricciones lineales múltiples: la prueba *F*

El estadístico *t* relacionado con cualquier coeficiente de MCO se emplea para probar si el correspondiente parámetro poblacional desconocido es igual a una constante dada (el que por lo general, aunque no siempre, es cero). Se acaba de mostrar cómo probar hipótesis acerca de una combinación lineal de β_j mediante una reordenación de la ecuación y realizando la regresión usando variables transformadas. Pero hasta ahora, sólo se han visto hipótesis en las que interviene una *única* restricción. Con frecuencia, se desea probar hipótesis *múltiples* acerca de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$. Se empieza con el caso principal en el que se prueba si un conjunto de variables independientes no tiene efecto parcial sobre la variable dependiente.

Prueba para las restricciones de exclusión

Ya se ha visto cómo probar si una determinada variable no tiene efecto parcial sobre la variable dependiente: usando el estadístico *t*. Ahora, se desea probar si un *grupo* de variables no tiene efecto sobre la variable dependiente. De manera más concreta, la hipótesis nula es que un conjunto de variables no tiene efecto sobre *y*, una vez que otro conjunto de variables ha sido controlado.

Como ejemplo de la utilidad de probar la significancia de un grupo de variables, considere el modelo siguiente que explica los sueldos de los jugadores de la liga mayor de béisbol:

$$\begin{aligned}\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\ + \beta_4 hrunsyr + \beta_5 rbisyr + u,\end{aligned} \tag{4.28}$$

donde *salary* es el sueldo total en 1993, *years* es años en la liga, *gamesyr* es el promedio de partidos jugados por año, *bavg* es el promedio de bateo a lo largo de la carrera de un jugador (por ejemplo, $bavg = 250$), *hrunsyr* es cuadrangulares por año y *rbisyr* es carreras impulsadas por año. Suponga que se desea probar la hipótesis nula de que, una vez controlados años en la liga y partidos por año, los estadísticos que miden el desempeño —*bavg*, *hrunsyr* y *rbisyr*— no tienen efectos sobre el sueldo. En esencia, esta hipótesis nula establece que la productividad como la miden las estadísticas de béisbol no tiene efecto sobre el sueldo.

En términos de los parámetros del modelo, la hipótesis nula se establece de la manera siguiente:

$$H_0: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0. \tag{4.29}$$

La hipótesis nula (4.29) constituye tres **restricciones de exclusión**: si (4.29) es verdadera, entonces, una vez que *years* y *gamesyr* han sido controladas *bavg*, *hrunsyr* y *rbisyr* no tienen ningún efecto sobre log(*salary*) y, por tanto, deben excluirse del modelo. Este es un ejemplo de un conjunto de **restricciones múltiples** ya que a los parámetros de (4.28) se les pone más de una restricción; ejemplos más generales de restricciones múltiples se verán más adelante. A una prueba de restricciones múltiples se le llama **prueba de hipótesis múltiple** o **prueba de hipótesis conjunta**.

¿Cuál será la alternativa a (4.29)? Si lo que se quiere tener en mente es que "los estadísticos de desempeño importan, aun después de controlar los años en la liga y los partidos por año", entonces la alternativa apropiada es simplemente

$$\mathrm{H}_1\text{: } \mathrm{H}_0 \text{ no es verdadera.} \tag{4.30}$$

La alternativa (4.30) se satisface si por lo menos uno de los β_3, β_4 o β_5 es diferente de cero (Uno o todos pueden ser diferentes de cero). La prueba que se estudia aquí permite detectar cualquier violación a H_0. La prueba también es válida cuando la alternativa es algo así como H_1: $\beta_3 > 0$, o $\beta_4 > 0$ o $\beta_5 > 0$, pero bajo tales alternativas no es la mejor prueba utilizable. Aquí no se cuenta con el espacio ni los conocimientos estadísticos previos necesarios para ver pruebas de mayor potencia para alternativas de una cola múltiples.

¿Qué hay que hacer para probar (4.29) contra (4.30)? Uno está inclinado a pensar que para probar (4.29) pueden usarse los estadísticos *t* de las variables *bavg*, *hrunsyr* y *rbisyr* para determinar si cada variable es significativa *individualmente*. Esta opción no es la adecuada. Un estadístico *t* determinado sirve para probar una hipótesis que no coloque ninguna restricción a otros parámetros. Además, se tendrían tres resultados —uno por cada estadístico *t*—. ¿Cuál sería el criterio para rechazar (4.29) al nivel de, por ejemplo, 5%? ¿Sería necesario que los tres estadísticos *t* fueran significativos al nivel de 5% o bastaría con que uno lo fuera? Estas son preguntas difíciles, pero por fortuna no tienen que ser contestadas. Además, emplear por separado estadísticos *t* para probar una hipótesis múltiple como la (4.29) puede resultar muy confuso. Se necesita una manera de probar *conjuntamente* las restricciones de exclusión.

Para dar un ejemplo, se estima la ecuación (4.28] empleando los datos del archivo MLB1. RAW. Se obtiene

$$\begin{aligned}
\widehat{\log(salary)} = {} & \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\, years + \underset{(.0026)}{.0126}\, gamesyr \\
& + \underset{(.00110)}{.00098}\, bavg + \underset{(.0161)}{.0144}\, hrunsyr + \underset{(.0072)}{.0108}\, rbisyr \\
n = 353,\ & \mathrm{SRC} = 183.186,\ R^2 = .6278,
\end{aligned} \tag{4.31}$$

donde SRC es la suma de los residuales cuadrados (Esta suma se empleará más adelante). Para facilitar comparaciones futuras, se han dejado varios dígitos después del punto decimal en la SRC y en la R^2. La ecuación (4.31) indica que, mientras *years* y *gamesyr* son estadísticamente significativos, ninguna de las variables *bavg*, *hrunsyr* y *rbisyr* tiene un estadístico *t* estadísticamente significativo, al nivel de significancia de 5%, frente a la alternativa de dos colas (El estadístico *t* de *rbisyr* es el más cercano a ser significativo; su valor-*p* de dos colas es .134). Por tanto, con base en los tres estadísticos *t*, parece que no puede rechazarse H_0.

Esta conclusión resulta ser errónea. Para ver esto, se debe obtener un estadístico de prueba para restricciones múltiples cuya distribución sea conocida y esté tabulada. La suma de los residuales cuadrados proporciona una base adecuada para probar hipótesis múltiples. Se mostrará que también *R*-cuadrada puede emplearse en el caso especial en que se prueben restricciones de exclusión.

La suma de los residuales cuadrados de (4.31) no dice nada acerca de la veracidad de la hipótesis en (4.29). Pero, el factor que sí dice algo es el aumento de la SRC cuando se eliminan del modelo las variables *bavg*, *hrunsyr* y *rbisyr*. Recuerde que, como las estimaciones de MCO se eligen de manera que se minimice la suma de los residuales cuadrados, *siempre* que se eliminan variables del modelo, la SRC aumenta; esto es un hecho algebraico. La pregunta es si este aumento es suficientemente grande, *en relación* con la SRC del modelo que tiene todas las variables, como para que se rechace la hipótesis nula.

El modelo sin las tres variables en cuestión es simplemente

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u. \tag{4.32}$$

En el contexto de las pruebas de hipótesis, la ecuación (4.32) es el **modelo restringido** para probar (4.29); al modelo (4.28) se le llama **modelo no restringido**. El modelo restringido siempre tiene menos parámetros que el no restringido.

Al estimar el modelo restringido con los datos del archivo MLB1.RAW, se obtiene

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(.11)}{11.22} + \underset{(.0125)}{.0713}\, years + \underset{(.0013)}{.0202}\, gamesyr \\ n &= 353,\ \text{SRC} = 198.311,\ R^2 = .5971.\end{aligned} \tag{4.33}$$

Como se esperaba, la SRC en (4.33) es mayor que la SRC en (4.31) y la R-cuadrada del modelo restringido es menor que la R-cuadrada del modelo no restringido. Lo que hay que decidir es si el aumento de la SRC, al pasar del modelo no restringido al modelo restringido (183.186 a 198.311) es suficientemente grande para rechazar (4.29). Como en todas las pruebas, la respuesta depende del nivel de significancia de la prueba. Pero la prueba no puede realizarse al nivel de significancia deseado hasta que se tenga un estadístico cuya distribución se conozca y pueda ser tabulada, bajo H_0. Por tanto, se necesita una manera de combinar la información de las dos SRC para obtener un estadístico de prueba que tenga una distribución conocida bajo H_0.

También puede obtenerse la prueba para el caso general; esto no es más difícil. Se escribe el modelo *no restringido* que tiene k variables independientes de la manera siguiente

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u; \tag{4.34}$$

la cantidad de parámetros en el modelo no restringido es $k + 1$. (Recuerde sumar uno por el intercepto.) Suponga que se tienen que probar q restricciones de exclusión: es decir, la hipótesis nula dice que q de las variables en (4.34) tienen coeficientes igual a cero. Para simplificar la notación, suponga que estas son las últimas q variables de la lista de variables independientes: $x_{k-q+1}, \ldots, x_k$. (Por supuesto que el orden de las variables es arbitrario y no tiene importancia.) La hipótesis nula es

$$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \tag{4.35}$$

la cual impone q restricciones de exclusión al modelo (4.34). La alternativa a (4.35) es simplemente que esta hipótesis es falsa; esto significa que por lo menos uno de los parámetros que aparecen en (4.35) es diferente de cero. Imponiendo las restricciones bajo H_0, se obtiene el modelo restringido:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u. \tag{4.36}$$

En esta subsección, se supondrá que tanto en el modelo no restringido como en el restringido existe un intercepto, pues este es el caso que se presenta con más frecuencia en la práctica.

Ahora, en cuanto al estadístico de prueba mismo. Antes, se dijo que mirando el aumento relativo de la SRC al pasar del modelo no restringido al modelo restringido podría obtenerse información para probar la hipótesis (4.35). El ***estadístico F*** (o *cociente F*) se define como

$$F \equiv \frac{(\text{SRC}_r - \text{SRC}_{nr})/q}{\text{SRC}_{nr}/(n-k-1)}, \tag{4.37}$$

donde SRC_r es la suma de residuales cuadrados del modelo restringido y SRC_{nr} es la suma de residuales cuadrados del modelo no restringido.

Pregunta 4.4

Considere relacionar el rendimiento individual en un examen estandarizado, *score* (puntuación), con otras variables diversas. Las variables relacionadas con la escuela son tamaño promedio del grupo (*classize*), gasto por estudiante (*expend*), compensación promedio a los profesores (*tchcomp*) y matrícula total (*enroll*). Otras variables específicas al estudiante son ingreso familiar (*faminc*), educación de la madre (*motheduc*), educación del padre (*fatheduc*) y número de hermanos (*siblings*). El modelo es

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + \beta_5 faminc + \beta_6 motheduc + \beta_7 fatheduc + \beta_8 siblings + u.$$

Dé la hipótesis nula que dice que una vez controlados los factores relacionados con la escuela, las variables específicas del estudiante no tienen efecto en el desempeño en la prueba estandarizada, ¿cuáles son los valores de k y de q en este ejemplo? Escriba la versión restringida del modelo.

Se observará de inmediato que, como SRC_r no puede ser menor que SRC_{nr}, el estadístico F *siempre* es no negativo (y casi siempre estrictamente positivo). Por tanto, si se obtiene un estadístico F negativo, algo debe de estar mal; por lo general es que se ha invertido el orden de las SRC en el numerador de F. Además, la SRC en el denominador de F es la SRC del modelo *no restringido*. La manera más sencilla de recordar dónde van las SRC es entender a F como una medida del aumento relativo de la SRC al pasar del modelo no restringido al restringido.

La diferencia de las SRC en el numerador de F es dividida entre q, la cantidad de restricciones impuestas al pasar del modelo no restringido al restringido (se eliminan q variables independientes). Por tanto, se puede escribir

$$q = \textbf{grados de libertad en el numerador} = gl_r - gl_{nr}, \qquad \textbf{4.38}$$

lo que indica que q es la diferencia en grados de libertad entre el modelo restringido y el no restringido. (Recuerde que gl = cantidad de observaciones − cantidad de parámetros estimados.) Como el modelo restringido tiene menos parámetros —y cada modelo se estima empleando las mismas n observaciones— gl_r es siempre mayor que gl_{nr}.

La SRC en el denominador de F es dividida entre los grados de libertad en el modelo no restringido:

$$n - k - 1 = \textbf{grados de libertad en el denominador} = gl_{nr}. \qquad \textbf{4.39}$$

En realidad, el denominador de F es precisamente el estimador insesgado de $\sigma^2 = \text{Var}(u)$ en el modelo no restringido.

En una aplicación específica, calcular el estadístico F es más sencillo que cuando se emplea la notación un poco complicada usada para describir el caso general. Primero, se obtienen los grados de libertad del modelo no restringido, gl_{nr}. Después, se cuentan las variables que han sido excluidas en el modelo restringido; esto es q. Las SRC se obtienen de los resultados de las regresiones de MCO, y de esta manera calcular el estadístico F es sencillo.

En la regresión acerca del sueldo en la liga mayor de béisbol, $n = 353$, y el modelo completo (4.28) contiene seis parámetros. Por tanto, $n - k - 1 = gl_{nr} = 353 - 6 = 347$. El modelo restringido (4.32) contiene tres variables independientes menos que (4.28) y de esta manera $q = 3$. Así, se tienen ya todos los ingredientes para calcular el estadístico F; pero esto se pospondrá hasta saber para qué sirve.

Para emplear el estadístico F es necesario conocer su distribución de muestreo bajo la hipótesis nula para de esta manera poder elegir los valores críticos y la regla de rechazo. Se puede demostrar que, bajo H_0 (y suponiendo que se satisfacen los supuestos del MLC), F está distribuida como una variable aleatoria F con $(q,n - k - 1)$ grados de libertad. Esto se escribe como

$$F \sim F_{q,n-k-1}.$$

La distribución de $F_{q,n-k-1}$ se tabula con facilidad y puede encontrarse en tablas estadísticas (vea la tabla G.3), y lo que es mejor, en software estadístico.

Aquí no se deducirá la distribución F porque esto es matemáticamente complejo. De manera básica, se puede demostrar que la ecuación (4.37) es en realidad el cociente de dos variables aleatorias independientes *ji*-cuadradas, divididas entre sus respectivos grados de libertad. La variable aleatoria *ji*-cuadrada del numerador tiene q grados de libertad y la variable aleatoria *ji*-cuadrada del denominador tiene $n - k - 1$ grados de libertad. Esta es la definición de una variable aleatoria con distribución F (vea el apéndice B).

De acuerdo con la definición de F es bastante claro que H_0 se rechazará en favor de H_1 cuando F sea suficientemente "grande". Qué tan grande depende del nivel de significancia elegido. Suponga que se ha elegido 5% como nivel de significancia. Sea c el percentil 95 en la distribución $F_{q,n-k-1}$. Este valor crítico depende de q (los *gl* en el numerador) y de $n - k - 1$ (los *gl* en el denominador). Es muy importante que los grados de libertad en el numerador y en el denominador sean los correctos.

En la tabla G.3 se dan los valores críticos correspondientes a 10, 5 y 1% de la distribución F. La regla de rechazo es sencilla. Una vez obtenida c, H_0 se rechaza a favor de H_1 al nivel de significancia elegido si

$$F > c. \tag{4.40}$$

Si el nivel de significancia es 5%, $q = 3$ y $n - k - 1 = 60$, el valor crítico es $c = 2.76$. Al nivel de 5%, H_0 se rechaza si el valor calculado para el estadístico F es mayor a 2.76. En la figura 4.7 se muestra el valor crítico correspondiente a 5% y la región de rechazo. Con los mismos grados de libertad, el valor crítico correspondiente a 1% es 4.13.

En la mayoría de las aplicaciones, el número de grados de libertad en el numerador (q) será claramente menor que el número de grados de libertad en el denominador ($n - k - 1$). Es posible que aquellas aplicaciones en las que $n - k - 1$ sea pequeño no tengan éxito debido a que tal vez los parámetros del modelo no restringido no sean estimados con precisión. Cuando el número de *gl* del denominador es aproximadamente 120, la distribución F ya no es sensible a él. (Esto es exactamente lo mismo que ocurre con la distribución t en la que a medida que el número de *gl* aumenta ésta se aproxima a la distribución normal estándar.) De manera que para los *gl* del denominador hay una entrada en la tabla que es $gl = \infty$, y es la que se usa con muestras grandes (porque entonces $n - k - 1$ es grande). Lo mismo ocurre cuando los *gl* en el numerador son grandes, pero es raro que esto se presente en las aplicaciones.

Cuando se rechaza H_0 se dice que $x_{k-q+1}, \dots, x_k$ son **estadísticamente significativas conjuntamente** (o *conjuntamente significativas*) al nivel de significancia correspondiente. Esta prueba por sí sola no permite decir cuáles de las variables tienen efecto parcial sobre y; puede ser que todas tengan efectos sobre y o que sólo algunas lo tengan. Si la hipótesis nula no se rechaza, entonces las variables son **conjuntamente no significativas**, lo que suele justificar que sean eliminadas del modelo.

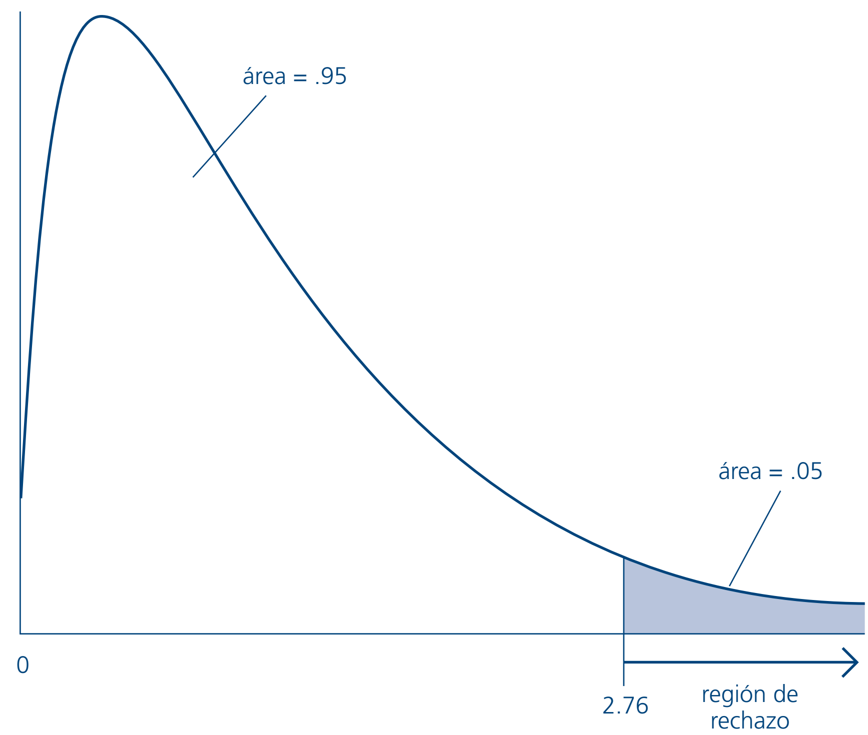
En el caso del ejemplo de la liga mayor de béisbol, en el que se tienen 3 grados de libertad en el numerador y 347 grados de libertad en el denominador, el valor crítico correspondiente a 5% es 2.60 y el valor crítico correspondiente a 1% es 3.78. H_0 se rechaza al nivel de 1% si F es mayor a 3.78, y al nivel de 5% si F es mayor a 2.60.

Ahora ya se está en condiciones de probar la hipótesis con la que se inició esta sección: una vez controladas *years* y *gamesyr* las variables *bavg*, *hrunsyr* y *rbisyr* no tienen efectos sobre los sueldos de los jugadores. En la práctica, es más fácil calcular primero $(SRC_r - SRC_{nr})/SRC_{nr}$ y multiplicar el resultado por $(n - k - 1)/q$; la razón de presentar la fórmula como en (4.37) es que esto facilita emplear el número correcto de grados de libertad en el numerador y en el denominador. Empleando las SRC de (4.31) y (4.33), se tiene

$$F = \frac{(198.311 - 183.186)}{183.186} \cdot \frac{347}{3} \approx 9.55.$$

FIGURA 4.7

Valor crítico correspondiente a 5% y región de rechazo en la distribución $F_{3,60}$.



Este número es claramente mayor que el valor crítico correspondiente a 1% en la distribución *F* con 3 y 347 grados de libertad, y por tanto se rechaza la hipótesis de que *bavg*, *hrunsyr* y *rbisyr* no tienen efectos en los sueldos.

El resultado de la prueba conjunta puede parecer sorprendente considerando los insignificantes estadísticos *t* de estas tres variables. Lo que ocurre es que las variables *hrunsyr* y *rbisyr* están fuertemente correlacionadas y esta multicolinealidad dificulta descubrir el efecto parcial de cada variable; esto se refleja en los estadísticos *t* individuales. El estadístico *F* prueba si estas tres variables (incluyendo *bavg*) son significativas *conjuntamente*, y la multicolinealidad entre *hrunsyr* y *rbisyr* al probar esta hipótesis es poco importante. En el problema 4.16, se pide al lector reestimar este modelo eliminando *rbisyr*, en cuyo caso *hrunsyr* se vuelve muy significativa. Lo mismo ocurre con *rbisyr* cuando *hrunsyr* se elimina del modelo.

El estadístico *F* suele ser útil para probar la exclusión de un grupo de variables cuando esas variables están fuertemente correlacionadas. Por ejemplo, suponga que se desea probar si el desempeño de una empresa afecta los sueldos de los directores generales o CEO. Existen muchas maneras de medir el desempeño de una empresa y puede que de antemano no resulte claro cuál de ellas sea la más apropiada. Dado que es posible que las medidas del desempeño de una empresa estén altamente correlacionadas, esperar hallar una sola medida significativa puede ser pedir demasiado debido a la multicolinealidad. Pero puede emplearse una prueba *F* para determinar si, como grupo, las variables de desempeño de la empresa afectan el sueldo.

Relación entre los estadísticos F y t

En esta sección se ha visto el empleo del estadístico F para probar si un grupo de variables debe ser incluido en un modelo. ¿Qué pasa si el estadístico F se emplea para probar la significancia de una *sola* variable independiente? Sin duda, este caso no queda excluido de acuerdo con lo visto previamente. Por ejemplo, se puede tomar como hipótesis nula H_0: $\beta_k = 0$ con $q = 1$ (para probar la restricción de exclusión de que x_k debe ser eliminada del modelo). De acuerdo con la sección 4.2, se sabe que para probar esta hipótesis puede emplearse el estadístico t para β_k. Entonces, la pregunta es: ¿se tienen dos maneras distintas de probar hipótesis acerca de un único coeficiente? La respuesta es no. Se puede demostrar que el estadístico F para probar la exclusión de una sola variable es igual al *cuadrado* del estadístico t correspondiente. Como t^2_{n-k-1} tiene una distribución $F_{1,n-k-1}$, los dos métodos conducen al mismo resultado, siempre que la alternativa sea de dos colas. Para la prueba de una sola hipótesis, el estadístico t es más flexible debido a que puede emplearse en pruebas contra alternativas de una cola. Dado que los estadísticos t son más fáciles de obtener que los F, en realidad no hay razón para emplear un estadístico F en la prueba de hipótesis de un solo parámetro.

En las regresiones sobre el sueldo de los jugadores de béisbol de la liga mayor se vio que dos (o más) variables, cada una con un estadístico t no significativo, juntas pueden ser muy significativas. También puede ocurrir que en un grupo de varias variables explicativas, una de ellas tenga un estadístico t significativo, pero que el grupo de variables sea conjuntamente no significativo a los niveles de significancia acostumbrados. ¿Cómo debe entenderse este tipo de resultados? Suponga que en un modelo que tiene muchas variables explicativas no se puede rechazar la hipótesis nula de que β_1, β_2, β_3, β_4 y β_5 sean todas igual a cero al nivel de 5%, sin embargo, el estadístico t para $\hat{\beta}_1$ sea significativo al nivel de 5%. Es claro que no se puede tener $\beta_1 \neq 0$ y, al mismo tiempo, que β_1, β_2, β_3, β_4 y β_5 ¡sean todas igual a cero! En una prueba es posible que se agrupen variables no significativas con una variable significativa y se concluya que todo el conjunto de variables es conjuntamente no significativo. (Estos conflictos entre una prueba t y una prueba conjunta F proporcionan otro ejemplo de por qué no debe decirse que se "acepta" la hipótesis nula; sólo debe decirse que no se puede rechazar.) El estadístico F sirve para detectar si un conjunto de coeficientes es distinto de cero, pero nunca será la mejor prueba para determinar si un solo coeficiente es distinto de cero. Para probar una sola hipótesis la prueba t es más adecuada. (En términos estadísticos, un estadístico F para restricciones conjuntas en las que se incluya $\beta_1 = 0$ tendrá menos potencia para detectar $\beta_1 \neq 0$ que el estadístico t usual. Vea en la sección C.6 del apéndice C una discusión sobre la potencia de una prueba.)

Por desgracia, el hecho de que algunas veces una variable estadísticamente significativa pueda ocultarse entre algunas variables no significativas puede llevar a abuso cuando los resultados de la regresión no se presentan con cuidado. Por ejemplo, suponga que, en un estudio para determinar las tasas de aprobación de crédito en una ciudad, x_1 sea la proporción de hogares negros en la ciudad. Suponga que las variables x_2, x_3, x_4 y x_5 sean proporciones de hogares con base en el grupo de edad del jefe de familia. Para explicar las tasas de aprobación de crédito, se incluirán medidas del ingreso, la riqueza, las tasas de crédito, etc. Suponga que la edad del jefe de familia tenga efecto sobre las tasas de aprobación de crédito, una vez controladas otras variables. Aun cuando la raza tenga un efecto marginalmente significativo es posible que las variables raza y edad sean conjuntamente no significativas. Alguien que desee concluir que la raza no es un factor significativo podría reportar algo así como "Las variables raza y edad fueron incluidas en la ecuación, pero resultaron conjuntamente no significativas al nivel de 5%". Tal vez una cuidadosa revisión evite esta clase de conclusiones equivocadas, pero el lector debe ser advertido de que tales resultados pueden encontrarse.

Con frecuencia, cuando una variable estadísticamente muy significativa se prueba junto con otro grupo de variables, éste es conjuntamente significativo. En tales casos no hay ninguna inconsistencia lógica al rechazar ambas hipótesis nulas.

Forma *R*-cuadrada del estadístico *F*

Para probar restricciones de exclusión, suele ser más adecuado emplear una forma del estadístico *F* que puede calcularse empleando las *R*-cuadradas de los modelos restringido y no restringido. Una razón para esto es que la *R*-cuadrada es siempre un valor que va desde cero hasta uno, mientras que las SRC pueden ser muy grandes dependiendo de las unidades de medición de *y*, lo que hace que cuando se emplean las SRC los cálculos sean bastante tediosos. Empleando el hecho de que $\text{SRC}_r = \text{STC}(1 - R_r^2)$ y $\text{SRC}_{nr} = \text{STC}(1 - R_{nr}^2)$, puede hacerse una sustitución en (4.37) y obtenerse

$$F = \frac{(R_{nr}^2 - R_r^2)/q}{(1 - R_{nr}^2)/(n - k - 1)} = \frac{(R_{nr}^2 - R_r^2)/q}{(1 - R_{nr}^2)/gl_{nr}} \tag{4.41}$$

(observe que los términos STC se cancelan en todas partes). A esto se le conoce como **forma *R*-cuadrada del estadístico *F***. [En este punto, hay que advertir que aunque la ecuación (4.41) es muy útil para probar restricciones de exclusión, no puede emplearse para probar todas las restricciones lineales. Como se verá cuando se estudien las restricciones lineales generales, algunas veces se necesita la forma suma de residuales cuadrados del estadístico *F*.]

Dado que en casi todas las regresiones se reporta la *R*-cuadrada (mientras que la SRC no), es fácil usar las *R*-cuadradas de los modelos restringido y no restringido para probar la exclusión de algunas variables. Debe ponerse especial atención en el orden de las *R*-cuadradas en el numerador: la *R*-cuadrada *no restringida* va primero [compare esto con las SRC en (4.37)]. Como $R_{nr}^2 > R_r^2$, esto demuestra, una vez más, que *F* siempre será positiva.

Al usar la forma *R*-cuadrada en esta prueba para excluir un conjunto de variables, es importante *no* elevar al cuadrado las *R*-cuadradas antes de hacer la sustitución en la fórmula (4.41); la elevación al cuadrado ya está hecha. Todas las regresiones dan R^2 y con estos números se sustituye directamente en la fórmula (4.41). En el ejemplo del sueldo en el béisbol, usando (4.41) se obtiene el estadístico *F*:

$$F = \frac{(.6278 - .5971)}{(1 - .6278)} \cdot \frac{347}{3} \approx 9.54,$$

que es muy cercano a lo que se obtuvo antes (La diferencia se debe al error de redondeo).

Ejemplo 4.9

[Educación de los padres en una ecuación de peso al nacer]

A continuación se presenta otro ejemplo para calcular un estadístico *F*. Considere el siguiente modelo para explicar el peso de un niño al nacer en términos de diversos factores:

$$\begin{aligned} bwght &= \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc \\ &\quad + \beta_4 motheduc + \beta_5 fatheduc + u, \end{aligned} \tag{4.42}$$

donde

bwght = peso al nacer, en libras.
cigs = número promedio de cigarros diarios que fumó la madre durante el embarazo.
parity = orden de nacimiento del niño.
faminc = ingreso anual de la familia.
motheduc = años de escolaridad de la madre.
fatheduc = años de escolaridad del padre.

Se probará la hipótesis nula que establece que, una vez controlados *cigs*, *parity* y *faminc*, la educación de los padres no tiene efecto sobre el peso al nacer. Esta hipótesis se enuncia como H_0: $\beta_4 = 0$, $\beta_5 = 0$ y por tanto se tienen $q = 2$ restricciones de exclusión a probar. En el modelo no restringido (4.42) hay $k + 1 = 6$ parámetros; de manera que el número de *gl* en el modelo no restringido es $n - 6$, donde n es el tamaño de la muestra.

Para probar esta hipótesis se emplearán los datos del archivo BWGHT.RAW. En este conjunto de datos se encuentra información sobre 1,388 nacimientos, pero hay que tener cuidado de contar las observaciones que se usarán en la hipótesis nula. Resulta que en 197 de estos nacimientos falta por lo menos la información muestral acerca de una de las dos variables *motheduc* y *fatheduc*; estas observaciones no pueden usarse al estimar el modelo no restringido. Entonces, en realidad, se tienen 1,191 observaciones y, por tanto, hay $1{,}191 - 6 = 1{,}185$ *gl* en el modelo no restringido. Se debe tener cuidado de emplear estas *mismas* 1,191 observaciones al estimar el modelo restringido (y no las 1,388 observaciones disponibles). En general, al estimar el modelo restringido para calcular el estadístico F, deben emplearse las mismas observaciones usadas para estimar el modelo no restringido; de no ser así, la prueba no es válida. En los casos en que no falten datos, éste no será un problema.

El número de *gl* en el numerador es 2 y en el denominador es 1,185; de acuerdo con la tabla G.3, el valor crítico correspondiente a 5% es $c = 3.0$. Para abreviar, en lugar de dar todos los resultados, se presentan sólo las R-cuadradas. La R-cuadrada para el modelo completo es $R^2_{nr} = .0387$. Eliminando de la regresión *motheduc* y *fatheduc* la R-cuadrada disminuye a $R^2_r = .0364$. Por tanto, el estadístico F es $F = [(.0387 - .0364)/(1 - .0387)](1{,}185/2) = 1.42$; como este valor es mucho menor que el valor crítico correspondiente a 5%, no se puede rechazar H_0. En otras palabras, *motheduc* y *fatheduc* son conjuntamente no significativas en la ecuación del peso al nacer.

Cálculo de los valores-*p* para pruebas *F*

Los valores-p son en especial útiles en la presentación de los resultados de las pruebas F. Como la distribución F depende de los *gl* tanto en el numerador como en el denominador, es difícil saber qué tan fuerte o qué tan débil es la evidencia contra una hipótesis nula mirando sólo los valores del estadístico F y uno o dos valores críticos.

En el contexto de una prueba F, el valor-p se define de la manera siguiente:

$$\text{valor-}p = \mathrm{P}(\mathcal{F} > F), \tag{4.43}$$

donde, con objeto de enfatizar, $\mathcal{F}$ denota una variable aleatoria F con $(q, n - k - 1)$ grados de libertad y F es el valor del estadístico de prueba. El valor-p tiene la misma interpretación que en el caso de los estadísticos: es la probabilidad de observar valores de F por lo menos tan grandes como el encontrado, *dado* que la hipótesis nula sea verdadera. Si el valor-p es pequeño, esto es una evidencia contra H_0. Por ejemplo, un valor-$p = .016$ significa que la posibilidad de observar un valor de F tan grande como el encontrado, siendo verdadera la hipótesis nula, es sólo de 1.6%; en tales casos suele rechazarse H_0. Si valor-$p = .314$, entonces la posibilidad de observar un valor del estadístico F tan grande como el encontrado, siendo verdadera la hipótesis nula, es de 31.4%. La mayoría de las personas considerarán esto como una evidencia bastante débil contra H_0.

Pregunta 4.5

Los datos del archivo ATTEND.RAW se usaron para estimar las dos ecuaciones

$$\begin{aligned}\widehat{atndrte} &= \underset{(2.87)}{47.13} + \underset{(1.09)}{13.37}\, priGPA \\ n &= 680,\ R^2 = .183\end{aligned}$$

y

$$\begin{aligned}\widehat{atndrte} &= \underset{(3.88)}{75.70} + \underset{(1.08)}{17.26}\, priGPA - \underset{(?)}{1.72}\, ACT, \\ n &= 680,\ R^2 = .291,\end{aligned}$$

donde, como siempre, los errores estándar están entre paréntesis; en esta ecuación falta el error estándar correspondiente a *ACT*. ¿Cuál es el estadístico t correspondiente al coeficiente de *ACT*? (*Sugerencia*: calcule primero el estadístico F para la significancia de *ACT*.)

Como en las pruebas t, una vez calculado el valor-p, la prueba F puede llevarse a cabo a cualquier nivel de significancia. Por ejemplo, si el valor-p = .024, H_0 se rechaza al nivel de significancia de 5% pero no al de 1%.

En el ejemplo 4.9, el valor-p en la prueba F es .238 y de esta manera la hipótesis nula que establece que tanto $\beta_{motheduc}$ como $\beta_{fatheduc}$ son cero, no se rechaza ni siquiera al nivel de significancia de 20%.

Muchos paquetes para econometría pueden probar restricciones de exclusión múltiple. Estos paquetes tienen varias ventajas sobre el cálculo manual de los estadísticos: es menos probable que se cometa un error, los valores-p se calculan de manera automática y el problema de datos faltantes, como en el ejemplo 4.9, es resuelto por estos paquetes sin mayor trabajo de nuestra parte.

El estadístico F para la significancia general de una regresión

La mayoría de los paquetes para regresión prueba de manera rutinaria un conjunto especial de restricciones de exclusión. Estas restricciones tienen una misma interpretación, sin importar cuál sea el modelo. En un modelo con k variables independientes, la hipótesis nula puede escribirse de la manera siguiente:

$$H_0\text{: } x_1, x_2, \ldots, x_k \text{ no ayudan a explicar } y.$$

Esta hipótesis nula es, de alguna manera, muy pesimista. Establece que *ninguna* de las variables explicativas tiene efecto sobre y. Enunciada en términos de los parámetros, la hipótesis nula dice que todos los parámetros de pendiente son cero:

$$H_0\text{: } \beta_1 = \beta_2 = \ldots = \beta_k = 0, \tag{4.44}$$

y la alternativa es que por lo menos uno de los β_j es distinto de cero. Otra manera de enunciar la hipótesis nula es que H_0: $E(y|x_1, x_2, \ldots, x_k) = E(y)$, de manera que conocer los valores de x_1, $x_2, \ldots, x_k$ no afecta el valor esperado de y.

En (4.44) hay k restricciones, y al imponerlas se obtiene el modelo restringido

$$y = \beta_0 + u; \tag{4.45}$$

todas las variables independientes han sido eliminadas de la ecuación. Ahora, la R-cuadrada en la estimación de (4.45) es cero; nada de la variación de y está siendo explicada porque no hay variables explicativas. Por tanto, el estadístico F para probar (4.44) puede escribirse como

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \tag{4.46}$$

donde R^2 es la R-cuadrada usual de la regresión de y sobre $x_1, x_2, \ldots, x_k$.

La mayoría de los paquetes de regresión generan de manera automática el estadístico F presentado en (4.46), lo que hace tentador usar este estadístico para probar restricciones de exclusión general. Esta tentación debe evitarse. El estadístico F dado en (4.41) es el que se usa para las restricciones de exclusión general; este estadístico depende de las R-cuadradas de los modelos restringido y no restringido. La forma especial de (4.46) sólo es válida para probar la exclusión conjunta de *todas* las variables independientes. A esto suele conocérsele como determinación de la **significancia general de la regresión**.

Si no se puede rechazar (4.44), entonces no hay evidencia de que alguna de las variables independientes ayude a explicar y. Esto, por lo general significa que se deben buscar otras variables para explicar y. En el ejemplo 4.9, el estadístico F para probar (4.44) es aproximadamente 9.55 con $k = 5$ y $n - k - 1 = 1{,}185$ gl. El valor-p es cero a cuatro decimales después del punto decimal, de manera que (4.44) se rechaza fuertemente. Por lo tanto, se concluye que las variables en la ecuación para *bwght sí* explican algo de la variación en *bwght*. La cantidad explicada no es grande: sólo 3.87%. La aparentemente pequeña R-cuadrada da como resultado un estadístico F muy significativo. Esta es la razón por la que al probar la significancia conjunta debe calcularse el estadístico F y no atender sólo a la magnitud de la R-cuadrada.

En ocasiones, el punto principal del estudio es el estadístico F para la hipótesis de que todas las variables independientes son conjuntamente no significativas. En el problema 4.10 se pide al lector que emplee datos sobre el rendimiento de acciones para probar si el rendimiento de acciones a un horizonte de cuatro años es predecible con base en la información conocida al principio del periodo. Bajo la *hipótesis de eficiencia de los mercados*, los rendimientos no deben ser predecibles; la hipótesis nula es precisamente (4.44).

Prueba para las restricciones generales lineales

Probar restricciones de exclusión es, con mucho, la aplicación más importante de los estadísticos F. Sin embargo, algunas veces las restricciones implicadas por una teoría son más complicadas que la sola exclusión de algunas variables independientes. También en este caso puede emplearse el estadístico F.

Como ejemplo, considere la ecuación siguiente:

$$\log(price) = \beta_0 + \beta_1\log(assess) + \beta_2\log(lotsize) + \beta_3\log(sqrft) + \beta_4 bdrms + u, \tag{4.47}$$

donde

price = precio de la casa.
assess = avalúo de la casa (antes de que fuera vendida).
lotsize = tamaño del terreno, en pies.
sqrft = superficie en pies cuadrados.
bdrms = cantidad de recámaras.

Ahora suponga que se desea probar si el avalúo de la casa es razonable. Si es este el caso, entonces una variación de 1% en *assess* deberá corresponder a una variación de 1% en *price*; es decir, $\beta_1 = 1$. Además, una vez controlado el avalúo de la casa, *lotsize*, *sqrft* y *bdrms* no deberán ayudar a explicar log(*price*). Estas hipótesis juntas pueden expresarse como

$$\mathrm{H}_0\colon \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0. \tag{4.48}$$

Hay que probar cuatro restricciones; tres son de exclusión, pero $\beta_1 = 1$ no lo es. ¿Cómo puede probarse esta hipótesis usando el estadístico F?

Como en el caso de la restricción de exclusión, se estima el modelo no restringido, (4.47) en este caso, y después se imponen las restricciones en (4.48) para así obtener el modelo restringido. El segundo paso es un poco complicado. Pero lo único que hay que hacer es sustituir las restricciones. Si (4.47) se escribe como

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + u, \tag{4.49}$$

entonces el modelo restringido es $y = \beta_0 + x_1 + u$. Ahora, para imponer la restricción de que el coeficiente de x_1 es la unidad, hay que estimar el modelo siguiente:

$$y - x_1 = \beta_0 + u. \tag{4.50}$$

Este es un modelo con un intercepto (β_0) pero con una variable independiente distinta de la que aparece en (4.49). El procedimiento para calcular el estadístico F es el mismo: estimar (4.50), obtener la SRC (SRC_r) y emplearla con la SRC no restringida de (4.49) en el estadístico F dado en (4.37). Se están probando $q = 4$ restricciones, y en el modelo no restringido hay $n - 5$ gl. El estadístico F es simplemente $[(\text{SRC}_r - \text{SRC}_{ur})/\text{SRC}_{ur}][(n - 5)/4]$.

Antes de ejemplificar esta prueba usando un conjunto de datos, es necesario hacer hincapié en un punto: en este ejemplo no puede emplearse la forma R-cuadrada del estadístico F debido a que la variable dependiente en (4.50) es diferente de la variable independiente en (4.49). Esto significa que la suma total de cuadrados de las dos regresiones será diferente y entonces (4.41) ya no es equivalente a (4.37). Como regla general, si en la regresión restringida se necesita una variable dependiente diferente debe emplearse la forma SRC del estadístico F.

El modelo no restringido estimado empleando los datos del archivo HPRICE1.RAW es

$$
\begin{aligned}
\widehat{\log(price)} = \underset{(.570)}{.264} &+ \underset{(.151)}{1.043}\log(assess) + \underset{(.0386)}{.0074}\log(lotsize) \\
&- \underset{(.1384)}{.1032}\log(sqrft) + \underset{(.0221)}{.0338}\,bdrms \\
n = 88,\ \text{SRC} = 1.822,\ R^2 = .773.&
\end{aligned}
$$

Empleando por separado estadísticos t para probar cada una de las hipótesis en (4.48), no puede rechazarse cada una. Pero la racionalidad de la evaluación es una hipótesis conjunta, de manera que las restricciones deben probarse conjuntamente. La SRC del modelo restringido es $\text{SRC}_r = 1.880$, y entonces el estadístico F es $[(1.880 - 1.822)/1.822](83/4) = .661$. El valor crítico de 5% para una distribución F con gl (4,83) gl es aproximadamente 2.50, por lo que H_0 no puede rechazarse. En esencia, no hay ninguna evidencia contra la hipótesis de que el avalúo de la casa sea razonable.

4.6 Informe de los resultados de la regresión

Para concluir este capítulo se proporcionan algunos lineamientos de cómo emplear los resultados de la regresión múltiple cuando se trata de proyectos empíricos algo complicados. Se espera que esto le ayude a leer trabajos publicados sobre las ciencias sociales aplicadas y, al mismo tiempo, lo prepare para escribir sus propias publicaciones empíricas. Este tema será ampliado en el resto del libro al dar los resultados de diversos ejemplos, pero muchos de los puntos clave pueden ser vistos ahora.

Naturalmente, siempre deben darse los coeficientes estimados de MCO. Los coeficientes estimados de las variables clave del análisis deben *interpretarse* (para lo que suele ser necesario conocer las unidades de medición de las variables). Por ejemplo, ¿una estimación es una elasticidad o tiene alguna otra interpretación que requiera explicación? La importancia económica o práctica de las estimaciones de las variables clave debe ser analizada.

Siempre deben darse los errores estándar junto con los coeficientes estimados. Algunos autores prefieren dar los estadísticos t en lugar de los errores estándar (y algunas veces incluso el valor absoluto de los estadísticos t). Aunque, en realidad, esto no es ningún problema, se prefiere dar los errores estándar. Primero, esto obliga a reflexionar con cuidado acerca de la hipótesis nula que se está probando; la hipótesis nula no siempre es que los parámetros poblacionales sean cero. Segundo, si se tienen los errores estándar es más fácil calcular intervalos de confianza.

La R-cuadrada de la regresión debe incluirse siempre; como se ha visto, además de proporcionar una medida de la bondad de ajuste, facilita los cálculos del estadístico F en las restricciones de exclusión. Suele ser una buena idea dar la suma de los residuales cuadrados y el error estándar de la regresión, pero esto no es crucial. La cantidad de observaciones empleadas para estimar cualquier ecuación debe aparecer cerca de la ecuación estimada.

Si sólo se estima un par de modelos, los resultados pueden resumirse en forma de ecuación, como se ha hecho aquí hasta ahora. Sin embargo, en muchas publicaciones se estiman varias ecuaciones con conjuntos diferentes de variables independientes. Puede ser que se estime una misma ecuación con diferentes grupos de personas o incluso tener ecuaciones que expliquen diferentes variables dependientes. En tales casos, es mejor resumir los resultados en una o en varias tablas. En la tabla debe indicarse con claridad cuál es la variable dependiente, y las variables independientes deben presentarse en una lista en la primera columna. Los errores estándar (o estadísticos t) pueden colocarse entre paréntesis debajo de los coeficientes estimados.

Ejemplo 4.10

[Conmutación entre sueldo y pensión para profesores]

Sea *totcomp* la compensación total promedio anual de un profesor, incluyendo sueldo y todas las prestaciones o beneficios (pensión, seguro médico, etc.). Extendiendo la ecuación estándar para salario, la compensación total debe ser una función de la productividad y de otras características. Como es usual, se usa la forma logarítmica:

$$\log(totcomp) = f(\textit{características de productividad, otros factores}),$$

donde $f(\cdot)$ es una función (por ahora no especificada). Se escribe

$$totcomp = salary + benefits = salary\left(1 + \frac{benefits}{salary}\right).$$

Esta ecuación muestra que la compensación total es el producto de dos términos: *salary* (es decir, el sueldo) y $1 + b/s$, donde b/s es un abreviación para el "cociente beneficios-sueldo". Tomando logaritmos de esta ecuación se tiene $\log(totcomp) = \log(salary) + \log(1 + b/s)$. Ahora, como en el caso de que b/s sea "pequeño", $\log(1 + b/s) \approx b/s$, se empleará esta aproximación. Esto conduce al modelo econométrico

$$\log(salary) = \beta_0 + \beta_1(b/s) + \textit{otros factores}.$$

Probar la conmutación entre sueldo y beneficios (*salary-benefits*) es entonces lo mismo que hacer una prueba de H_0: $\beta_1 = -1$ contra H_1: $\beta_1 \neq -1$.

Para probar esta hipótesis se emplean los datos del archivo MEAP93.RAW. Estos datos son promediados con base en el nivel escolar y no se observan muchos otros factores que podrían afectar la compensación total. Se incluirán variables de control para el tamaño de la escuela (*enroll*) y la cantidad de personal por cada mil estudiantes (*staff*), además de otros datos como las tasas de deserción (*droprate*) y la tasa de graduación (*gradrate*). En la muestra el b/s promedio es .205 y el valor mayor es .450.

En la tabla 4.1 se dan las ecuaciones estimadas, donde los errores estándar aparecen entre paréntesis debajo de los coeficientes estimados. La variable clave es b/s, el cociente beneficios-sueldo.

De acuerdo con la primera columna de la tabla 4.1, se ve que, cuando no se controla ninguno de los otros factores, el coeficiente de MCO para b/s es $-.825$. El estadístico t para probar la hipótesis nula H_0: $\beta_1 = -1$ es $t = (-.825 + 1)/.200 = .875$ y, por tanto, la regresión simple no permite rechazar H_0. Después de agregar las variables de control para tamaño de la escuela y cantidad de personal (que corresponde

TABLA 4.1

Prueba de la conmutación entre sueldo y beneficios

Variable dependiente: log(*salary*)			
Variables independientes	**(1)**	**(2)**	**(3)**
b/s	−.825 (.200)	−.605 (.165)	−.589 (.165)
log(*enroll*)	——	.0874 (.0073)	.0881 (.0073)
log(*staff*)	——	−.222 (.050)	−.218 (.050)
droprate	——	——	−.00028 (.00161)
gradrate	——	——	.00097 (.00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observaciones *R*-cuadrada	408 .040	408 .353	408 .361

Pregunta 4.6

La adición de *droprate* y *gradrate*, ¿cómo afecta a la conmutación estimada entre sueldo-beneficios?, ¿son estas variables conjuntamente significativas al nivel de 5%?, ¿y al nivel de 10%?

aproximadamente a la cantidad de estudiantes por profesor), el coeficiente estimado de *b/s* se convierte en −.605. Ahora la prueba de $\beta_1 = -1$ da un estadístico t de aproximadamente 2.39; por tanto, H_0 se rechaza al nivel de 5% contra la alternativa de dos colas. Las variables log(*enroll*) y log(*staff*) son muy significativas estadísticamente.

RESUMEN

En este capítulo se ha visto el importante tema de la inferencia estadística, que permite inferir algo acerca del modelo poblacional a partir de una muestra aleatoria. A continuación se resumen los puntos principales:

1. Bajo los supuestos del modelo lineal clásico RLM.1 a RLM.6 los estimadores de MCO están distribuidos normalmente.
2. Bajo los supuestos del MLC, los estadísticos t tienen distribuciones t bajo la hipótesis nula.
3. Los estadísticos t se usan para probar hipótesis acerca de un solo parámetro contra alternativas de una o de dos colas usando pruebas de una o dos colas, respectivamente. La hipótesis nula más común es H_0: $\beta_j = 0$, pero algunas veces, en H_0, se desean probar otros valores de β_j.
4. En las pruebas de hipótesis clásicas, primero se elige un nivel de significancia, el cual, junto con los *gl* y la hipótesis alternativa, determinan el valor crítico contra el cual se compara el estadístico t. En una prueba t es más informativo calcular el valor-p —el menor nivel de

significancia al que puede rechazarse la hipótesis nula— de manera que la hipótesis pueda probarse a cualquier nivel de significancia.

5. Bajo los supuestos del MLC pueden construirse intervalos de confianza para cada β_j. Estos IC pueden emplearse para probar cualquier hipótesis nula respecto a β_j contra una alternativa de dos colas.
6. Pruebas para una sola hipótesis que se refiera a más de un β_j pueden realizarse escribiendo el modelo de una manera tal que contenga el parámetro de interés. Después, puede emplearse el estadístico *t* estándar.
7. El estadístico *F* se usa para probar restricciones de exclusión múltiple y existen dos formas equivalentes de la prueba. Una se basa en las SRC de los modelos restringido y no restringido. Otra, más fácil de usar, se basa en las *R*-cuadradas de los dos modelos.
8. Al calcular el estadístico *F*, el número de *gl* en el numerador es la cantidad de restricciones que se prueban, mientras que el número de *gl* en el denominador es el número de grados de libertad en el modelo no restringido.
9. La alternativa en las pruebas *F* es de una cola. En el método clásico se especifica un nivel de significancia el cual, junto con los *gl* en el numerador y los *gl en* el denominador, determina el valor crítico. La hipótesis nula se rechaza cuando el estadístico, *F*, es mayor al valor crítico, *c*. Otra alternativa es calcular el valor-*p* para resumir las evidencias contra H_0.
10. Restricciones lineales múltiples generales pueden probarse empleando la forma suma de residuales cuadrados del estadístico *F*.
11. El estadístico *F* para la significancia general de una regresión prueba la hipótesis nula de que *todos* los parámetros de pendiente son cero, sin restringir el intercepto. Bajo H_0, las variables explicativas no tienen ningún efecto sobre el valor esperado de *y*.

Supuestos del modelo lineal clásico

Ahora es un momento adecuado para revisar el conjunto completo de supuestos del modelo lineal clásico (MLC) para la regresión de corte transversal. Después de cada supuesto se encuentra un comentario acerca de su papel en el análisis de regresión múltiple.

Supuesto RLM.1 (Lineal en los parámetros)

El modelo poblacional puede escribirse como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros (constantes) desconocidos de interés y u es un error aleatorio no observable o término de perturbación.

El supuesto RLM.1 describe la relación poblacional que se espera estimar y explícitamente establece a las β_j —los efectos poblacionales *ceteris paribus* de x_j sobre y— como los parámetros de interés.

Supuesto RLM.2 (Muestreo aleatorio)

Se tiene una muestra aleatoria de n observaciones, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, \ldots, n\}$, que satisface el modelo poblacional del supuesto RLM.1.

Este supuesto del muestreo aleatorio significa que se tienen datos que pueden emplearse para estimar las β_j, y que estos datos han sido elegidos de manera que sean representativos de la población descrita en el supuesto RLM.1.

Supuesto RLM.3 (Colinealidad no perfecta)

En la muestra (y por tanto en la población), ninguna de las variables independientes es constante y no hay una relación *lineal* constante entre las variables independientes.

Una vez que se tiene una muestra de datos, se necesita saber que éstos pueden emplearse para calcular las estimaciones de MCO, las $\hat{\beta}_j$. Este es el papel del supuesto RLM.3: si en cada variable independiente hay variaciones muestrales y no existe una relación lineal exacta entre las variables independientes, las $\hat{\beta}_j$ pueden calcularse.

Supuesto RLM.4 (Media condicional cero)

Para cualesquiera valores de las variables explicativas, el error u tiene un valor esperado de cero. En otras palabras, $\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$.

Como se analizó, suponer que los efectos no observables no están, en promedio, relacionados con las variables explicativas es clave para obtener la primera propiedad estadística de cada estimador de MCO: cada estimador es insesgado respecto al parámetro poblacional correspondiente. Claro, todos los supuestos anteriores se emplean para demostrar el insesgamiento.

Supuesto RLM.5 (Homocedasticidad)

Para cualesquiera valores de las variables explicativas, el error u tiene la misma varianza. En otras palabras,

$$\mathrm{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2.$$

En comparación con el supuesto RLM.4, el supuesto de homocedasticidad tiene importancia secundaria; en particular, el supuesto RLM.5 no tiene relación con el insesgamiento de las $\hat{\beta}_j$. Sin embargo, la homocedasticidad tiene dos consecuencias importantes: 1) permite obtener fórmulas para la varianza de muestreo cuyos componentes son fáciles de caracterizar; 2) bajo los supuestos de Gauss-Markov RLM.1 a RLM.5 puede concluirse que los estimadores de MCO son los que tienen menor varianza entre *todos* los estimadores lineales insesgados.

Supuesto RLM.6 (Normalidad)

El error poblacional u es *independiente* de las variables explicativas $x_1, x_2, \ldots, x_k$ y está distribuido normalmente siendo su media cero y su varianza σ^2: $u \sim \mathrm{Normal}(0, \sigma^2)$.

En este capítulo se agregó el supuesto RLM.6 para obtener las distribuciones de muestreo exactas de los estadísticos t y de los estadísticos F de manera que puedan realizarse pruebas de hipótesis exactas. En el siguiente capítulo se verá que RLM.6 puede omitirse si se tiene un tamaño de muestra razonablemente grande. El supuesto RLM. 6 sí implica una propiedad de eficiencia más fuerte de los MCO: los estimadores de MCO son los que tienen la menor varianza entre todos los estimadores insesgados; el grupo de comparación ya no está restringido a los estimadores lineales en la $\{y_i: i = 1, 2, \ldots, n\}$.

TÉRMINOS CLAVE

Alternativa de dos colas
Alternativa de una cola
Cociente t
Conjuntamente no significativas
Estadísticamente no significativa
Estadísticamente significativa
Estadísticamente significativas conjuntamente
Estadístico F
Estadístico t
Estimadores insesgados de varianza mínima
Forma R-cuadrada del estadístico F
Grados de libertad en el denominador
Grados de libertad en el numerador

Hipótesis alternativa
Hipótesis nula
Intervalo de confianza (IC)
Modelo lineal clásico
Modelo no restringido
Modelo restringido
Nivel de significancia
Prueba de dos colas
Prueba de hipótesis conjunta
Prueba de hipótesis múltiple
Prueba de una cola
Regla de rechazo
Restricciones de exclusión
Restricciones múltiples
Significancia económica
Significancia general de la regresión
Significancia práctica
Supuesto de normalidad
Supuestos del modelo lineal clásico (MLC)
Valor crítico
Valor-*p*

PROBLEMAS

4.1 ¿Cuál de las causas siguientes puede hacer que los estadísticos *t* usuales de MCO no sean válidos (es decir, que no tengan una distribución *t* bajo H_0)?

i) Heterocedasticidad.

ii) Que exista un coeficiente de correlación muestral de .95 entre dos variables independientes del modelo.

iii) Omitir una variable explicativa importante.

4.2 Considere una ecuación para explicar los sueldos de los directores generales o CEO en términos de las ventas anuales de la empresa, el rendimiento sobre capital (*roe*, en forma de porcentaje), y el rendimiento de las acciones de la empresa (*ros*, en forma de porcentaje):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

i) En términos de los parámetros del modelo, establezca la hipótesis nula de que, controlando *sales* y *roe*, *ros* no tiene efecto en el sueldo de los CEO. Establezca la alternativa de que un mejor desempeño de las acciones de la empresa incrementa el sueldo de los CEO.

ii) Con los datos de CEOSAL1.RAW, empleando MCO se obtuvo la ecuación siguiente:

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(.32)}{4.32} + \underset{(.035)}{.280} \log(sales) + \underset{(.0041)}{.0174}\, roe + \underset{(.00054)}{.00024}\, ros \\ n &= 209,\ R^2 = .283.\end{aligned}$$

¿Cuál es el porcentaje de aumento de *salary* que se pronostica si *ros* aumenta 50 puntos? ¿Tienen *ros* un efecto práctico grande sobre *salary*?

iii) Pruebe la hipótesis nula que dice que *ros* no tiene efecto sobre *salary* contra la alternativa que dice que *ros* tiene efecto positivo. Realice la prueba al nivel de significancia de 10%.

iv) ¿Incluiría usted *ros* en el modelo final que explica las compensaciones de los CEO en términos del desempeño de la empresa? Explique.

4.3 La variable *rdintens* representa el gasto en investigación y desarrollo (I & D) dado como porcentaje de las ventas. Las ventas (*sales*) se miden en millones de dólares. La variable *profmarg* representa la ganancia como porcentaje de las ventas.

Empleando los datos del archivo RDCHEM.RAW de 32 empresas de la industria química, se estimó la ecuación siguiente:

$$\begin{aligned}\widehat{rdintens} &= \underset{(1.369)}{.472} + \underset{(.216)}{.321} \log(sales) + \underset{(.046)}{.050}\, profmarg \\ n &= 32,\ R^2 = .099.\end{aligned}$$

i) Interprete el coeficiente de log(*sales*). En particular, si *sales* aumenta 10%, ¿cuál es la variación estimada en puntos porcentuales en *rdintens*? ¿Es este efecto económicamente grande?

ii) Pruebe la hipótesis de que la intensidad de la I & D no varía con *sales* contra la alternativa de que aumenta con las ventas. Realice la prueba a los niveles de significancia de 5 y 10%.

iii) Interprete el coeficiente de *profmarg*. ¿Es este coeficiente económicamente grande?

iv) ¿Tiene *profmarg* un efecto estadístico significativo sobre *rdintens*?

4.4 En una ciudad estudiantil, ¿influye la población de estudiantes sobre las rentas de las viviendas? Sea *rent* la renta mensual promedio en una ciudad estudiantil de Estados Unidos. Sean *pop* el total de la población en esa ciudad, *avginc* el ingreso promedio en la ciudad y *pctstu* la población de estudiantes dada como porcentaje del total de la población. Un modelo para probar esta relación es

$$\log(rent) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(avginc) + \beta_3 pctstu + u.$$

i) Dé la hipótesis nula que establece que el tamaño del cuerpo estudiantil en relación con la población no tiene efecto *ceteris paribus* sobre las rentas mensuales. Mencione la alternativa que establece que sí tiene efecto.

ii) ¿Qué signo espera que tengan β_1 y β_2?

iii) La ecuación estimada empleando datos de 1990 del archivo RENTAL.RAW sobre 64 ciudades estudiantiles es:

$$\begin{aligned}\widehat{\log(rent)} = \underset{(.844)}{.043} + \underset{(.039)}{.066}\log(pop) + \underset{(.081)}{.507}\log(avginc) + \underset{(.0017)}{.0056}\,pctstu\\ n = 64,\ R^2 = .458.\end{aligned}$$

¿Qué está equivocado en la siguiente afirmación: "Un aumento de 10% en la población corresponde aproximadamente a un aumento de 6.6% en la renta"?

iv) Pruebe la hipótesis establecida en el inciso i) al nivel de 1%

4.5 Considere la ecuación estimada del ejemplo 4.3, la cual se emplea para estudiar el efecto de faltar a clases en el promedio general (GPA) en la universidad:

$$\begin{aligned}\widehat{colGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{.412}\,hsGPA + \underset{(.011)}{.015}\,ACT - \underset{(.026)}{.083}\,skipped\\ n = 141,\ R^2 = .234.\end{aligned}$$

i) Empleando la aproximación normal estándar, encuentre el intervalo de confianza de 95% para β_{hsGPA}.

ii) ¿Se puede rechazar la hipótesis H_0: $\beta_{hsGPA} = .4$ contra la alternativa de dos colas al nivel de 5%?

iii) ¿Se puede rechazar la hipótesis H_0: $\beta_{hsGPA} = 1$ contra la alternativa de dos colas al nivel de 5%?

4.6 En la sección 4.5 se empleó como ejemplo la prueba de si son razonables los avalúos de precios de casas. Ahí se empleó el modelo log-log para *price* y *assess* [vea la ecuación (4.47)]. Aquí se emplea la formulación nivel-nivel.

i) En el modelo de regresión simple

$$price = \beta_0 + \beta_1 assess + u,$$

el avalúo es racional si $\beta_1 = 1$ y $\beta_0 = 0$. La ecuación estimada es

$$\begin{aligned}\widehat{price} = \underset{(16.27)}{-14.47} + \underset{(.049)}{.976}\,assess\\ n = 88,\ \text{SRC} = 165{,}644.51,\ R^2 = .820.\end{aligned}$$

Primero pruebe la hipótesis H_0: $\beta_0 = 0$ contra la alternativa de dos colas. Después pruebe H_0: $\beta_1 = 1$ contra la alternativa de dos colas. ¿Qué concluye?

ii) Para probar la hipótesis conjunta de que $\beta_0 = 0$ y $\beta_1 = 1$, se necesita la SRC del modelo restringido. Esto es equivalente a calcular $\sum_{i=1}^{n}(price_i - assess_i)^2$, donde $n = 88$, ya que los residuales en el modelo restringido son precisamente las diferencias $price_i - assess_i$. (No se necesita ninguna estimación del modelo restringido ya que ambos parámetros son especificados bajo H_0.) Con esto se obtiene SRC = 209,448.99. Realice la prueba F de la hipótesis conjunta.

iii) Ahora, pruebe H_0: $\beta_2 = 0$, $\beta_3 = 0$ y $\beta_4 = 0$ en el modelo

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u.$$

La R-cuadrada de la estimación de este modelo usando las mismas 88 casas es .829.

iv) Si la varianza de *price* cambia con *assess*, *lotsize*, *sqrft* o *bdrms*, ¿qué se puede decir acerca de la prueba F del inciso iii)?

4.7 En el ejemplo 4.7 se usaron datos sobre empresas no sindicalizadas para estimar la relación entre las tasas de piezas defectuosas (*scrap*) y otras características de la empresa. Este ejemplo se verá ahora más de cerca y se emplearán todas las empresas disponibles.

i) El modelo poblacional estimado en el ejemplo 4.7 puede escribirse como

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

Empleando las 43 observaciones disponibles de 1987, la ecuación estimada es

$$\begin{aligned}\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales) + \underset{(.360)}{.992}\, \log(employ)\end{aligned}$$
$$n = 43,\ R^2 = .310.$$

Compare esta ecuación con la estimada empleando sólo las 29 empresas no sindicalizadas de la muestra.

ii) Muestre que el modelo poblacional puede expresarse también de la manera siguiente

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

donde $\theta_3 = \beta_2 + \beta_3$. [*Sugerencia:* recuerde que $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.] Interprete la hipótesis H_0: $\theta_3 = 0$.

iii) Estimando la ecuación del inciso ii) se obtiene

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales/employ) + \underset{(.205)}{.041}\, \log(employ)$$
$$n = 43,\ R^2 = .310.$$

Controlando la capacitación de los trabajadores (*hrsemp*) y el cociente ventas-empleado (*sales/employ*), ¿tienen las firmas más grandes tasas de piezas defectuosas estadísticamente significativas mayores?

iv) Pruebe la hipótesis de que un aumento de 1% en *sales/employ* corresponde a una disminución de 1% en la tasa de piezas defectuosas.

4.8 Considere el siguiente modelo de regresión múltiple con tres variables independientes, bajo los supuestos RLM.1 a RLM.6 del modelo lineal clásico:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Usted desea probar la hipótesis nula H_0: $\beta_1 - 3\beta_2 = 1$.

i) Sean $\hat{\beta}_1$ y $\hat{\beta}_2$ los estimadores de MCO de β_1 y β_2. Encuentre $\mathrm{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ en términos de las varianzas de $\hat{\beta}_1$ y $\hat{\beta}_2$ y de la covarianza entre ellas. ¿Cuál es el error estándar de $\hat{\beta}_1 - 3\hat{\beta}_2$?

ii) Dé el estadístico t para probar H_0: $\beta_1 - 3\beta_2 = 1$.

iii) Defina $\theta_1 = \beta_1 - 3\beta_2$ y $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Escriba una ecuación de regresión en la que se incluyan β_0, θ_1, β_2 y β_3 que le permita obtener directamente $\hat{\theta}_1$ y su error estándar.

4.9 En el problema 3.3, se estimó la ecuación

$$\begin{aligned}\widehat{sleep} = \underset{(112.28)}{3{,}638.25} - \underset{(.017)}{.148}\ totwrk - \underset{(5.88)}{11.13}\ educ + \underset{(1.45)}{2.20}\ age\\ n = 706,\ R^2 = .113,\end{aligned}$$

donde ahora se dan los errores estándar junto con sus coeficientes estimados.

i) ¿Es, *educ* o *age* individualmente significativa al nivel de 5% contra una alternativa de dos colas? Presente su trabajo.

ii) Eliminando del modelo *educ* y *age* se obtiene

$$\begin{aligned}\widehat{sleep} = \underset{(38.91)}{3{,}586.38} - \underset{(.017)}{.151}\ totwrk\\ n = 706,\ R^2 = .103.\end{aligned}$$

Al nivel de significancia de 5%, ¿son *educ* y *age* conjuntamente significativas en la ecuación original? Justifique su respuesta.

iii) ¿Afecta mucho a la disyuntiva entre sueño (*sleep*) y trabajo (tot*wrk*) incluir *educ* y *age* en el modelo?

iv) Suponga que la ecuación para sueño contiene heterocedasticidad. ¿Qué significa esto en relación con las pruebas calculadas en los incisos i) y ii)?

4.10 El análisis de regresión puede emplearse para probar si el mercado emplea de manera eficiente la información sobre valuación de acciones. En concreto, sea *return* el rendimiento total de conservar una acción de una empresa durante el periodo de cuatro años que va desde fines de 1990 hasta fines de 1994. La *hipótesis de los mercados eficientes* dice que estos rendimientos no deben estar relacionados de manera sistemática con la información conocida en 1990. Si las características conocidas de una empresa al principio del periodo ayudaran para predecir los rendimientos de las acciones, entonces esta información podría usarse para elegir las acciones.

Para 1990, sea *dkr* el cociente de deuda sobre capital de una empresa, *eps* sean las ganancias por acción, *netinc* sea el ingreso neto y *salary* la compensación total del director general.

i) Empleando los datos en el archivo RETURN.RAW, se estimó la ecuación siguiente:

$$\begin{aligned}\widehat{return} = \underset{(6.89)}{-14.37} + \underset{(.201)}{.321}\ dkr + \underset{(.078)}{.043}\ eps - \underset{(.0047)}{.0051}\ netinc + \underset{(.0022)}{.0035}\ salary\\ n = 142,\ R^2 = .0395.\end{aligned}$$

Pruebe si al nivel de significancia de 5% las variables explicativas son conjuntamente significativas. ¿Es alguna de las variables explicativas individualmente significativa?

ii) Ahora se estima el modelo empleando la forma logarítmica de *netinc* y de *salary*:

$$\begin{aligned}\widehat{return} = \underset{(39.37)}{-36.30} + \underset{(.203)}{.327}\ dkr + \underset{(.080)}{.069}\ eps - \underset{(3.39)}{4.74}\ \log(netinc) + \underset{(6.31)}{7.24}\ \log(salary)\\ n = 142,\ R^2 = .0330.\end{aligned}$$

¿Se modifica alguna de sus conclusiones del inciso i)?

iii) En esta muestra, algunas de las empresas tienen deuda cero y otras tienen ganancias negativas. ¿Debe tratar de emplearse en el modelo log(*dkr*) o log(*eps*) para ver si con esto mejora el ajuste? Explique.

iv) En general, ¿es fuerte o débil la evidencia para la predictibilidad de los rendimientos de las acciones?

4.11 La tabla siguiente se obtuvo empleando los datos del archivo CEOSAL2.RAW:

Variable dependiente: log(*salary*)			
Variables independientes	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observaciones *R*-cuadrada	177 .281	177 .304	177 .353

La variable *mktval* es el valor de mercado de la empresa, *profmarg* es la ganancia expresada como porcentaje de las ventas, *ceoten* son los años como director general de la empresa y *comten* son los años de antigüedad en la empresa del director general.

i) Analice el efecto de *profmarg* sobre el sueldo del director general.

ii) ¿Tiene un efecto significativo el valor de mercado? Explique.

iii) Interprete los coeficientes de *ceoten* y de *comten*. ¿Son estas variables explicativas estadísticamente significativas?

iv) ¿Qué opina del hecho de que una mayor antigüedad en la empresa, manteniendo todos los demás factores constantes, corresponda a un sueldo menor?

EJERCICIOS EN COMPUTADORA

C4.1 El modelo siguiente puede usarse para estudiar si los gastos de campaña afectan los resultados de las elecciones:

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

donde *voteA* es el porcentaje de votos recibidos por el candidato A, *expendA* y *expendB* son los gastos de campaña del candidato A y del candidato B y *prtystrA* es una medida de la fortaleza del partido del candidato A (el porcentaje de votos que obtuvo el partido de A en la elección presidencial más reciente).

i) ¿Cuál es la interpretación de β_1?

ii) En términos de los parámetros, establezca la hipótesis nula de que un aumento de 1% en los gastos de A es compensado por un aumento de 1% en los gastos de B.

iii) Estime el modelo dado usando los datos del archivo VOTE1.RAW y presente los resultados de la manera usual. ¿Los gastos de A afectan los resultados de las elecciones? ¿Y los gastos de B? ¿Puede usar estos resultados para probar la hipótesis del inciso ii)?

iv) Estime el modelo que proporciona directamente el estadístico t para probar la hipótesis del inciso ii). ¿Qué concluye usted? (Use una alternativa de dos colas.)

C4.2 Para este ejercicio utilice los datos del archivo LAWSCH85.RAW.

i) Usando el modelo del problema 3.4, establezca y pruebe la hipótesis nula de que el ranking entre las escuelas de derecho no tiene efecto *ceteris paribus* sobre el sueldo inicial medio.

ii) ¿Son las características de los estudiantes de nuevo ingreso —a saber, *LSAT* y *GPA*— significativas de manera individual o conjunta para explicar el sueldo (*salary*)? (Asegúrese de tomar en cuenta que hay datos faltantes en *LSAT* y *GPA*.)

iii) Pruebe si el tamaño del grupo (*clsize*) o el tamaño de la facultad (*faculty*) necesitan ser agregados a esta ecuación; realice una prueba sencilla. (Tenga cuidado de tomar en cuenta que hay datos faltantes en *clsize* y *faculty*.)

iv) ¿Qué factores pueden influir en la posición en el ranking de una escuela de leyes que no están incluidos en la regresión del sueldo?

C4.3 Vuelva al problema **3.14.** Como variable dependiente emplee ahora el log del precio de la vivienda:

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u.$$

i) Usted desea estimar y obtener un intervalo de confianza para la variación porcentual del precio (*price*) cuando una casa tiene una recámara adicional de 150 pies cuadrados. En forma decimal, esto es $\theta_1 = 150\beta_1 + \beta_2$. Emplee los datos del archivo HPRICE1.RAW para estimar θ_1.

ii) Exprese β_2 en términos de θ_1 y β_1 y sustituya esto en la ecuación de $\log(price)$.

iii) Emplee el inciso ii) para obtener el error estándar de $\hat{\theta}_1$ y con este error estándar construya un intervalo de confianza de 95%.

C4.4 En el ejemplo 4.9 puede estimarse la versión restringida del modelo empleando todas las 1,388 observaciones muestrales. Calcula R-cuadrada de la regresión de *bwght* sobre *cigs*, *parity* y *faminc* empleando todas las observaciones. Compare esto con la R-cuadrada dada para el modelo restringido del ejemplo 4.9.

C4.5 Para este ejercicio emplee los datos en el archivo MLB1.RAW.

i) Emplee el modelo estimado en la ecuación (4.31) y elimine la variable *rbisyr*. ¿Qué pasa con la significancia estadística de *hrunsyr*? ¿Qué pasa con el tamaño del coeficiente de *hrunsyr*?

ii) Al modelo del inciso i) agregue las variables *runsyr* (carreras por año), *fldperc* (porcentaje de fildeo) y *sbasesyr* (bases robadas por año). ¿Cuáles de estos factores son significativos individualmente?

iii) En el modelo del inciso ii) pruebe la significancia conjunta de *bavg*, *fldperc* y *sbasesyr*.

C4.6 Para este ejercicio emplee los datos del archivo WAGE2.RAW.

i) Considere la ecuación estándar para salario

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Establezca la hipótesis nula de que un año más de experiencia en la fuerza de trabajo general tiene el mismo efecto sobre $\log(wage)$ que un año más de antigüedad en el empleo actual.

ii) Al nivel de significancia de 5% pruebe la hipótesis nula del inciso i) contra la alternativa de dos colas construyendo un intervalo de confianza del 95%. ¿Qué es lo que concluye usted?

C4.7 Vaya al ejemplo empleado en la sección 4.4. Use los datos del archivo TWOYEAR.RAW.

i) La variable *phsrank* es el percentil de la persona en el ranking del bachillerato (Cuanto más alto sea el número es mejor. Por ejemplo, 90 significa que usted está considerado mejor que el 90% de los que terminaron el bachillerato en ese año). Determine el *phsrank* menor, mayor y promedio en la muestra.

ii) Agregue a la ecuación (4.26) la variable *phsrank* y dé las estimaciones de MCO en la forma usual. ¿Es *phsrank* estadísticamente significativa? ¿Cuál es el valor de 10 puntos porcentuales en el ranking del bachillerato en términos de salario?

iii) La adición de la variable *phsrank* a (4.26) ¿modifica sustancialmente las conclusiones sobre el rendimiento correspondiente a dos y cuatro años de universidad? Explique.

iv) Este conjunto de datos contiene una variable llamada *id*. Explique por qué espera usted que si agrega *id* a la ecuación (4.17) o (4.26), esta variable será estadísticamente no significativa. ¿Cuál es el valor-*p* de dos colas?

C4.8 La base de datos 401KSUBS.RAW contiene información acerca de riqueza financiera neta (*nettfa*), edad de la persona entrevistada (*age*), ingreso familiar anual (*inc*), tamaño de la familia (*fsize*), y participación en ciertos planes de pensiones para estadounidenses. Las variables riqueza e ingreso están dadas en miles de dólares. Para esta ecuación, emplee sólo los datos de hogares de una sola persona (es decir $fsize = 1$).

i) ¿Cuántos hogares de una sola persona hay en esta base de datos?

ii) Use MCO para estimar el modelo

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u,$$

y dé los resultados empleando el formato habitual. Compruebe que sólo utiliza los hogares de una sola persona que hay en la muestra. Interprete los coeficientes de pendiente. ¿Hay algo que sorprenda en las estimaciones de pendiente?

iii) ¿Tiene algún significado interesante el intercepto en la regresión del inciso ii)? Explique.

iv) Encuentre el valor-*p* de la prueba H_0: $\beta_2 = 1$ contra H_0: $\beta_2 < 1$. ¿Rechaza usted H_0 al nivel de significancia de 1%?

v) Si realiza una regresión simple de *nettfa* sobre *inc*, ¿es el coeficiente estimado de *inc* muy diferente del estimado en el inciso ii)? Justifique su respuesta.

C4.9 Para responder esta pregunta emplee los datos del archivo DISCRIM.RAW. (Vea también el ejercicio para computadora C3.8 del capítulo 3.)

i) Use MCO para estimar el modelo

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u,$$

y dé los resultados en la forma habitual. ¿Es $\hat{\beta}_1$ estadísticamente diferente de cero al nivel de significancia de 5% contra una alternativa de dos colas? ¿Y al nivel de significancia de 1%?

ii) ¿Cuál es la correlación entre log(*income*) y *prppov*? En cualquier caso, ¿es cada variable estadísticamente significativa? Dé los valores-*p* de dos colas.

iii) En la regresión del inciso i) agregue la variable log(*hseval*) (log del valor promedio de la vivienda en cada zona postal). Interprete su coeficiente y dé el valor-*p* de dos colas para H_0: $\beta_{\log(hseval)} = 0$.

iv) En la regresión del inciso iii), ¿qué ocurre con la significancia estadística individual de log(*income*) y de *prppov*? ¿Son estas variables conjuntamente significativas? (Calcule un valor-*p*.) ¿Qué concluye de sus respuestas?

v) Dados los resultados de las regresiones anteriores, ¿cuál daría usted como la más confiable para determinar si la constitución racial de una zona postal influye en los precios locales de la comida rápida?

C4.10 Con ayuda de los datos del archivo ELEM94_95.RAW responda las preguntas siguientes. Los hallazgos pueden compararse con los de la tabla 4.1. La variable dependiente *lavgsal* es el logaritmo del sueldo promedio del profesor y *bs* es el cociente de beneficios promedio entre sueldo promedio (por escuela).

i) Realice una regresión simple de *lavgsal* sobre *bs*. ¿Es la pendiente estimada estadísticamente distinta de cero? ¿Es estadísticamente diferente de -1?

ii) Agregue las variables *lenrol* (log de la matrícula escolar) y *lstaff* (log de la cantidad de personal) a la regresión del inciso i). ¿Qué pasa con el coeficiente de *bs*? Compare esta situación con la de la tabla 4.1

iii) ¿A qué se debe que el error estándar del coeficiente de *bs* sea menor en el inciso ii) que en el inciso i)? (*Sugerencia*: ¿qué ocurre con la varianza del error frente a la multicolinealidad cuando se agregan *lenrol* y *lstaff*?)

iv) ¿A qué se debe que el coeficiente de *lstaff* sea negativo? ¿Es este coeficiente grande en magnitud?

v) Ahora, agregue a la regresión la variable *lunch* (porcentaje de alumnos que reciben desayuno gratuito). Manteniendo otros factores constantes, ¿están siendo compensados los profesores por enseñar a estudiantes con antecedentes desventajosos? Explique.

vi) En general, es el tipo de resultados que usted encuentra en ELEM94_95.RAW consistente con el tipo de resultados en la tabla 4.1?

CAPÍTULO 5

Análisis de regresión múltiple: MCO asintóticos

En los capítulos 3 y 4 se ha presentado lo que se conoce como propiedades de las *muestras finitas*, *muestras pequeñas* o propiedades *exactas* de los estimadores de MCO del modelo poblacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u. \quad (5.1)$$

Por ejemplo, la insesgadez o insesgamiento de MCO (obtenida en el capítulo 3) bajo los primeros cuatro supuestos de Gauss-Markov es una propiedad de las muestras finitas porque se satisface para *cualquier* tamaño de muestra n (sujeta a la leve restricción de que n debe ser al menos igual a la cantidad total de parámetros en el modelo de regresión, $k + 1$). De igual manera, el hecho de que MCO sea el mejor estimador lineal insesgado bajo el conjunto completo de supuestos de Gauss-Markov (RLM.1 a RLM.5) es una propiedad de las muestras finitas.

En el capítulo 4, al modelo lineal clásico se le agregó el supuesto RLM.6, que establece que el término del error u tiene una distribución normal y es independiente de las variables explicativas. Esto permitió obtener las distribuciones de muestreo *exactas* de los estimadores de MCO (condicionales en las variables explicativas de la muestra). En particular, el teorema 4.1 mostró que los estimadores de MCO tienen distribuciones de muestreo normales, lo que condujo directamente a las distribuciones t y F para los estadísticos t y F. Si el error no está distribuido normalmente, la distribución de un estadístico t no es con exactitud t y un estadístico F no tiene una distribución F exacta para ningún tamaño de muestra.

Además de las propiedades de muestras finitas, es importante conocer las **propiedades asintóticas** o **propiedades de muestras grandes** de los estimadores y de los estadísticos de prueba. Estas propiedades no están definidas para un tamaño de muestra determinado; se definen a medida que el tamaño de la muestra crece sin límite. Por fortuna, bajo los supuestos que se han hecho aquí, los MCO tienen propiedades satisfactorias para muestras grandes. Un hallazgo importante es que aun eliminando el supuesto de normalidad (supuesto RLM.6), los estadísticos t y F tienen distribuciones que son *aproximadamente* distribuciones t y F, al menos cuando las muestras son de tamaño grande. Esto se analiza con más detalle en la sección 5.2 después de ver la consistencia de MCO en la sección 5.1.

5.1 Consistencia

Aunque la insesgadez de los estimadores es importante, no siempre puede lograrse. Por ejemplo, como se vio en el capítulo 3, el error estándar de la regresión, $\hat{\sigma}$, no es un estimador insesgado de σ, la desviación estándar del error u en el modelo de regresión múltiple. Aunque bajo RLM.1

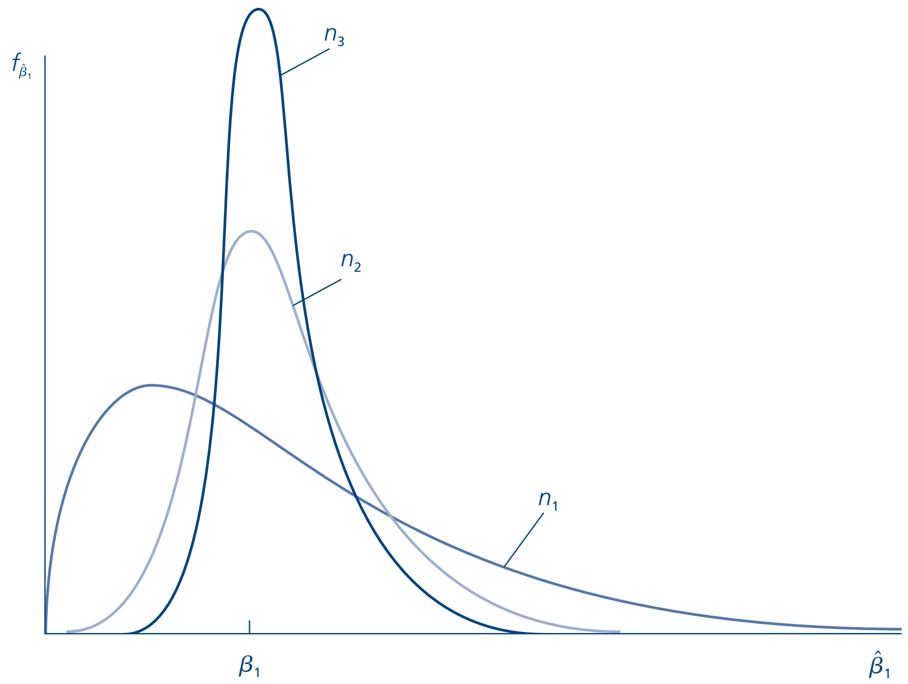
a RLM.4 los estimadores de MCO son insesgados, en el capítulo 11 se encontrará que existen regresiones de series de tiempo en las que no lo son. Además, en la parte 3 del libro, se encuentran otros estimadores que a pesar de ser sesgados son útiles.

Aunque no todos los estimadores útiles son insesgados, casi todos los economistas están de acuerdo en que la **consistencia** es un requisito mínimo para un estimador. El Premio Nobel, Clive W. J. Granger, especialista en econometría, observó, "Si usted no puede obtener un resultado correcto a medida que n tiende a infinito, es mejor que se dedique a otra cosa". Lo que quiere decir es que si el estimador de un determinado parámetro poblacional no es consistente, entonces se está perdiendo el tiempo.

Hay varias maneras de describir la consistencia. En el apéndice C se presentan definiciones y resultados formales; aquí, se empleará la comprensión intuitiva. En concreto, sea $\hat{\beta}_j$ el estimador de MCO de β_j para alguna j. Para cada n, $\hat{\beta}_j$ tiene una distribución de probabilidad (que representa los valores que puede tomar en distintas muestras aleatorias de tamaño n). Como $\hat{\beta}_j$ es insesgado bajo los supuestos RLM.1 a RLM.4, la media de esta distribución es β_j. Si este estimador es consistente, entonces a medida que el tamaño de la muestra aumente, la distribución de $\hat{\beta}_j$ se estrechará cada vez más en torno a β_j. A medida que n tiende a infinito, la distribución de $\hat{\beta}_j$ se colapsa hacia un solo punto β_j. En efecto, esto significa que si es posible recolectar tantos datos como se desee, entonces puede hacerse que el estimador esté arbitrariamente cerca de β_j. Esta convergencia se muestra en la figura 5.1.

FIGURA 5.1

Distribuciones de muestreo de $\hat{\beta}_1$ para muestras de tamaños $n_1 < n_2 < n_3$.



Naturalmente, en cualquier aplicación, se tiene un tamaño de muestra fijo, lo que hace que una propiedad asintótica, como la consistencia, sea difícil de comprender. La consistencia implica realizar un experimento para saber lo que ocurre a medida que el tamaño de la muestra aumenta (mientras, al mismo tiempo, se obtienen numerosas muestras aleatorias para cada tamaño de muestra). Si el obtener cada vez más datos no lleva a acercarse al valor del parámetro de interés, entonces el procedimiento de estimación que se está usando no es adecuado.

Por fortuna, el mismo conjunto de supuestos implica tanto insesgadez como consistencia de MCO. Esto se resume en un teorema.

Teorema 5.1 (Consistencia de MCO)

Bajo los supuestos RLM.1 a RLM.4, el estimador de MCO $\hat{\beta}_j$ es consistente para β_j, para toda $j = 0,1, ..., k$.

Una demostración general de este resultado se obtiene con más facilidad empleando los métodos del álgebra matricial descritos en los apéndices D y E. Sin embargo, el teorema 5.1 puede demostrarse sin dificultad para el caso del modelo de regresión simple. La atención se centrará en el estimador de pendiente, $\hat{\beta}_1$.

La demostración empieza igual que la de la insesgadez: se da la fórmula para $\hat{\beta}_1$, y después se sustituye en $y_i = \beta_0 + \beta_1 x_{i1} + u_i$:

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i\right) \Bigg/ \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right) \\ &= \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i\right) \Bigg/ \left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right).\end{aligned} \tag{5.2}$$

Al numerador y al denominador se les puede aplicar la ley de los grandes números, con lo que convergen en probabilidad a las cantidades poblacionales $\mathrm{Cov}(x_1,u)$ y $\mathrm{Var}(x_1)$, respectivamente. Siempre que $\mathrm{Var}(x_1) \neq 0$ —supuesto RLM.3— pueden emplearse las propiedades de los *límites de probabilidad* (véase el apéndice C) y obtener

$$\begin{aligned}\mathrm{plim}\ \hat{\beta}_1 &= \beta_1 + \mathrm{Cov}(x_1,u)/\mathrm{Var}(x_1) \\ &= \beta_1 \text{ debido a que } \mathrm{Cov}(x_1,u) = 0.\end{aligned} \tag{5.3}$$

Se ha empleado el hecho, visto en los capítulos 2 y 3, de que $\mathrm{E}(u|x_1) = 0$ (supuesto RLM.4) implica que x_1 y u no están correlacionados (tienen covarianza cero).

Como cuestión técnica, para asegurar que los límites de probabilidad existen, se debe suponer que $\mathrm{Var}(x_1) < \infty$ y $\mathrm{Var}(u) < \infty$ (lo que significa que sus distribuciones de probabilidad no están muy dispersas), pero no habrá mucha preocupación por los casos en los que no se satisfagan estos supuestos.

Los argumentos anteriores, y en particular la ecuación (5.3), muestran que basta que se suponga correlación cero para que MCO sea consistente en el caso de la regresión simple. Esto también es cierto en el caso general, lo cual se enuncia en el siguiente supuesto.

Supuesto RLM.4′ (Media cero y correlación cero)

$\mathrm{E}(u) = 0$ y $\mathrm{Cov}(x_j, u) = 0$, para $j = 1, 2, ..., k$.

El supuesto RLM.4′ es más débil que el RLM.4 en el sentido de que el último implica el primero. Una manera de caracterizar el supuesto de media condicional cero, $E(u|x_1, ..., x_k) = 0$, es que *alguna* función de las variables explicativas está correlacionada con u. El supuesto RLM.4′ sólo requiere que ninguna x_j esté correlacionada con u (y que u tenga media cero en la población). En el capítulo 2 en realidad los estimadores de MCO para la regresión simple se motivaron empleando el supuesto RLM.4′, y en el caso de la regresión múltiple las condiciones de primer orden para MCO, dadas en la ecuación (3.13), son simplemente los análogos muestrales de los supuestos de correlación cero poblacionales (y del supuesto de media cero). Por tanto, de alguna manera, el supuesto RLM.4′ es más natural porque lleva de modo directo a las estimaciones de MCO. Además, cuando se piensa en violaciones del supuesto RLM.4 en general se piensa en que $\text{Cov}(x_j, u) \neq 0$ para alguna j. ¿A qué se debe que el supuesto RLM.4 haya sido el empleado hasta ahora? Hay dos razones, las cuales ya han aparecido antes. La primera es que MCO resulta ser sesgado (pero consistente) bajo el supuesto RLM.4′ si $E(u|x_1, ..., x_k)$ depende de alguna de las x_j. Como antes la atención se centró en las propiedades de muestreo de los estimadores de MCO para muestras finitas, o exactas, se necesitó el supuesto más fuerte de media condicional cero.

La segunda, y tal vez la más importante, es que el supuesto de media condicional cero significa que se ha modelado de manera adecuada la función de regresión poblacional (FRP). Es decir, bajo el supuesto RLM.4 se puede escribir

$$E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

con lo que se pueden obtener los efectos parciales de las variables explicativas sobre el valor promedio o esperado de y. En cambio, si sólo se supone el supuesto RLM.4′, $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ no necesita representar la función de regresión poblacional y surge la posibilidad de que algunas funciones no lineales de x_j, como por ejemplo x_j^2, puedan estar correlacionadas con el error u. Una situación como ésta significa que en el modelo se han descuidado no linealidades que podrían ayudar a explicar mejor y; si se sabe esto, por lo general se incluirán funciones no lineales. En otras palabras, la mayoría de las veces se espera obtener una buena estimación de la FRP, con lo que el supuesto de media condicional cero es natural. De cualquier manera, el supuesto de correlación cero, que es más débil, resulta ser útil para interpretar la estimación MCO de un modelo lineal como la mejor aproximación lineal a la FRP. Este supuesto también se emplea en temas más avanzados como en el capítulo 15, donde no se tiene ningún interés en modelar una FRP. Para un análisis más amplio de este punto, algo sutil, vea Wooldridge (2002, capítulo 4).

Obtención de la inconsistencia en MCO

Así como el que no se satisfaga $E(u|x_1, ..., x_k) = 0$ ocasiona sesgo en los estimadores de MCO, la correlación entre u y *cualquiera* de las $x_1, x_2, ..., x_k$ causa, por lo general, que *todos* los estimadores de MCO sean inconsistentes. Esta sencilla pero importante observación suele resumirse de la manera siguiente: *si el error está correlacionado con cualquiera de las variables independientes, entonces MCO es sesgado e inconsistente*. Esto es muy lamentable porque significa que cualquier sesgo persistirá a medida que el tamaño de la muestra aumente.

En el caso de la regresión simple, la inconsistencia puede obtenerse de la primera parte de la ecuación (5.3), que se satisface, ya sea que u y x_1 estén o no no correlacionadas. La **inconsistencia** en $\hat{\beta}_1$ (algunas veces llamada **sesgo asintótico**) es

$$\text{plim}\,\hat{\beta}_1 - \beta_1 = \text{Cov}(x_1, u)/\text{Var}(x_1). \tag{5.4}$$

Como $\text{Var}(x_1) > 0$, la inconsistencia en $\hat{\beta}_1$ es positiva si x_1 y u están correlacionadas positivamente y es negativa si x_1 y u están correlacionadas negativamente. Si la covarianza entre x_1 y u es pequeña con relación a la varianza de x_1, la inconsistencia puede ser insignificante; por desgracia, no se puede estimar qué tan grande es la covarianza debido a que no se conoce u.

La ecuación (5.4) puede emplearse para obtener un análogo asintótico del sesgo de la variable omitida (vea la tabla 3.2 en el capítulo 3). Suponga que el modelo verdadero,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

satisface los primeros cuatro supuestos de Gauss-Markov. Entonces v tiene media cero y no está correlacionada con x_1 ni con x_2. Si $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ denotan los estimadores de MCO en la regresión de y sobre x_1 y x_2 entonces el teorema 5.1 implica que estos estimadores son consistentes. Si se omite x_2 de la regresión y se realiza la regresión simple de y sobre x_1, entonces $u = \beta_2 x_2 + v$. Sea $\tilde{\beta}_1$ el estimador de pendiente en la regresión simple. Entonces

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \beta_2 \delta_1, \tag{5.5}$$

donde

$$\delta_1 = \text{Cov}(x_1, x_2)/\text{Var}(x_1). \tag{5.6}$$

Por tanto, para fines prácticos, puede considerarse que la inconsistencia es lo mismo que el sesgo. La diferencia es que la inconsistencia se expresa en términos de la varianza poblacional de x_1 y de la covarianza poblacional entre x_1 y x_2, mientras que el sesgo se basa en sus contrapartes muestrales (porque se condiciona sobre los valores de x_1 y x_2 en la muestra).

Si x_1 y x_2 no están correlacionadas (en la población), entonces $\delta_1 = 0$ y $\tilde{\beta}_1$ es un estimador consistente de β_1 (aunque no necesariamente insesgado). Si x_2 tiene un efecto parcial positivo sobre y, de manera que $\beta_2 > 0$, y x_1 y x_2 están correlacionadas positivamente, de manera que $\delta_1 > 0$, entonces la inconsistencia en $\tilde{\beta}_1$ es positiva. Y así sucesivamente la dirección de la inconsistencia o sesgo asintótico puede obtenerse de la tabla 3.2. Si la covarianza entre x_1 y x_2 es pequeña con relación a la varianza de x_1, la inconsistencia puede ser pequeña.

Ejemplo 5.1

[Precios de viviendas y distancia a un incinerador]

Sea y el precio de una casa (*price*), sea x_1 la distancia de la casa a un nuevo incinerador de basura (*distance*), y sea x_2 la "calidad" de la casa (*quality*). La variable *quality* se deja vaga de manera que pueda comprender cosas como tamaño de la casa y del terreno, cantidad de recámaras y de baños, así como aspectos intangibles como lo atractivo de la zona. Si el incinerador hace que se deprecien las casas, entonces β_1 será positivo: permaneciendo todo lo demás igual, una casa que esté más lejos del incinerador valdrá más. Por definición, β_2 es positivo, ya que las casas de mayor calidad tienen un precio más alto, permaneciendo todo lo demás igual. Si el incinerador se construyó más alejado de las mejores casas, en promedio, entonces *distance* y *quality* estarán correlacionadas positivamente y de esta manera $\delta_1 > 0$. Una regresión simple de *price* sobre *distance* [o log(*price*) sobre log(*distance*)] tenderá a sobrestimar el efecto del incinerador: $\beta_1 + \beta_2\delta_1 > \beta_1$.

Pregunta 5.1

Suponga que el modelo

$$score = \beta_0 + \beta_1 skipped + \beta_2 priGPA + u$$

en el que *score* es la puntuación en un examen final, *skipped* es la cantidad de clases perdidas y *priGPA* es el promedio general hasta antes del semestre actual, satisface los cuatro primeros postulados de Gauss-Markov. Si $\tilde{\beta}_1$ proviene de la regresión simple de *score* sobre *skipped*, ¿cuál es la dirección del sesgo asintótico de $\tilde{\beta}_1$?

Un punto importante acerca de la inconsistencia de los estimadores de MCO es que, por definición, el problema no se resuelve adicionando observaciones a la muestra. Cuando mucho, con más datos el problema empeora: a medida que el tamaño de la muestra aumenta, los estimadores de MCO se acercan cada vez más a $\beta_1 + \beta_2\delta_1$.

En el caso general de k regresores, obtener el signo y la magnitud de la inconsistencia es más difícil, como lo es obtener el sesgo. Es necesario recordar que si en el modelo de la ecuación (5.1), por ejemplo, x_1 está correlacionada con u, pero las otras variables independientes no están correlacionadas con u, *todos* los estimadores de MCO son por lo general inconsistentes. Por ejemplo, en el caso $k = 2$,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

suponga que x_2 no está correlacionada con u pero x_1 sí. Entonces los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ de MCO serán, en general, inconsistentes ambos. (El intercepto también será inconsistente.) La inconsistencia de $\hat{\beta}_2$ surge cuando x_1 y x_2 están correlacionadas, como suele ser el caso. Si x_1 y x_2 no están correlacionadas, entonces cualquier correlación entre x_1 y u *no* da como resultado la inconsistencia de $\hat{\beta}_2$: plim $\hat{\beta}_2 = \beta_2$. Además, la inconsistencia de $\hat{\beta}_1$ es la misma que en (5.4). Lo mismo es válido para el caso general: si x_1 está correlacionada con u, pero x_1 y u no están correlacionadas con las demás variables independientes, entonces sólo $\hat{\beta}_1$ será inconsistente y la inconsistencia estará dada por (5.4). El caso general es muy parecido al de la variable omitida visto en la sección 3A.4 del apéndice 3A.

5.2 Normalidad asintótica e inferencia con muestras grandes

La consistencia de un estimador es una propiedad importante, pero sola no permite realizar inferencias estadísticas. Saber que a medida que el tamaño de la muestra aumenta un estimador se acerca cada vez más al valor poblacional no permite probar hipótesis acerca de los parámetros. Para probar hipótesis, se necesita la distribución de muestreo del estimador de MCO. El teorema 4.1 muestra que bajo los supuestos RLM.1 a RLM.6 del modelo lineal clásico, estas distribuciones de muestreo son normales. Este resultado es la base para obtener las distribuciones t y F que con frecuencia se emplean en la econometría aplicada.

La normalidad exacta de los estimadores de MCO depende de manera crucial de la normalidad, en la población, de la distribución del error u. Si los errores $u_1, u_2, ..., u_n$ son tomados de manera aleatoria de una distribución distinta a la normal, los $\hat{\beta}_j$ no estarán distribuidos normalmente, lo que significa que los estadísticos t no tendrán distribuciones t y que los estadísticos F no tendrán distribuciones F. Este es un problema que podría ser serio porque las inferencias dependen de que puedan obtenerse los valores críticos o los valores-p a partir de las distribuciones t y F respectivamente.

Recuerde que el supuesto RLM.6 es equivalente a decir que la distribución de y dadas $x_1, x_2, ..., x_k$ es normal. Como y sí se puede observar y u no, en una aplicación determinada, es mucho más fácil pensar en si la distribución de y es normal. En efecto, se han visto ya algunos ejemplos en los que y en definitiva no puede tener una distribución condicional normal. Una variable aleatoria distribuida normalmente tiene una distribución simétrica respecto a su media, puede tomar cualquier valor positivo o negativo (pero con probabilidad cero) y más de 95% del área bajo la distribución se encuentra a no más de dos desviaciones estándar de la media.

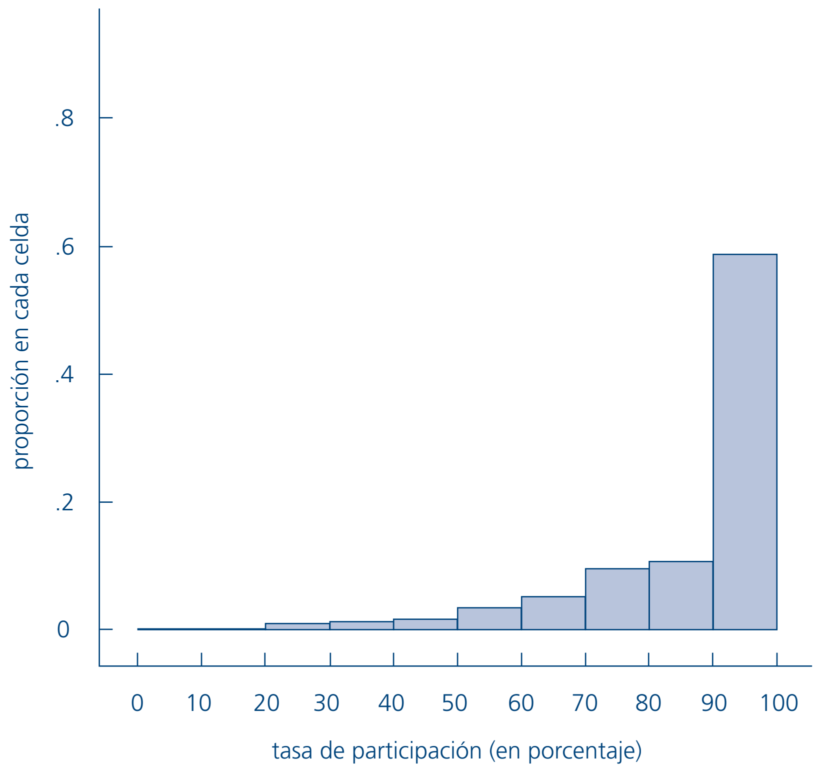
En el ejemplo 3.5 se estimó un modelo que explicaba la cantidad de detenciones de cada hombre joven en un determinado año (*narr86*). En la población, la mayoría de los hombres no son arrestados durante un año y la inmensa mayoría es arrestada cuando mucho una vez. (En la muestra de los 2,725 hombres de la base de datos CRIME1.RAW, menos de 8% fue arrestado más de una vez durante 1986.) Como en 92% de la muestra la variable *narr86* toma sólo dos valores, no se puede decir que esta variable esté distribuida normalmente en la población.

En el ejemplo 4.6, se estimó un modelo que explicaba la participación porcentual (*prate*) en los planes de pensiones 401(k). En la figura 5.2, la distribución de frecuencia (conocida también como *histograma*) muestra que la distribución de *prate* es muy asimétrica hacia la derecha y no distribuida normalmente. En efecto, más de 40% de las observaciones de *prate* corresponden al valor 100, lo cual indica 100% de participación. Esto viola el supuesto de normalidad, incluso condicionada a las variables explicativas.

Se sabe que la normalidad no juega ningún papel en la insesgadez de MCO y tampoco afecta las conclusiones de que MCO es el mejor estimador lineal insesgado bajo los supuestos de Gauss-Markov. Pero la inferencia exacta basada en los estadísticos t y F requiere RLM.6. Significa esto

FIGURA 5.2

Histograma de *prate* obtenido de los datos del archivo 401k.RAW.



que, ¿en el análisis de *prate* que se realizará en el ejemplo 4.6, debe abandonarse el estadístico t para determinar qué variables son estadísticamente significativas? Por fortuna, la respuesta a esta pregunta es *no*. Aunque las y_i no provienen de una distribución normal, puede emplearse el teorema del límite central del apéndice C para concluir que los estimadores de MCO satisfacen la **normalidad asintótica**, lo cual significa que están distribuidos de manera aproximadamente normal cuando se tienen muestras de tamaño suficientemente grande.

Teorema 5.2 (Normalidad asintótica de MCO)

Bajo los supuestos RLM.1 a RLM.5 de Gauss-Markov,

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \text{Normal}(0,\sigma^2/a_j^2)$, donde $\sigma^2/a_j^2 > 0$ es la **varianza asintótica** de $\sqrt{n}\,(\hat{\beta}_j - \beta_j)$; los coeficientes de pendiente, $a_j^2 = \text{plim}\left(n^{-1}\sum_{i=1}^{n}\hat{r}_{ij}^2\right)$, donde $\hat{r}_{ij}$ son los residuales de regresar x_j sobre las otras variables independientes. Se dice que $\hat{\beta}_j$ está *distribuida en forma asintóticamente normal* (vea el apéndice C);

(ii) $\hat{\sigma}^2$ es un estimador consistente de $\sigma^2 = \text{Var}(u)$;

(iii) Para cada j,

$$(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1), \tag{5.7}$$

donde $\text{ee}(\hat{\beta}_j)$ es el error estándar usual de MCO.

La demostración de la normalidad asintótica es un poco complicada y se esboza en el apéndice para el caso de la regresión simple. El inciso ii) es consecuencia de la ley de los grandes números y el iii) de los incisos i) y ii) y de las propiedades asintóticas que se analizan en el apéndice C.

El teorema 5.2 es útil porque se abandona el supuesto de normalidad RLM.6; la única restricción a la distribución del error es que tenga varianza finita, algo que se supondrá siempre. También se ha supuesto la media condicional cero (RLM.4) y la homocedasticidad de u (RLM.5).

Observe que en (5.7) aparece la distribución normal estándar, y no la distribución t_{n-k-1}. Esto se debe a que la distribución es sólo aproximada. En cambio, en el teorema 4.2 la distribución del cociente en (5.7) fue *exactamente* t_{n-k-1} para cualquier tamaño de muestra. Desde una perspectiva práctica, esta diferencia es irrelevante. En realidad, es igual de legítimo escribir

$$(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j) \overset{a}{\sim} t_{n-k-1}, \tag{5.8}$$

ya que a medida que aumentan los grados de libertad t_{n-k-1} se aproxima a la distribución normal estandar.

La ecuación (5.8) indica que la prueba t y la construcción de intervalos de confianza se realizan *exactamente* como bajo los supuestos del modelo lineal clásico. Esto significa que el análisis de variables dependientes como *prate* y *narr86* no tiene que cambiar en absoluto si se satisfacen los supuestos de Gauss-Markov: en ambos casos se tienen al menos 1,500 observaciones, lo que sin duda es suficiente para justificar la aproximación del teorema del límite central.

Si el tamaño de la muestra no es muy grande, entonces la distribución t puede ser una aproximación inadecuada para la distribución del estadístico t cuando u no está distribuida de manera normal. Por desgracia, no hay una indicación de qué tan grande debe ser el tamaño de la muestra para que la aproximación sea suficientemente buena. Algunos econometristas piensan que $n = 30$ es satisfactorio, pero esto puede no ser suficiente para todas las distribuciones de u.

Dependiendo de la distribución de u puede que para que el teorema del límite central proporcione una aproximación útil sean necesarias más observaciones. Además, la calidad de la aproximación no sólo depende de n, sino también de los gl, $n - k - 1$: cuando en el modelo hay más variables independientes, para emplear la aproximación t en general se necesita un tamaño de muestra mayor. Los métodos de inferencia para grados de libertad pequeños y errores no normales quedan fuera del alcance de este libro. Aquí sólo se usará el estadístico t como se ha usado siempre sin preocuparse por el supuesto de normalidad.

Es muy importante ver que el teorema 5.2 *sí* requiere el supuesto de homocedasticidad (junto con el de media condicional cero). Si $\mathrm{Var}(y|\mathbf{x})$ no es constante, el estadístico t usual y los intervalos de confianza no son válidos, sin importar qué tan grande sea el tamaño de la muestra; el teorema del límite central no nos saca de apuros cuando hay heterocedasticidad. A esto se debe que el capítulo 8 se dedique por completo a analizar lo que se puede hacer en presencia de heterocedasticidad.

Una conclusión del teorema 5.2 es que $\hat{\sigma}^2$ es un estimador consistente de σ^2; sabíamos del Teorema 3.3 que $\hat{\sigma}^2$ es insesgada para σ^2 bajo los supuestos de Gauss-Markov. La consistencia implica que $\hat{\sigma}$ es un estimador consistente de σ lo cual es importante para establecer el resultado de normalidad asintótica de la ecuación (5.7).

Recuerde que $\hat{\sigma}$ aparece en el error estándar de cada $\hat{\beta}_j$. En efecto, la varianza estimada de $\hat{\beta}_j$ es

$$\widehat{\mathrm{Var}(\hat{\beta}_j)} = \frac{\hat{\sigma}^2}{\mathrm{STC}_j(1 - R_j^2)}, \tag{5.9}$$

Pregunta 5.2

En un modelo de regresión con un tamaño de muestra grande, ¿cuál es un intervalo de confianza aproximado de 95% para $\hat{\beta}_j$ bajo RLM.1 a RLM.5? A esto se le conoce como **intervalo de confianza asintótico**.

donde STC_j es la suma total de cuadrados de x_j en la muestra, y R_j^2 es la R-cuadrada de regresar x_j sobre todas las demás variables independientes. En la sección 3.4, se estudiaron cada uno de los componentes de (5.9), los cuales se exponen ahora en el contexto del análisis asintótico. A medida que aumenta el tamaño de la muestra, $\hat{\sigma}^2$ converge en probabilidad a la constante σ^2. Además, R_j^2 se aproxima a un número estrictamente entre cero y la unidad (de manera que $1 - R_j^2$ converge a algún número entre cero y uno). La varianza de muestreo de x_j es STC_j/n, y de esta manera, a medida que el tamaño de la muestra aumenta, STC_j/n converge a $\mathrm{Var}(x_j)$. Esto significa que STC_j aumenta aproximadamente a la misma velocidad que el tamaño de la muestra: $\mathrm{STC}_j \approx n\sigma_j^2$, donde σ_j^2 es la varianza poblacional de x_j. Combinando estos hechos se encuentra que $\widehat{\mathrm{Var}(\hat{\beta}_j)}$ se reduce a cero a la velocidad de $1/n$; a esto se debe que las muestras de tamaño grande sean mejores.

Cuando u no está distribuida normalmente, a la raíz cuadrada de (5.9) se le suele llamar **error estándar asintótico** y al estadístico t **estadístico t asintótico**. Como estas cantidades son las mismas que se encontraron en el capítulo 4, aquí se les llamará errores estándar y estadísticos t, entendiéndose que algunas veces sólo están justificados con muestras grandes. Un comentario similar vale también para los **intervalos de confianza asintóticos** construidos a partir de errores estándar asintóticos.

Empleando los argumentos anteriores acerca de la varianza estimada, se puede escribir

$$\mathrm{ee}(\hat{\beta}_j) \approx c_j/\sqrt{n}, \tag{5.10}$$

donde c_j es una constante positiva que *no* depende del tamaño de la muestra. La ecuación (5.10) es sólo una aproximación, pero es una regla práctica útil: puede esperarse que los errores estándar disminuyan a una velocidad inversamente proporcional a la *raíz cuadrada* del tamaño de la muestra.

Ejemplo 5.2

[Errores estándar en una ecuación para el peso al nacer]

Los datos del archivo BWGHT.RAW se emplean para estimar una relación en la que el logaritmo del peso al nacer es la variable dependiente, y la cantidad de cigarros fumados por día (*cigs*) así como el logaritmo del ingreso familiar son las variables independientes. La cantidad total de observaciones es 1,388. Empleando la primera mitad de las observaciones (694), el error estándar de $\hat{\beta}_{cigs}$ es aproximadamente .0013. Empleando todas las observaciones, el error estándar es más o menos .00086. El cociente del último error estándar entre el primero es .00086/.0013 $\approx$ .662. Esta cantidad es bastante cercana a $\sqrt{694/1,388} \approx .707$, el cociente que se obtiene con la aproximación dada en (5.10). En otras palabras, la ecuación (5.10) implica que el error estándar que se obtiene empleando el tamaño de muestra mayor debe ser aproximadamente 70.7% del error estándar que se obtiene empleando la muestra menor. Este porcentaje está bastante cercano al 66.2% que se calculó empleando el cociente de los errores estándar.

La normalidad asintótica de los estimadores de MCO implica también que, en las muestras grandes, los estadísticos *F* tienen distribuciones aproximadamente *F*. Por tanto, para probar restricciones de exclusión u otras hipótesis múltiples, no cambia nada de lo hecho antes.

Otras pruebas con muestras grandes: el estadístico del multiplicador de Lagrange

Una vez en el campo del análisis asintótico, pueden emplearse otros estadísticos para probar hipótesis. En la mayoría de los casos no hay razón para ir más allá de los estadísticos usuales *t* y *F*: como se acaba de ver, en muestras grandes estos estadísticos tienen justificación sin el supuesto de normalidad. De cualquier modo, algunas veces es útil tener otras maneras de probar restricciones de exclusión múltiple; ahora se verá el **estadístico del multiplicador de Lagrange** (***ML***), el cual ha obtenido cierta popularidad en la econometría moderna.

El nombre "estadístico del multiplicador de Lagrange" proviene de la optimización restringida, un tema que queda fuera del alcance de este libro [Vea Davidson y MacKinnon (1993)]. También se emplea el nombre **estadístico de puntuación**, el cual proviene de la optimización empleando el cálculo. Por fortuna, en el marco de la regresión lineal, es sencillo motivar el estadístico *ML* sin tener que profundizar en matemática compleja.

La forma del estadístico *ML* que se obtiene aquí se apoya en los supuestos de Gauss-Markov, los mismos supuestos que justifican el estadístico *F* en muestras grandes. El supuesto de normalidad no es necesario.

Para obtener el estadístico *ML* considere el modelo de regresión múltiple usual con *k* variables independientes:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u. \qquad \mathbf{5.11}$$

Se desea probar, por ejemplo, si las *q* últimas variables tienen parámetros poblacionales igual a cero: la hipótesis nula es

$$\mathrm{H}_0\colon \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \qquad \mathbf{5.12}$$

que impone *q* restricciones de exclusión al modelo (5.11). Como en las pruebas *F*, la alternativa a (5.12) es que por lo menos uno de los parámetros es distinto de cero.

El estadístico *ML* sólo requiere la estimación del modelo *restringido*. Así, supóngase que se ha corrido la regresión

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}, \qquad \textbf{5.13}$$

donde "~" significa que las estimaciones corresponden al modelo restringido. En particular, $\tilde{u}$ indica los residuales del modelo restringido. (Como de costumbre, esto es sólo una manera abreviada de indicar que se obtiene el residual restringido de cada observación en la muestra.)

Si las variables omitidas x_{k-q+1} a x_k en realidad tienen coeficientes poblacionales iguales a cero, entonces, por lo menos de manera aproximada, $\tilde{u}$ no deberá estar correlacionada con ninguna de estas variables de la muestra. Esto sugiere correr una regresión de estos residuales sobre las variables independientes que se excluyen bajo H_0, que es, más o menos, lo que hace la prueba *ML*. Sin embargo, resulta que para obtener un estadístico de prueba útil, deben incluirse en la regresión *todas* las variables independientes. (Se deben incluir todos los regresores porque, en general, los regresores omitidos en el modelo restringido están correlacionados con el regresor que aparece en el modelo restringido.) Por tanto, se corre la regresión

$$\tilde{u} \text{ sobre } x_1, x_2, \ldots, x_k. \qquad \textbf{5.14}$$

Este es un ejemplo de una **regresión auxiliar**, que se emplea para calcular un estadístico de prueba, pero cuyos coeficientes no son de interés directo.

¿Cómo se emplea el resultado de la regresión de (5.14) para probar (5.12)? Si (5.12) es verdadera, la *R*-cuadrada de (5.14) debe tener un valor "cercano" a cero, sujeto al error de muestreo, porque $\tilde{u}$ no estará correlacionada aproximadamente con ninguna de las variables independientes. Como siempre en las pruebas de hipótesis, la cuestión es cómo determinar cuándo es suficientemente grande el estadístico para rechazar la hipótesis nula al nivel de significancia elegido. Resulta que, bajo la hipótesis nula, el tamaño de la muestra multiplicado por la *R*-cuadrada usual de la regresión auxiliar (5.14) tiene una distribución asintótica como variable aleatoria ji-cuadrada con q grados de libertad. Esto conduce a un procedimiento sencillo para probar la significancia de un conjunto de q variables independientes.

Estadístico del multiplicador de Lagrange para *q* restricciones de exclusión:

i) Regresar y sobre el conjunto *restringido* de variables independientes y guardar los residuales, $\tilde{u}$.

ii) Regresar $\tilde{u}$ sobre *todas* las variables independientes y obtener la *R*-cuadrada, por ejemplo, R_u^2 (para distinguirla de la *R*-cuadrada obtenida con y como variable dependiente).

iii) Calcular $ML = nR_u^2$ [el tamaño de la muestra multiplicado por la *R*-cuadrada obtenida en el paso (ii)].

iv) Comparar *ML* con el valor crítico adecuado, c, de una distribución χ_q^2; si $ML > c$ se rechaza la hipótesis nula. Es todavía mejor obtener el valor-*p* como la probabilidad de que una variable aleatoria χ_q^2 sea mayor que el valor del estadístico de prueba. Si el valor-*p* es menor que el nivel de significancia elegido, entonces se rechaza H_0 Si no es así, no se puede rechazar H_0. La regla de rechazo es en esencia la misma que en la prueba *F*.

Debido a su forma, al estadístico *ML* también se le conoce como estadístico ***n-R*-cuadrada**. A diferencia del estadístico *F*, los grados de libertad en el modelo no restringido no juegan ningún papel al llevar a cabo la prueba *ML*. Lo único que importa es la cantidad de restricciones que se prueba (q), la magnitud de la *R*-cuadrada auxiliar (R_u^2) y el tamaño de la muestra (n). El número de *gl* en el modelo no restringido no juega ningún papel debido a la naturaleza asintótica del estadístico *ML*. Pero hay que tener cuidado de multiplicar R_u^2 por el tamaño de la muestra para obtener *ML*; un valor aparentemente bajo de la *R*-cuadrada puede conducir a una significancia conjunta si n es grande.

Antes de dar un ejemplo es necesario hacer una advertencia. Si en el paso i), se regresa de manera equivocada y sobre todas las variables independientes y se obtienen los residuales de esta regresión no restringida para emplearlos después en el paso ii), no se obtiene un estadístico interesante: la *R*-cuadrada que se obtiene ¡será exactamente cero!

Esto se debe a que MCO elige las estimaciones de manera que los residuales no estén correlacionados en la muestra con todas las variables independientes incluidas [vea las ecuaciones en (3.13)]. Por tanto, sólo se puede probar (5.12) regresando los residuales restringidos sobre *todas* las variables independientes. (Regresar los residuales restringidos sobre el conjunto restringido de variables independientes producirá también $R^2 = 0$.)

Ejemplo 5.3

[Modelo económico para la delincuencia]

Se ejemplifica la prueba *ML* empleando una pequeña extensión del modelo para la delincuencia visto en el ejemplo 3.5:

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

donde

narr86 = número de veces que un hombre ha sido arrestado.
pcnv = proporción de detenciones anteriores que condujeron a una condena.
avgsen = duración promedio de las condenas anteriores cumplidas.
tottime = total del tiempo que ha pasado el hombre en prisión antes de 1986 y desde los 18 años de edad.
ptime86 = tiempo en meses pasado en prisión en 1986.
qemp86 = cantidad de trimestres que la persona tuvo empleo en 1986.

Se emplea el estadístico *ML* para probar la hipótesis nula de que *avgsen* y *tottime* no tienen efectos sobre *narr86* una vez controlados los otros factores.

En el paso i), se estima el modelo restringido regresando *narr86* sobre *pcnv*, *ptime86* y *qemp86*; las variables *avgsen* y *tottime* se excluyen de esta regresión. De esta regresión se obtienen los residuales $\tilde{u}$ de 2,725 a partir de ellos. A continuación se corre la regresión de

$$\tilde{u} \text{ sobre } pcnv,\ ptime86,\ qemp86,\ avgsen, \text{ y } tottime; \tag{5.15}$$

como siempre, el orden en que aparezcan las variables independientes es irrelevante. Con esta segunda regresión se obtiene R_u^2, que resulta ser aproximadamente .0015. Este valor puede parecer pequeño, pero para obtener el estadístico *ML* debe ser multiplicado por n: $ML = 2{,}725(.0015) \approx 4.09$. El valor crítico correspondiente a 10% en una distribución ji-cuadrada con dos grados de libertad es aproximadamente 4.61 (redondeado a dos cifras decimales; vea la tabla G.4). Por tanto, al nivel de significancia de 10% no se puede rechazar la hipótesis nula de que $\beta_{avgsen} = 0$ y $\beta_{tottime} = 0$. El valor-*p* es $P(\chi_2^2 > 4.09) \approx .129$, de manera que H_0 se rechazará al nivel de significancia de 15%.

Como comparación, la prueba *F* para la significancia conjunta de *avgsen* y *tottime* da un valor-*p* de aproximadamente .131, que está bastante cerca al obtenido empleando el estadístico *ML*. Esto no debe extrañar, ya que, asintóticamente, los dos estadísticos tienen la misma probabilidad del error tipo I. (Es decir, los dos rechazan con la misma frecuencia la hipótesis nula cuando ésta es verdadera.)

Como indica el ejemplo anterior, cuando se tiene una muestra grande, es raro que existan discrepancias importantes entre los resultados de las pruebas *ML* y *F*. Aquí se usará sobre todo el estadístico *F* ya que este es el calculado de manera rutinaria por la mayoría de los paquetes para regresión. Pero debe conocerse el estadístico *ML* pues es empleado en trabajos de aplicación.

Un último comentario sobre el estadístico *ML*. Como ocurre con el estadístico *F*, se debe tener cuidado de emplear las mismas observaciones en los pasos i) y ii). Si faltan datos de alguna de las variables independientes que se excluyen bajo la hipótesis nula, los residuales del paso i) deben obtenerse de una regresión sobre el conjunto de datos reducido.

5.3 Eficiencia asintótica de MCO

Se sabe que, bajo los supuestos de Gauss-Markov, los estimadores de MCO son los mejores estimadores lineales insesgados. MCO es también **asintóticamente eficiente** dentro de una determinada clase de estimadores bajo los supuestos de Gauss-Markov. Un tratamiento general requiere álgebra matricial y análisis asintótico avanzado. Primero se describe el resultado en el caso de la regresión simple.

En el modelo

$$y = \beta_0 + \beta_1 x + u, \tag{5.16}$$

u tiene media condicional cero bajo RLM.4: $\mathrm{E}(u|x) = 0$. Esto abre una variedad de estimadores consistentes para β_0 y β_1; como de costumbre, la atención se enfoca al parámetro de pendiente, β_1. Sea $g(x)$ una función de x; por ejemplo, $g(x) = x^2$ o $g(x) = 1/(1 + |x|)$. Entonces u no está correlacionado con $g(x)$ (vea la propiedad CE.5 en el apéndice B). Sea $z_i = g(x_i)$ para todas las observaciones i. Entonces el estimador

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n}(z_i - \bar{z})y_i\right) \Bigg/ \left(\sum_{i=1}^{n}(z_i - \bar{z})x_i\right) \tag{5.17}$$

es consistente para β_1, siempre que $g(x)$ y x estén correlacionadas. [Recuerde que es posible que $g(x)$ y x no estén correlacionadas debido a que la correlación mide dependencia *lineal*.] Para ver esto, se puede sustituir $y_i = \beta_0 + \beta_1 x_i + u_i$ y escribir $\tilde{\beta}_1$ como

$$\tilde{\beta}_1 = \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})u_i\right) \Bigg/ \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i\right). \tag{5.18}$$

Ahora puede aplicarse la ley de los grandes números al numerador y al denominador, los cuales convergen en probabilidad a $\mathrm{Cov}(z,u)$ y a $\mathrm{Cov}(z,x)$, respectivamente. Siempre que $\mathrm{Cov}(z,x) \neq 0$ —de manera que z y x estén correlacionadas— se tiene

$$\mathrm{plim}\ \tilde{\beta}_1 = \beta_1 + \mathrm{Cov}(z,u)/\mathrm{Cov}(z,x) = \beta_1,$$

debido a que $\mathrm{Cov}(z,u) = 0$ bajo RLM.4.

Mostrar que $\tilde{\beta}_1$ es asintóticamente normal es más difícil. Pero, empleando argumentos similares a los que se presentan en el apéndice, se puede demostrar que $\sqrt{n}\,(\tilde{\beta}_1 - \beta_1)$ es asintóticamente normal con media cero y varianza asintótica $\sigma^2\mathrm{Var}(z)/[\mathrm{Cov}(z,x)]^2$. La varianza asintótica de los estimadores de MCO se obtiene cuando $z = x$, en cuyo caso, $\mathrm{Cov}(z,x) = \mathrm{Cov}(x,x) = \mathrm{Var}(x)$. Por tanto la varianza asintótica de $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$, donde $\hat{\beta}_1$ es el estimador de MCO, es $\sigma^2\mathrm{Var}(x)/[\mathrm{Var}(x)]^2 = \sigma^2/\mathrm{Var}(x)$. Ahora, la desigualdad de Cauchy-Schwartz (vea el apéndice B.4) implica que $[\mathrm{Cov}(z,x)]^2 \leq \mathrm{Var}(z)\mathrm{Var}(x)$, lo cual implica que la varianza asintótica de $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$ no es mayor a la de $\sqrt{n}\,(\tilde{\beta}_1 - \beta_1)$. Para el caso de la regresión simple se ha demostrado que, bajo los supuestos de Gauss-Markov, los estimadores de MCO tienen una varianza asintótica menor que cualquier estimador de la forma (5.17). [El estimador en (5.17) es un ejemplo de un *estimador de variables instrumentales*, los cuales se estudiarán ampliamente en el capítulo 15.] Si el supuesto de homocedasticidad no se satisface, entonces existen estimadores de la forma (5.17) que tienen una varianza asintótica menor que los de MCO. Esto se verá en el capítulo 8.

El caso general es similar pero matemáticamente mucho más difícil. En el caso de k regresores, la clase de estimadores consistentes se obtiene generalizando las condiciones de primer orden de MCO:

$$\sum_{i=1}^{n} g_j(\boldsymbol{x}_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik}) = 0,\ j = 0, 1, \ldots, k, \qquad \textbf{5.19}$$

donde $g_j(\boldsymbol{x}_i)$ denota cualquier función de todas las variables explicativas para la observación i. Como se puede ver, comparando (5.19) con las condiciones de primer orden de MCO en (3.13), los estimadores de MCO se obtienen cuando $g_0(\boldsymbol{x}_i) = 1$ y $g_j(\boldsymbol{x}_i) = x_{ij}$ para $j = 1, 2, \ldots, k$. La clase de estimadores en (5.19) es infinita porque se puede emplear cualquier función de las x_{ij} que se desee.

Teorema 5.3 (Eficiencia asintótica de MCO)

Bajo los supuestos de Gauss-Markov, sean $\tilde{\beta}_j$ los estimadores que resuelven las ecuaciones de la forma (5.19) y sean $\hat{\beta}_j$ los estimadores de MCO. Entonces para $j = 0, 1, 2, \ldots, k$, los estimadores de MCO tienen las menores varianzas asintóticas: $\mathrm{Avar}\,\sqrt{n}\,(\hat{\beta}_j - \beta_j) \leq \mathrm{Avar}\,\sqrt{n}\,(\tilde{\beta}_j - \beta_j)$.

Probar la consistencia de los estimadores en (5.19) o demostrar que son asintóticamente normales, es matemáticamente complicado. [Vea Wooldridge (2002, capítulo 5).]

RESUMEN

Las afirmaciones subyacentes al material de este capítulo son bastante técnicas, pero sus consecuencias prácticas son inmediatas. Se ha mostrado que los primeros cuatro supuestos de Gauss-Markov implican que MCO es consistente. Además, todos los métodos de prueba y construcción de intervalos de confianza estudiados en el capítulo 4 son aproximadamente válidos sin el supuesto de que los errores provengan de una distribución normal (o lo que es equivalente, que la distribución de y dadas las variables explicativas no sea normal). Esto significa que se puede aplicar MCO y emplear los métodos anteriores en una serie de aplicaciones en las que la variable dependiente no esté distribuida en forma ni incluso aproximadamente normal. Se mostró también que para probar restricciones de exclusión, el estadístico *ML* puede emplearse en lugar del estadístico *F*.

Antes de dejar este capítulo, debe hacerse notar que ejemplos como el 5.3 bien pueden tener problemas que *sí* requieran una atención especial. En el caso de una variable como *narr86* que en la mayoría de los hombres de la población toma los valores 0 o 1, puede que un modelo lineal no sea capaz de captar de forma adecuada la relación funcional entre *narr86* y las variables explicativas. Además, aun cuando un modelo lineal sí describa el valor esperado de las detenciones, la heterocedasticidad puede ser problema. Problemas como éste no disminuyen a medida que el tamaño de la muestra aumenta y en capítulos posteriores se volverá a ellos.

TÉRMINOS CLAVE

Asintóticamente eficiente
Consistencia
Error estándar asintótico
Estadístico de puntuación
Estadístico del multiplicador de Lagrange (*ML*)
Estadístico *n-R*-cuadrada
Estadístico *t* asintótico
Inconsistencia
Intervalo de confianza asintótico
Normalidad asintótica
Propiedades asintóticas
Propiedades de muestras grandes
Regresión auxiliar
Sesgo asintótico
Varianza asintótica

PROBLEMAS

5.1 En el modelo de regresión simple bajo RLM.1 a RLM.4, se argumentó que el estimador de pendiente, $\hat{\beta}_1$, es consistente para β_1. Empleando $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, muestre que $\hat{\beta}_0 = \beta_0$. [Es necesario emplear la consistencia de $\hat{\beta}_1$ y la ley de los grandes números, junto con el hecho de que $\beta_0 = \mathrm{E}(y) - \beta_1\mathrm{E}(x_1)$.]

5.2 Suponga que el modelo

$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisface los primeros cuatro supuestos de Gauss-Markov, donde *pctstck* es el porcentaje de la pensión de un trabajador invertida en el mercado de valores, *funds* de la cantidad de fondos mutualistas de donde el trabajador puede elegir y *risktol* es una medida de la tolerancia al riesgo (*risktol* grande significa que la persona tiene una alta tolerancia al riesgo). Si *funds* y *risktol* están correlacionadas positivamente, ¿cuál es la inconsistencia en $\tilde{\beta}_1$, el coeficiente de pendiente en la regresión simple de *pctstck* sobre *funds*?

5.3 La base de datos SMOKE.RAW contiene información sobre la conducta como fumadores y otras variables de los adultos solteros estadounidenses en una muestra aleatoria. La variable *cigs* es la cantidad (promedio) de cigarros fumados por día. ¿Piensa que la variable *cigs* tiene una distribución normal en la población de estadounidenses adultos? Explique.

5.4 En el modelo de regresión simple (5.16), bajo los primeros cuatro supuestos de Gauss-Markov, se mostró que los estimadores de la forma (5.17) son consistentes para la pendiente, β_1. Dado un estimador de este tipo, se define un estimador de β_0 mediante $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$. Muestre que $\tilde{\beta}_0 = \beta_0$.

EJERCICIOS EN COMPUTADORA

C5.1 Para este ejercicio emplee los datos del archivo WAGE1.RAW.

i) Estime la ecuación

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Guarde los residuales y trace un histograma.

ii) Repita el inciso i), pero empleando como variable dependiente log(*wage*).

iii) ¿Diría usted que el supuesto RLM.6 está más cerca de ser satisfecho en el modelo nivel-nivel o en el modelo log-nivel?

C5.2 Para este ejercicio emplee los datos del archivo GPA2.RAW.

i) Empleando las 4,137 observaciones estime la ecuación

$$colgpa = \beta_0 + \beta_1 hsperc + \beta_2 sat + u$$

y dé los resultados en la forma estándar.

ii) Empleando las primeras 2,070 observaciones estime de nuevo la ecuación del inciso i).

iii) Determine el cociente de los errores estándar obtenidos en los incisos i) y ii) para *hsperc*. Compare este resultado con el de (5.10).

C5.3 En la ecuación (4.42) del capítulo 4, calcule el estadístico *ML* para probar si *motheduc* y *fatheduc* son conjuntamente significativas. Al obtener los residuales del modelo restringido, tenga cuidado de que éste se estime empleando sólo aquellas observaciones para las que estén disponibles todas las variables en el modelo no restringido (vea el ejemplo 4.9).

C5.4 Existen varios estadísticos que son comúnmente empleados para detectar la falta de normalidad en las distribuciones poblacionales subyacentes. Aquí se estudiará uno de estos estadísticos que mide el sesgo de una distribución. Recuerde que todas las variables aleatorias distribuidas normalmente son simétricas respecto a su media; por tanto, si una variable aleatoria distribuida de manera simétrica se estandariza, por ejemplo $z = (y - \mu_y)/\sigma_y$, donde $\mu_y = \mathrm{E}(y)$ y $\sigma_y = \mathrm{de}(y)$, entonces z tiene media cero, varianza uno y $\mathrm{E}(z^3) = 0$. Dada una muestra de datos $\{y_i : i = 1, ..., n\}$, los valores y_i en la muestra pueden estandarizarse mediante $z_i = (y_i - \hat{\mu}_y)/\hat{\sigma}_y$, donde $\hat{\mu}_y$ es la media muestral y $\hat{\sigma}_y$ es la desviación estándar muestral. (Se ignora el hecho de que éstas sean estimaciones basadas en la muestra.) Un estadístico muestral que mide la asimetría es $n^{-1}\sum_{i=1}^{n} z_i^3$, o aquel en el que en lugar de n se emplea $(n - 1)$ como ajuste para los grados de libertad. Si y tiene una distribución normal en la población, la asimetría medida en los valores estandarizados de la muestra no debe ser significativamente diferente de cero.

i) Emplee primero la base de datos 401KSUBS.RAW, utilizando sólo aquellas observaciones para las que $fsize = 1$. Determine la asimetría de *inc*. Haga lo mismo con log(*inc*). ¿Qué variable tiene más asimetría y, por tanto, parece ser menos probable que esté distribuida normalmente?

ii) A continuación emplee BWGHT2.RAW. Determine la asimetría de *bwght* y de log(*bwght*). ¿Qué concluye?

iii) Evalúe la afirmación siguiente: "La transformación logarítmica siempre hace que una variable positiva parezca estar distribuida más normalmente".

iv) Si en el contexto de la regresión interesa el supuesto de normalidad, ¿se deben evaluar las distribuciones no condicionales de y y de $\log(y)$? Explique.

Apéndice 5A

Normalidad asintótica de MCO

A continuación se esboza una prueba de la normalidad asintótica de MCO [Teorema 5.2 i)] para el caso de la regresión simple. Se escribe el modelo de regresión simple como en la ecuación (5.16). A continuación, mediante el álgebra usual de la regresión simple, se puede escribir

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/s_x^2)\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right],$$

donde para denotar la varianza de muestreo de $\{x_i: i = 1, 2, ..., n\}$ se usa s_x^2. De acuerdo con la ley de los grandes números (vea el apéndice C), $s_x^2 \xrightarrow{p} \sigma_x^2 = \mathrm{Var}(x)$. El supuesto RLM.3 descarta la colinealidad perfecta, lo que significa que $\mathrm{Var}(x) > 0$ (x_i varía en la muestra, y por tanto x no es constante en la población). Después, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i = n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i + (\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$, donde $\mu = \mathrm{E}(x)$ es la media poblacional de las x. Ahora $\{u_i\}$ es una secuencia de variables aleatorias i.i.d. cuya media es cero y cuya varianza es σ^2, y de esta manera, a medida que $n \to \infty$ $n^{-1/2}\sum_{i=1}^{n}u_i$ converge a la distribución Normal$(0,\sigma^2)$; esto es precisamente el teorema del límite central del apéndice C. De acuerdo con la ley de los grandes números, $\mathrm{plim}(\mu - \bar{x}) = 0$. Un resultado estándar de la teoría asintótica es que si $\mathrm{plim}(w_n) = 0$ y z_n tiene una distribución asintótica normal, entonces $\mathrm{plim}(w_n z_n) = 0$. [Vea Wooldridge (2002, capítulo 3 para una mayor discusión).] Esto implica que $(\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$ tiene plim igual a cero. Después, $\{(x_i - \mu)u_i: i = 1, 2, ...\}$ es una secuencia indefinida de variables aleatorias i.i.d. cuya media es cero —dado que u y x no están correlacionadas bajo el supuesto RLM.4— y varianza $\sigma^2\sigma_x^2$ de acuerdo con el supuesto RLM.5 de homocedasticidad. Por tanto, $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ tiene una distribución asintótica Normal$(0,\sigma^2\sigma_x^2)$. Se acaba de demostrar que la diferencia entre $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ y $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ tiene plim igual a cero. Un resultado de la teoría asintótica es que si z_n tiene una distribución asintótica normal y $\mathrm{plim}(v_n - z_n) = 0$, entonces v_n tiene la misma distribución asintótica normal que z_n. Se sigue que $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ también tiene una distribución asintótica Normal$(0,\sigma^2\sigma_x^2)$. De todo esto se obtiene

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/\sigma_x^2)\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right] + [(1/s_x^2) - (1/\sigma_x^2)]\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right],$$

y como $\mathrm{plim}(1/s_x^2) = 1/\sigma_x^2$, el plim del segundo término es cero. Por tanto, la distribución asintótica de $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ es Normal$(0,\{\sigma^2\sigma_x^2\}/\{\sigma_x^2\}^2)$ = Normal$(0,\sigma^2/\sigma_x^2)$. Con esto termina la demostración para el caso de la regresión simple, ya que en este caso $a_1^2 = \sigma_x^2$. Para el caso general vea Wooldridge (2002, capítulo 4).

CAPÍTULO 6

Análisis de regresión múltiple: temas adicionales

En este capítulo se reúnen varios temas sobre el análisis de regresión múltiple que no pudieron ser vistos de forma adecuada en los capítulos anteriores. Estos temas no son tan fundamentales como el material visto en los capítulos 3 y 4, pero son importantes para la aplicación de la regresión múltiple a una amplia gama de problemas empíricos.

6.1 Efectos del escalamiento de datos sobre los estadísticos de MCO

En el capítulo 2 sobre regresión bivariada se analizaron de forma breve los efectos que la modificación de las unidades de medición tienen sobre el intercepto y la pendiente estimadas por MCO. También se mostró que la modificación de las unidades de medición no afecta la *R*-cuadrada. Ahora se vuelve al tema del escalamiento de datos y se examinan los efectos que escalar las variables dependiente o independiente tiene sobre los errores estándar, los estadísticos *t,* los estadísticos *F* y los intervalos de confianza.

Se encontrará que todo lo que se espera que ocurra, en efecto ocurre. Cuando se reescalan las variables, tanto los coeficientes como los errores estándar, los intervalos de confianza, los estadísticos *t* y los estadísticos *F* se modifican de una manera tal que se preservan todos los efectos medidos y los resultados de las pruebas. Aunque esto no es ninguna sorpresa —preocuparía que no fuera así— es útil ver de forma explícita lo que ocurre. Con frecuencia, el escalamiento de datos se emplea con fines cosméticos, por ejemplo para reducir la cantidad de ceros después del punto decimal en un coeficiente estimado. Eligiendo de manera adecuada las unidades de medición puede mejorarse la apariencia de una ecuación estimada sin modificar nada esencial.

Este problema podría tratarse de manera general, pero se ilustra mucho mejor con ejemplos. Asimismo, no tiene objeto introducir aquí una notación abstracta.

Se comienza con una ecuación en la que se relaciona el peso de los niños al nacer con la cantidad de cigarros fumados y el ingreso familiar:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc, \qquad \boxed{\textbf{6.1}}$$

donde

$bwght$ = peso del niño al nacer, en onzas.
$cigs$ = cantidad de cigarros por día que fumó la madre durante el embarazo.
$faminc$ = ingreso familiar anual, en miles de dólares.

TABLA 6.1

Efectos del escalamiento de datos

Variable dependiente	**(1) *bwght***	**(2) *bwghtlbs***	**(3) *bwght***
Variables independientes			
cigs	−.4634 (.0916)	−.0289 (.0057)	——
packs	——	——	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercepto	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observaciones	1,388	1,388	1,388
R-cuadrada	.0298	.0298	.0298
SRC	557,485.51	2,177.6778	557,485.51
EER	20.063	1.2539	20.063

En la primera columna de la tabla 6.1 se dan las estimaciones de esta ecuación obtenidos empleando los datos del archivo BWGHT.RAW. Los errores estándar se dan entre paréntesis. El estimado de *cigs* indica que si una mujer fuma cinco cigarros más por día, se pronostica que el peso del niño al nacer será aproximadamente .4634(5) = 2.317 onzas menos. El estadístico t para *cigs* es −5.06, de manera que esta variable es estadísticamente muy significativa.

Supóngase que ahora el peso al nacer se mide en libras, y no en onzas. Sea *bwghtlbs* = *bwght*/16 el peso al nacer en libras. ¿Qué pasa con los estadísticos de MCO si en la ecuación se emplea ésta como variable dependiente? El efecto sobre los coeficientes estimados se encuentra con facilidad mediante una sencilla manipulación de la ecuación (6.1). Toda la ecuación se divide entre 16:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)faminc.$$

Como el lado izquierdo de la ecuación es el peso al nacer dado en libras, resulta que cada uno de los coeficientes nuevos será el correspondiente coeficiente anterior dividido entre 16. Para comprobar esto, en la columna (2) de la tabla 6.1 se da la regresión de *bwghtlbs* sobre *cigs* y *faminc* A cuatro cifras decimales, el intercepto y las pendientes de la columna (2) son las de la columna (1) divididas entre 16. Por ejemplo, ahora el coeficiente de *cigs* es −.0289; esto significa que si *cigs* fuera mayor en cinco unidades, el peso al nacer sería .0289(5) = .1445 *libras* menor. En términos de onzas, se tiene .1445(16) = 2.312, lo cual, debido al error de redondeo, es un poco diferente del 2.317 obtenido antes. El punto es que una vez que los efectos se transforman a las mismas unidades, se obtiene exactamente el mismo resultado, al margen de la manera en que se mida la variable dependiente.

¿Qué ocurre con la significancia estadística? Como era de esperarse, modificar la variable dependiente de onzas a libras no tiene ningún efecto sobre la importancia estadística de las variables independientes. En la columna (2) los errores estándar son 16 veces menores que en la columna (1). Con unos cuantos cálculos rápidos se ve que los estadísticos t de la columna (2) son, en realidad, idénticos a los estadísticos t de la columna (1). Los extremos de los intervalos de confianza en la columna (2) son exactamente los extremos de los intervalos de confianza en la columna (1) divididos entre 16. Esto se debe a que los IC se modifican por el mismo factor que los errores estándar. [Recuerde que aquí el IC de 95% es $\hat{\beta}_j \pm 1.96\ \mathrm{ee}(\hat{\beta}_j)$.]

En términos de la bondad de ajuste, las R-cuadradas de las dos regresiones son idénticas, como debería ser. Observe que la suma de residuales cuadrados, SRC, y el error estándar de la regresión, EER, difieren entre estas ecuaciones. Estas diferencias se explican con facilidad. Sea $\hat{u}_i$ el residual de la observación i en la ecuación original (6.1). Entonces, cuando la variable independiente es *bwghtlbs* el residual es simplemente $\hat{u}_i/16$. De manera que en la segunda ecuación el residual *cuadrado* es $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$. A esto se debe que en la columna (2) la suma de los residuales cuadrados sea igual a la SRC de la columna (1) dividida entre 256.

Dado que el EER $= \hat{\sigma} = \sqrt{\mathrm{SSR}/(n-k-1)} = \sqrt{\mathrm{SSR}/1{,}385}$, en la columna (2) el EER es 16 veces menor que en la columna (1). Otra manera de ver esto es que el error en la ecuación, en la que como variable dependiente se emplea *bwghtlbs,* tiene una desviación estándar 16 veces menor que la desviación estándar del error original. Esto no significa que al cambiar la manera en que se mide el peso al nacer se reduzca el error; el que el EER sea menor sólo refleja la diferencia en las unidades de medición.

A continuación se regresa la variable dependiente a sus unidades originales: *bwght* se mide en onzas. Ahora se modifican las unidades de medición de una de las variables independientes, *cigs*. Se define la variable *packs* como la cantidad de cajetillas de cigarros fumadas por día. De manera que, *packs* = *cigs*/20. ¿Qué pasa ahora con los coeficientes y con los demás estadísticos de MCO? Bueno, se puede escribir

$$\widehat{bwght} = \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2 faminc = \hat{\beta}_0 + (20\hat{\beta}_1)packs + \hat{\beta}_2 faminc.$$

Por tanto, el intercepto y el coeficiente de pendiente de *faminc* no cambia, pero el coeficiente de *packs* es 20 veces el de *cigs*. Esto es intuitivamente interesante. En la columna (3) de la tabla 6.1 se presentan los resultados de la regresión de *bwght* sobre *packs* y *faminc*. A propósito, recuerde que no tendría caso incluir en una misma ecuación *cigs* y *packs* esto induciría una multicolinealidad perfecta y no tendría ningún significado.

Además del coeficiente de *packs*, en la columna (3) no hay otro estadístico que sea diferente de los estadísticos de la columna (1): el error estándar de *packs* es 20 veces mayor que el de *cigs,* que aparece en la columna (1). Esto significa que el estadístico t para probar la significancia de los cigarros fumados es el mismo ya sea que éstos se midan en términos de cigarros individuales o de cajetillas. Esto es natural.

Pregunta 6.1

Suponga que en la ecuación original del peso al nacer (6.1), *faminc* se mide en dólares y no en miles de dólares. Por tanto, se define la variable *fincdol* = 1,000·*faminc*. ¿Cómo se modificarán los estadísticos de MCO al emplear *fincdol* en lugar de *faminc*? Para presentar los resultados de la regresión, ¿considera usted que es mejor medir el ingreso en dólares o en miles de dólares?

El ejemplo anterior explica con detalle la mayoría de las posibilidades que surgen cuando se modifican las unidades de medición de las variables dependiente e independiente. En economía suelen cambiarse las unidades de medición cuando se trata de cantidades de dólares, en especial cuando son cantidades muy grandes.

En el capítulo 2 se dijo que, si la variable dependiente aparece de forma logarítmica, modificar las unidades de medición no afecta el coeficiente de pendiente. Lo mismo es válido aquí: modificar las unidades de medición de la variable dependiente, cuando ésta aparece de forma logarítmica, no afecta a ninguna de las estimaciones de pendiente. Esto es consecuencia del sencillo hecho de que $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ para cualquier constante $c_1 > 0$. El nuevo intercepto será $\log(c_1) + \hat{\beta}_0$. De manera similar, modificar las unidades de medición de cualquiera de las x_j, cuando en la regresión aparece $\log(x_j)$, sólo afecta el intercepto. Esto corresponde a lo que ya se sabe acerca de cambios porcentuales y, en particular, de elasticidades: son invariantes a las unidades de medición ya sea de y o de x_j. Por ejemplo, si en (6.1) se hubiera especificado log(*bwght*), como la variable dependiente, se hubiera estimado la ecuación y después vuelta a estimar con log(*bwghtlbs*) como variable dependiente, en ambas regresiones los coeficientes de *cigs* y de *faminc* hubieran sido los mismos, sólo el intercepto hubiera sido diferente.

Coeficientes beta

En econometría, algunas veces una variable clave se mide en una escala que es difícil de interpretar. Los economistas laborales suelen incluir en las ecuaciones de salario puntuaciones de pruebas, y las escalas que se usan para las puntuaciones de estas pruebas suelen ser arbitrarias y difíciles de interpretar (¡por lo menos para los economistas!). En casi todos los casos lo que interesa es comparar la puntuación de un determinado individuo con la de la población. De esta manera, en lugar de preguntar por el efecto sobre el salario por hora si, por ejemplo, la puntuación en una prueba es 10 puntos superior, es más sutil preguntar qué pasa cuando la puntuación de la prueba es una *desviación estándar* superior.

Nada impide que se vea lo que ocurre con la variable dependiente cuando en un modelo estimado una variable independiente aumenta cierto número de desviaciones estándar, suponiendo que ya se haya obtenido la desviación estándar muestral (lo cual es fácil con la mayoría de los paquetes para regresión). Esto suele ser una buena idea. Así, por ejemplo, cuando se ve el efecto de la puntuación obtenida en un examen estándar, como en el SAT (examen de admisión estándar, en Estados Unidos) sobre el GPA (promedio general en la universidad), puede determinarse la desviación estándar del SAT y ver lo que ocurre cuando la puntuación en el SAT aumenta una o dos desviaciones estándar.

Algunas veces es útil obtener resultados de la regresión cuando *todas* las variables que intervienen, tanto la variable dependiente como las variables independientes, han sido *estandarizadas*. Una variable está estandarizada cuando se le resta su media y se divide entre su desviación estándar (vea el apéndice C). Esto significa que para cada una de las variables de la muestra se calcule su *valor-z*. Después se corre la regresión empleando los valores-z.

¿Por qué es útil la estandarización? Lo más fácil es partir de la ecuación original de MCO, con las variables en sus formas originales:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik} + \hat{u}_i. \quad \boxed{6.2}$$

Para hacer hincapié en que la estandarización se aplica a todos los valores muestrales se han incluido los subíndices correspondientes a la observación i. Ahora, si se obtiene el promedio de (6.2), se usa el hecho de que $\hat{u}_i$ tiene media muestral cero, y el resultado se sustrae de (6.2), se obtiene

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \dots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

Ahora, sea $\hat{\sigma}_y$ la desviación estándar muestral de la variable dependiente, sea $\hat{\sigma}_1$ la *ds* muestral de x_1, sea $\hat{\sigma}_2$ la *ds* muestral de x_2, etc. Entonces, mediante álgebra sencilla se obtiene la ecuación

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + \dots + (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{u}_i/\hat{\sigma}_y). \tag{6.3}$$

Cada variable en (6.3) ha sido estandarizada sustituyéndola por su valor-z, lo que ha dado como resultado nuevos coeficientes de pendiente. Por ejemplo, el coeficiente de pendiente de $(x_{i1} - \bar{x}_1)/\hat{\sigma}_1$ es $(\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1$. Este coeficiente es simplemente el coeficiente original, $\hat{\beta}_1$, multiplicado por el cociente de la desviación estándar de x_1 entre la desviación estándar de y. El intercepto ha desaparecido por completo.

Es útil reescribir (6.3), eliminando los subíndices i, como

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \dots + \hat{b}_k z_k + error, \tag{6.4}$$

donde z_y denota el valor-z de y, z_1 denota el valor-z de x_1, y así sucesivamente. Los nuevos coeficientes son

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \text{ para } j = 1, \dots, k. \tag{6.5}$$

A los $\hat{b}_j$ se les conoce como **coeficientes estandarizados** o **coeficientes beta**. (El último nombre es el más común, lo cual no es muy afortunado debido a que beta gorro es lo que se ha empleado para denotar las estimaciones *usuales* de MCO.)

La ecuación (6.4) confiere a los coeficientes beta un significado interesante: si x_1 aumenta en una desviación estándar, entonces $\hat{y}$ se modifica en $\hat{b}_1$ desviaciones estándar. De esta manera, los efectos se miden no en términos de las unidades originales de y o de x_j, sino de unidades de desviaciones estándar. Como esto vuelve irrelevantes las escalas de los regresores, esta ecuación coloca a las variables explicativas en igualdad de condiciones. En una ecuación de MCO usual, no es posible que simplemente mirando las distintas magnitudes de los coeficientes se concluya que la variable explicativa con el coeficiente mayor sea "la más importante". Se acaba de ver que cambiando las unidades de medición de las x_j pueden modificarse las magnitudes de los coeficientes según se desee. Pero, cuando se ha estandarizado cada una de las x_j, comparar las magnitudes de los coeficientes beta obtenidos es más informativo.

Aun cuando los coeficientes tengan una interpretación sencilla —por ejemplo, cuando la variable dependiente y las variables independientes de interés se encuentran en forma logarítmica, de manera que los coeficientes de MCO de interés son elasticidades estimadas— sigue siendo útil calcular los coeficientes beta. Aunque las elasticidades no tienen unidades de medición, una variación de una determinada variable explicativa de, por ejemplo, 10% puede representar una variación mayor o menor sobre el rango de esa variable que una variación de otra variable explicativa de 10%. Por ejemplo, en un estado en el que el ingreso sea grande pero la variación en el gasto por estudiante sea relativamente pequeña, no tendrá mucho sentido comparar la elasticidad del ingreso con la del gasto. Comparar las magnitudes de los coeficientes beta sí puede ser útil.

Para obtener los coeficientes beta, siempre pueden estandarizarse y, x_1, …, x_k y después correr la regresión de MCO de y sobre los valores-z de x_1, …, x_k —caso en el que no es necesario incluir el intercepto, ya que éste será cero—. Cuando son muchas las variables independientes esto puede ser tedioso. Algunos paquetes para regresión proporcionan los coeficientes beta mediante un sencillo comando. El ejemplo siguiente muestra el empleo de los coeficientes beta.

Ejemplo 6.1

[Efectos de la contaminación sobre el precio de la vivienda]

Para ejemplificar el uso de los coeficientes beta se emplearán los datos del ejemplo 4.5 (del archivo HPRICE2.RAW). Recuerde que la variable independiente clave es *nox*, una medida de la cantidad de óxido de nitrógeno en el aire de cada comunidad. Una manera de entender la magnitud del efecto de la contaminación —sin profundizar en los conocimientos subyacentes al efecto del óxido de nitrógeno en la calidad del aire— es calcular los coeficientes beta. (Otro método es el que se dio en el ejemplo 4.5: se obtuvo la elasticidad del precio respecto a *nox* empleando *price* y *nox* en forma logarítmica.)

La ecuación poblacional es el modelo nivel-nivel

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u,$$

donde todas las variables excepto *crime* ya fueron definidas en el ejemplo 4.5; *crime* es la cantidad de delitos reportado *per cápita*. En la ecuación siguiente se dan los coeficientes beta (así que cada variable ha sido transformada en su valor-*z*):

$$\widehat{zprice} = -.340\ znox - .143\ zcrime + .514\ zrooms - .235\ zdist - .270\ zstratio.$$

Esta ecuación indica que un aumento de *nox* en una desviación estándar hace que el precio disminuya .34 desviaciones estándar; un aumento de *crime* en una desviación estándar hace que el precio disminuya .14 desviaciones estándar. De manera que una variación relativa de la contaminación poblacional tiene un efecto mayor sobre el precio de la vivienda que la misma variación relativa de la delincuencia. El tamaño de la vivienda, medido por el número de habitaciones (*rooms*), es lo que tiene el mayor efecto estandarizado. Para saber cuál es el efecto de cada una de las variables independientes sobre el valor en dólares del precio mediano de la vivienda, es necesario emplear las variables no estandarizadas.

Emplear las variables estandarizada o no estandarizada no afecta a la significancia estadística: en ambos casos los estadísticos *t* son los mismos.

6.2 Más acerca de la forma funcional

En varios de los ejemplos anteriores ha aparecido el recurso más empleado en econometría para tomar en cuenta relaciones no lineales entre la variable explicada y las variables explicativas: el empleo de logaritmos para las variables, dependiente o independiente. También se han encontrado modelos que contienen términos cuadráticos para algunas de las variables explicativas; ahora se desea proporcionar un tratamiento sistemático de estos casos. En esta sección se verán algunas variaciones y extensiones de las formas funcionales, las cuales suelen encontrarse en el trabajo aplicado.

Más acerca del empleo de las formas funcionales logarítmicas

Se comenzará viendo de nuevo la interpretación de los parámetros en el modelo

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \tag{6.6}$$

en el que las variables han sido tomadas del ejemplo 4.5. Recuerde que en todo el libro $\log(x)$ es el logaritmo *natural* de x. El coeficiente β_1 es la elasticidad de *price* respecto a *nox* (contaminación). El coeficiente β_2 es la variación que hay en $\log(price)$, cuando $\Delta rooms = 1$; como se ha visto varias veces, multiplicado por 100 es el cambio porcentual aproximado de *price*. Recuerde que $100 \cdot \beta_2$ suele conocerse como la semielasticidad de *price* respecto a *rooms*.

Al estimar empleando los datos del archivo HPRICE2.RAW, se obtiene

$$
\begin{aligned}
\widehat{\log(price)} &= \underset{(0.19)}{9.23} - \underset{(.066)}{.718} \log(nox) + \underset{(.019)}{.306}\, rooms \\
n &= 506,\ R^2 = .514.
\end{aligned}
\tag{6.7}
$$

Por tanto, cuando *nox* aumenta 1%, *price* disminuye .718%, manteniendo *rooms* constante. Cuando *rooms* aumenta en uno, *price* aumenta aproximadamente 100(.306) = 30.6%.

La estimación de que una habitación más incremente el precio aproximadamente 30.6% resulta ser un poco inexacto en esta aplicación. El error de aproximación se debe a que a medida que la variación de $\log(y)$ es mayor, la aproximación $\%\Delta y \approx 100 \cdot \Delta\log(y)$ se hace cada vez más inexacta. Por fortuna, existe un cálculo sencillo que permite calcular el cambio porcentual exacto.

Para describir este procedimiento, se considerará el modelo general estimado

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2.$$

(Agregar más variables independientes no modifica el procedimiento.) Ahora, fijando x_1, se tiene $\Delta\widehat{\log(y)} = \hat{\beta}_2 \Delta x_2$. Al emplear propiedades algebraicas sencillas de las funciones exponencial y logarítmica se obtiene la variación porcentual *exacta* pronosticada para y

$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_2 \Delta x_2) - 1], \tag{6.8}$$

donde la multiplicación por 100 convierte la variación proporcional en una variación porcentual. Cuando $\Delta x_2 = 1$,

$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_2) - 1]. \tag{6.9}$$

Aplicado al ejemplo del precio de la vivienda con $x_2 = rooms$ y $\hat{\beta}_2 = .306$, $\%\Delta\widehat{price} = 100[\exp(.306) - 1] = 35.8\%$, que es claramente mayor que la variación porcentual aproximada, 30.6%, que se obtuvo directamente con la ecuación (6.7). {A propósito, éste no es un estimador insesgado porque $\exp(\cdot)$ es una función no lineal; sin embargo, es un estimador consistente de $100[\exp(\beta_2) - 1]$. Esto se debe a que el límite de probabilidad pasa por funciones continuas, mientras que el operador del valor esperado no. Vea el apéndice C.}

El ajuste en la ecuación (6.8) no es tan crucial con cambios porcentuales pequeños. Por ejemplo, cuando en la ecuación (6.7) se incluye el cociente estudiante-profesor (*stratio*), su coeficiente estimado es −.052, lo que significa que si *stratio* aumenta en uno, *price* disminuye aproximadamente 5.2%. La variación proporcional exacta es $\exp(-.052) - 1 \approx -.051$, o sea −5.1%. Por otro lado, si *stratio* aumenta en cinco, entonces la variación porcentual aproximada del precio (*price*) es −26%, mientras que la variación exacta que se obtiene con la ecuación (6.8) es $100[\exp(-.26) - 1] \approx -22.9\%$.

La aproximación logarítmica para las variaciones porcentuales tiene una ventaja que justifica el proporcionarla en los resultados aun cuando la variación porcentual sea grande. Para describir esta ventaja, considere de nuevo el efecto sobre el precio de una variación de uno en la cantidad de habitaciones. La aproximación logarítmica es simplemente el coeficiente de *rooms*, en la ecuación (6.7), multiplicado por 100, es decir, 30.6%. Se calculó también que una estimación de la variación porcentual exacta al *aumentar* la cantidad de habitaciones en uno *es* 35.8%. Pero, ¿qué ocurre si se desea estimar el cambio porcentual cuando la cantidad de habitaciones *disminuye* en uno? En la ecuación (6.8) se tendrán $\Delta x_2 = -1$ y $\hat{\beta}_2 = .306$, y de esta manera $\%\Delta\widehat{price} = 100[\exp(-.306) - 1] = -26.4$, es decir, una disminución de 26.4%. Observe que la aproximación basada en el empleo del coeficiente de *rooms* está entre 26.4 y 35.8 —un resultado que siempre se presenta—. En otras palabras, el uso simple del coeficiente (multiplicado por 100) da una estimación que está siempre entre los valores absolutos de las estimaciones correspondientes a un aumento y a una disminución. Si interesa en específico un aumento o una disminución, se emplean los cálculos basados en la ecuación (6.8).

Lo que se acaba de señalar acerca del cálculo de variaciones porcentuales es en esencia lo que se señala en la introducción a la economía cuando se trata del cálculo, por ejemplo, de elasticidades precio de la demanda con base en variaciones grandes del precio: el resultado depende de si se emplea el precio y la cantidad iniciales o finales en el cálculo de los cambios porcentuales. Emplear la aproximación logarítmica es similar en esencia a calcular la elasticidad arco de la demanda, en donde para calcular las variaciones porcentuales se emplean en los denominadores precios y cantidades promedio.

Se ha visto que emplear logaritmos naturales conduce a coeficientes con una interpretación interesante, y que pueden ignorarse las unidades de medición de las variables que aparecen en forma logarítmica porque los coeficientes de pendiente no varían ante un cambio de unidades. Existen algunas otras razones a las que se debe que los logaritmos sean tan empleados en las aplicaciones. Primero, cuando $y > 0$, los modelos en los que como variable dependiente se emplea $\log(y)$ suelen satisfacer mejor los supuestos del MLC que los modelos en los que se emplea y en forma lineal. Variables estrictamente positivas suelen tener distribuciones condicionales que son heterocedásticas o asimétricas; empleando logaritmos, ambos problemas pueden atenuarse, o incluso eliminarse.

Además, en algunos casos, el empleo de logaritmos suele estrechar el rango de la variable, en una cantidad considerable. Esto hace las estimaciones menos sensibles a observaciones atípicas (o extremas) de las variables dependiente o independiente. En el capítulo 9 se verá el tema de las observaciones atípicas.

Para el uso de logaritmos existen algunas reglas prácticas, aunque ninguna está escrita en piedra. Cuando una variable es una cantidad positiva en dólares, suele emplearse su logaritmo. Esto se ha visto en el caso de variables tales como sueldos, salarios, ventas y valor de mercado de las empresas. Variables como población, cantidad de empleados y matrícula escolar suelen aparecer en forma logarítmica; éstas tienen la característica común de ser cantidades enteras grandes.

Variables que se miden en años —como educación, experiencia, antigüedad, edad, etc.— usualmente aparecen en su forma original. Variables que son una proporción o un porcentaje —como tasa de desempleo, tasa de participación en un plan de pensiones, porcentaje de estudiantes que aprueban un examen estandarizado y tasa de detención por delitos reportados— pueden aparecer ya sea en su forma original o en forma logarítmica, aunque hay una tendencia a usarlas en su forma lineal. Esto se debe a que cualquier coeficiente de regresión relacionado con la variable *original* —sea la variable dependiente o independiente— tendrá una interpretación como variación en *puntos porcentuales*. (Vea en el apéndice A un repaso de la diferencia entre variación porcentual y variación en puntos porcentuales.) Si en una regresión se usa, por ejemplo, log(*unem*) donde *unem* es el porcentaje de individuos desempleados, se debe tener mucho cuidado de distinguir entre una variación de un punto porcentual y una variación porcentual. Recuerde, si *unem* varía de 8 a 9, este es un aumento de un punto porcentual, pero un aumento de 12.5% a partir del nivel inicial de desempleo. Usar el logaritmo significa que se está poniendo atención a la variación porcentual de la tasa de desempleo: $\log(9) - \log(8) \approx .118$ es decir 11.8%, que es la aproximación logarítmica del aumento real de 12.5 por ciento.

Pregunta 6.2

Suponga que la cantidad anual de detenciones de personas que conducen bajo los efectos del alcohol se determina mediante

$$\log(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2 age16\_25 + other\ factors,$$

donde *age 16_25* es la proporción de la población entre 16 y 25 años de edad. Demuestre que β_2 tiene la interpretación (*ceteris paribus*) siguiente: es el cambio porcentual en la variable *arrests* (arrestos) al aumentar el porcentaje de personas de 16 a 25 años de edad en un *punto porcentual*.

Una limitación de los logaritmos en que no pueden usarse si una variable toma valores negativos o cero. Cuando una variable y es no negativa, pero puede tomar el valor 0 suele emplearse $\log(1 + y)$. Las interpretaciones de la variación porcentual suelen preservarse, salvo cuando se trata de variaciones que comienzan en $y = 0$ (donde la variación porcentual incluso no está definida). En general, usar $\log(1 + y)$ y después interpretar las estimaciones como si la variable fuera $\log(y)$ es aceptable cuando los datos de y contienen relativamente poco ceros. Un ejemplo puede ser aquel en el que y es horas de capacitación por empleado en la población de empresas manufactureras, si una proporción grande de las empresas proporciona capacitación al menos a un trabajador. Sin embargo, técnicamente, $\log(1 + y)$ no puede estar distribuida normalmente (aunque puede ser menos heterocedástica que y). Otras posibilidades útiles, aunque más avanzadas, son los modelos Tobit y Poisson del capítulo 17.

Una desventaja de usar la variable dependiente de forma logarítmica es que es más difícil pronosticar la variable original. El modelo original permite pronosticar $\log(y)$, no y. Sin embargo, es bastante sencillo convertir un pronóstico para $\log(y)$ en un pronóstico para y (ver la sección 6.4). Un punto relacionado con esto es que *no* es correcto comparar las R-cuadradas de modelos en los que y es la variable dependiente en un caso y $\log(y)$ es variable dependiente en otro. Estas mediciones explican variaciones de variables diferentes. En la sección 6.4 se verá cómo calcular medidas comparables de bondad de ajuste.

Modelos con funciones cuadráticas

Las **funciones cuadráticas** se emplean también con bastante frecuencia en economía para captar efectos marginales crecientes o decrecientes. En el apéndice A se encuentra un repaso de las propiedades de las funciones cuadráticas.

El caso más simple, es aquel en el que y depende de un solo factor observado x, pero lo hace de forma cuadrática:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Por ejemplo, tómese $y = wage$ (salario) y $x = exper$ (experiencia). Como se vio en el capítulo 3, este modelo cae fuera del análisis de regresión simple pero puede ser tratado con facilidad empleando regresión múltiple.

Es importante recordar que β_1 no mide la variación en y respecto a x; no tiene ningún sentido mantener x^2 constante mientras se varía x. Si la ecuación estimada se expresa como

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \tag{6.10}$$

entonces se tiene la aproximación

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ de manera que } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \tag{6.11}$$

Esto indica que la pendiente de la relación entre x y y dependen del valor de x; la pendiente estimada es $\hat{\beta}_1 + 2\hat{\beta}_2 x$. Si se sustituye con $x = 0$, se ve que $\hat{\beta}_1$ puede interpretarse como la pendiente aproximada al pasar de $x = 0$ a $x = 1$. Después de eso, el segundo término, $2\hat{\beta}_2 x$, debe ser tomado en cuenta.

Si lo único que interesa es calcular el cambio predicho en y dado un valor inicial de x y una variación de x, puede emplearse directamente (6.10): no hay ninguna razón para emplear el cálculo de aproximación. Sin embargo, en general lo que más interesa es resumir con rapidez el efecto de x sobre y, y, en este caso, la interpretación de $\hat{\beta}_1$ y $\hat{\beta}_2$ en (6.11) proporciona este resumen. Típicamente, x se sustituye por su valor promedio en la muestra o por otros valores interesantes como la mediana o los valores de los cuartiles inferior o superior.

En muchas de las aplicaciones, $\hat{\beta}_1$ es positivo y $\hat{\beta}_2$ es negativo. Por ejemplo, empleando los datos de salario del archivo WAGE1.RAW, se obtiene

$$\begin{aligned}\widehat{wage} &= \underset{(.35)}{3.73} + \underset{(.041)}{.298}\ exper - \underset{(.0009)}{.0061}\ exper^2 \\ n &= 526,\ R^2 = .093.\end{aligned} \tag{6.12}$$

La ecuación estimada indica que *exper* tiene un efecto decreciente sobre *wage*. El primer año de experiencia vale 30¢ por hora ($.298). El segundo año de experiencia vale menos [aproximadamente $.298 - 2(.0061)(1) \approx .286$, es decir 28.6¢, de acuerdo con la aproximación en (6.11) con $x = 1$]. Al pasar de 10 a 11 años de experiencia, el aumento predicho en *wage* es aproximadamente $.298 - 2(.0061)(10) = .176$, es decir 17.6¢. Y así sucesivamente.

Cuando el coeficiente de x es positivo y el de x^2 es negativo la cuadrática tiene forma parabólica. Siempre existe un valor positivo de x para el que el efecto de x sobre y es cero; antes de este punto, x tiene un efecto positivo sobre y; después, x tiene un efecto negativo sobre y. En la práctica, es importante saber dónde se encuentra este punto de inflexión.

En la ecuación estimada (6.10) si $\hat{\beta}_1 > 0$ y $\hat{\beta}_2 < 0$, el punto de inflexión (o máximo de la función) siempre se alcanzará en el punto correspondiente al coeficiente de x sobre *el doble* del valor absoluto del coeficiente de x^2:

$$x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|. \tag{6.13}$$

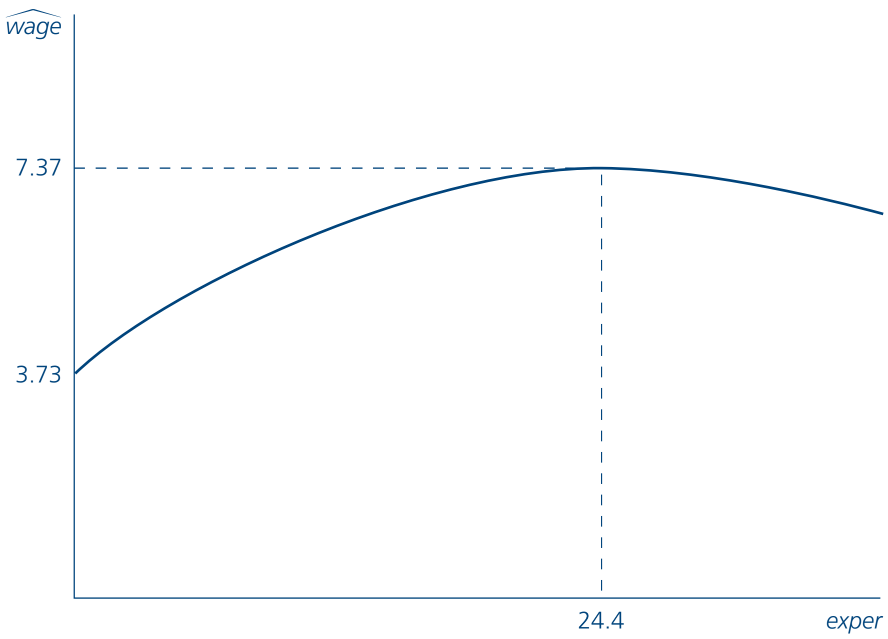
En el ejemplo del salario, $x^* = exper^*$ es $.298/[2(.0061)] \approx 24.4$. (Observe cómo al hacer este cálculo simplemente se ha eliminado el signo menos de $-.0061$.) Esta relación cuadrática se ilustra en la figura 6.1.

En la ecuación (6.12) del salario, el rendimiento de la experiencia se vuelve cero aproximadamente a los 24.4 años. ¿Cómo debe entenderse esto? Existen al menos tres explicaciones posibles. Primero, puede ser que en la muestra haya pocas personas que tengan más de 24 años de experiencia y por esto la parte de la curva a la derecha de 24 puede ignorarse. El costo de emplear ecuaciones cuadráticas para captar efectos decrecientes es que llega un momento en el que la forma cuadrática cambia de dirección. Si este punto está más allá de todas excepto un porcentaje pequeño de las personas de la muestra, entonces esto no debe preocupar mucho. Pero en la base de datos WAGE1.RAW, aproximadamente 28% de las personas de la muestra tienen más de 24 años de experiencia; este es un porcentaje muy alto como para ignorarlo.

Es posible que el rendimiento de *exper* realmente se vuelva negativo en algún punto, pero es difícil creer que esto ocurra a los 24 años de experiencia. Una posibilidad más creíble es que el efecto estimado de *exper* sobre *wage* esté sesgado debido a que no se hayan controlado otros factores, o debido a que la relación funcional entre *wage* y *exper* dada por la ecuación (6.12)

FIGURA 6.1

Relación cuadrática entre $\widehat{wage}$ y *exper*.



no sea del todo correcta. En el ejercicio para computadora C6.2 se pide al lector explorar estas posibilidades controlando la educación y empleando log(*wage*) como variable dependiente.

Cuando en un modelo hay una variable dependiente de forma logarítmica y una variable explicativa que aparece de forma cuadrática, debe tenerse cuidado al reportar los efectos parciales. El ejemplo siguiente muestra que las funciones cuadráticas pueden tener también forma de U, en lugar de la forma parabólica. En la ecuación (6.10) la forma de U surge cuando $\hat{\beta}_1$ es negativa y $\hat{\beta}_2$ es positiva; esto capta un efecto creciente de x sobre y.

Ejemplo 6.2

[Efectos de la contaminación sobre el precio de la vivienda]

Ahora se modificará el modelo del ejemplo 4.5 para el precio de la vivienda incluyendo un término cuadrático en *rooms*:

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2\log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \quad \mathbf{6.14}$$

El modelo estimado con los datos del archivo HPRICE2.RAW es

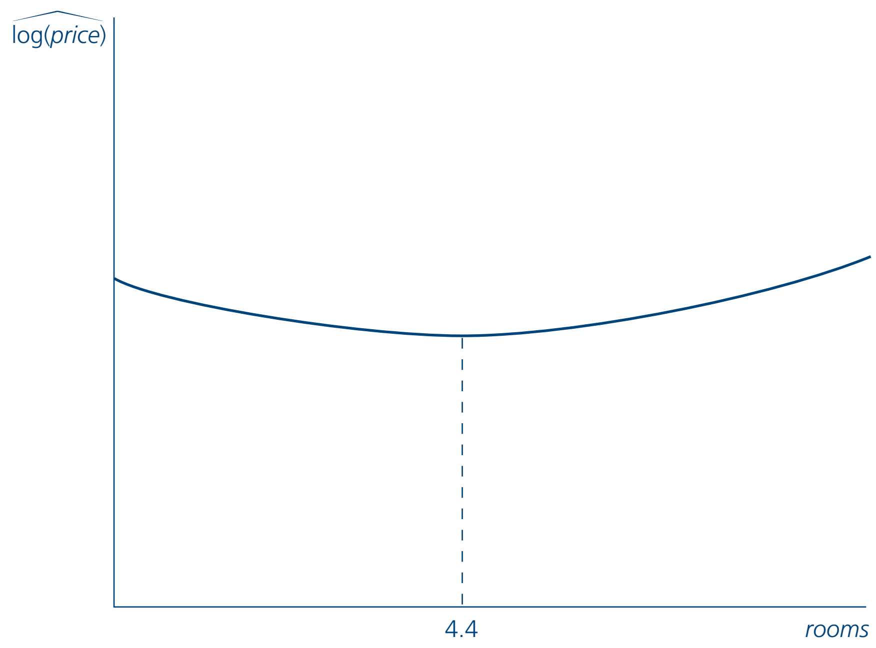
$$\begin{aligned}
\widehat{\log(price)} &= \underset{(.57)}{13.39} - \underset{(.115)}{.902} \log(nox) - \underset{(.043)}{.087} \log(dist) \\
&\quad - \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2 - \underset{(.006)}{.048}\, stratio \\
&\quad n = 506,\ R^2 = .603.
\end{aligned}$$

El estadístico t para el término cuadrático $rooms^2$ es aproximadamente 4.77, de manera que este término es estadísticamente muy significativo. Pero, ¿cuál es la interpretación del efecto de *rooms* sobre log(*price*)? En un principio, el efecto parece ser extraño. Como el coeficiente de *rooms* es negativo y el de $rooms^2$ es positivo, esta ecuación literalmente implica que, para valores bajos de *rooms*, una habitación más tiene un efecto *negativo* sobre log(*price*). En algún punto el efecto se vuelve positivo, y la forma cuadrática significa que la semielasticidad de *price* respecto a *rooms* aumenta a medida que *rooms* aumenta. Esta situación se muestra en la figura 6.2.

El valor del punto de inflexión de *rooms* se obtiene empleando la ecuación (6.13) (aun cuando $\hat{\beta}_1$ sea negativo y $\hat{\beta}_2$ sea positivo). El valor absoluto del coeficiente de *rooms*, .545, dividido entre el doble del coeficiente de $rooms^2$, .062, da $rooms^* = .545/[2(.062)] \approx 4.4$; este punto se indica en la figura 6.2.

FIGURA 6.2

$\widehat{\log(price)}$ como función cuadrática de *rooms*.



¿Puede creerse que partiendo de tres habitaciones y aumentando a cuatro en realidad se reduzca el valor esperado de una casa? Lo más seguro es que no. Resulta que sólo en cinco de las 506 comunidades de la muestra hay casas en las que la cantidad promedio de habitaciones sea 4.4 o menos, aproximadamente 1% de la muestra. Esta cantidad es tan pequeña que para fines prácticos la función cuadrática a la izquierda de 4.4 puede ignorarse. A la derecha de 4.4, se ve que agregar una habitación tiene un efecto creciente sobre el cambio porcentual del precio:

$$\widehat{\Delta\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

y de esta manera

$$\begin{aligned}\%\widehat{\Delta price} &\approx 100\{[-.545 + 2(.062)]rooms\}\Delta rooms \\ &= (-54.5 + 12.4\ rooms)\Delta rooms.\end{aligned}$$

Así, el aumento de *rooms* de, por ejemplo, cinco a seis, hace que el precio aumente aproximadamente $-54.5 + 12.4(5) = 7.5\%$; el aumento de seis a siete hace que el precio aumente aproximadamente $-54.5 + 12.4(6) = 19.9\%$. Este es un efecto creciente muy fuerte.

En general, ¿qué pasa si los coeficientes de los términos lineal y cuadrado tienen el *mismo* signo (ya sea ambos positivos o ambos negativos) y la variable explicativa sea necesariamente no negativa (como en el caso de *rooms* o de *exper*)? En cualquier caso, no hay un punto de inflexión en un valor de $x > 0$. Por ejemplo, si tanto β_1 como β_2 son positivos, el menor valor esperado de y se encuentra en $x = 0$ y el aumento de x siempre tiene un efecto positivo y creciente sobre y. (Esto ocurre también si $\beta_1 = 0$ y $\beta_2 > 0$, lo que significa que el efecto parcial en $x = 0$ es cero y crece a medida que crece x.) De manera similar, si tanto β_1 como β_2 son negativos, el mayor valor esperado de y se encuentra en $x = 0$, y aumentos en x tienen un efecto negativo sobre y, y la magnitud del efecto aumenta a medida que x es más grande.

Hay muchas otras posibilidades del empleo de funciones cuadráticas junto con logaritmos. Por ejemplo, una extensión de (6.14) que permite una elasticidad no constante entre *price* y *nox* es

$$\begin{aligned}\log(price) &= \beta_0 + \beta_1\log(nox) + \beta_2[\log(nox)]^2 \\ &+ \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u.\end{aligned} \tag{6.15}$$

Si $\beta_2 = 0$, entonces β_1 es la elasticidad de *price* respecto a *nox*. De otro modo, esta elasticidad depende del nivel de *nox*. Para ver esto, pueden combinarse los argumentos de los efectos parciales en los modelos cuadráticos y logarítmico para mostrar que

$$\%\Delta price \approx [\beta_1 + 2\beta_2\log(nox)]\%\Delta nox; \tag{6.16}$$

por tanto, la elasticidad de *price* respecto a *nox* es $\beta_1 + 2\beta_2\log(nox)$, de manera que dependen de $\log(nox)$.

Por último, en los modelos de regresión pueden incluirse otros términos polinomiales. Desde luego, el cuadrático es el que se encuentra con más frecuencia, pero de vez en cuando hay un término cúbico o incluso uno cuártico. Una forma funcional que suele ser razonable para una función de costo total es

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u.$$

Estimar un modelo así no ofrece complicaciones. Interpretar los parámetros es más complicado (aunque sencillo empleando el cálculo); aquí estos modelos no se estudiarán más.

Modelos con términos de interacción

Algunas veces es natural que el efecto parcial, la elasticidad o la semielasticidad de la variable dependiente respecto a una variable explicativa dependa de la magnitud de *otra* variable explicativa. Por ejemplo, en el modelo

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u,$$

el efecto parcial de *bdrms* (cantidad de recámaras) sobre *price* (precio) (manteniendo constantes todas las demás variables) es

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft. \tag{6.17}$$

Si $\beta_3 > 0$, entonces (6.17) implica que, en casas más grandes, una recámara más produce un aumento mayor en el precio. En otras palabras, existe un **efecto de interacción** entre la superficie en pies cuadrados y la cantidad de recámaras. Para sintetizar el efecto de *bdrms* sobre *price*, es necesario evaluar (6.17) en valores útiles de *sqrft*, como por ejemplo en el valor de la media, o en los cuartiles inferior o superior en la muestra. Si β_3 es o no cero es sencillo de probar.

Cuando se incluye un término de interacción puede ser complicado interpretar los parámetros de las variables originales. Por ejemplo, la ecuación anterior para el precio de las casas, (6.17), indica que β_2 es el efecto de *bdrms* sobre *price* para una casa con ¡cero pies cuadrados! Este efecto, por supuesto, no tiene mucho interés. En cambio, se debe tener cuidado de elegir valores útiles de *sqrft*, valores tales como la media o la mediana muestrales, en la versión estimada de la ecuación (6.17).

Con frecuencia es útil volver a parametrizar un modelo de manera que los coeficientes de las variables originales tengan un significado útil. Considere un modelo que tenga dos variables explicativas y una interacción:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

Como se acaba de decir, β_2 es el efecto parcial de x_2 sobre y cuando $x_1 = 0$. Con frecuencia esto no tiene ningún interés. En cambio, se puede parametrizar de nuevo el modelo y obtener

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u,$$

donde μ_1 es la media poblacional de x_1 y μ_2 es la media poblacional de x_2. Ahora, puede verse con facilidad que el coeficiente de x_2, δ_2, es el efecto parcial de x_2 sobre y en la media de x_1. (Multiplicando en la segunda ecuación la interacción y comparando los coeficientes se puede demostrar fácilmente que $\delta_2 = \beta_2 + \beta_3 \mu_1$. El parámetro δ_1 tiene una interpretación similar.) Por tanto, si se sustrae la media de las variables —en la práctica, esta es la media muestral— antes de crear el término de interacción, los coeficientes de las variables originales tienen una interpretación útil. Además, se obtienen de inmediato los errores estándar de los efectos parciales en los valores medios. Nada impide sustituir μ_1 o μ_2 por otros valores de las variables explicativas que puedan ser de interés. El ejemplo siguiente muestra cómo se pueden usar los términos de interacción.

Ejemplo 6.3

[Efectos de la asistencia a clases sobre el desempeño en el examen final]

Un modelo para explicar el resultado estandarizado de un examen final (*stndfnl*) en términos del porcentaje de asistencia a clases, el anterior promedio general de calificaciones y la puntuación en el ACT (examen de admisión a la universidad) es

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u. \quad \textbf{6.18}$$

(Se emplea la puntuación estandarizada obtenida en el examen por las razones vistas en la sección 6.1: facilita interpretar el desempeño de un estudiante en relación con el resto del grupo.) Además de los términos cuadráticos de *priGPA* y *ACT*, este modelo contiene un término de interacción entre *priGPA* y la tasa de asistencia a clases. La idea es que la asistencia a clases puede tener efectos diferentes en estudiantes con distinto desempeño en el pasado, medido mediante el *priGPA*. Lo que interesa es el efecto de la asistencia a clases sobre la puntuación en el examen final: $\Delta stndfnl/\Delta atndrte = \beta_1 + \beta_6 priGPA$.

Empleando las 680 observaciones del archivo ATTEND.RAW, de los estudiantes de una materia en principios de microeconomía, la ecuación estimada es

$$\begin{aligned}
\widehat{stndfnl} = {} & \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\, atndrte - \underset{(.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT \\
& + \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA \cdot atndrte \\
& n = 680,\ R^2 = .229,\ \bar{R}^2 = .222.
\end{aligned} \quad \textbf{6.19}$$

Esta ecuación debe interpretarse con extremo cuidado. Si sólo se observa el coeficiente de *atndrte* se concluirá, de manera incorrecta, que la asistencia a clases tiene efecto *negativo* sobre la puntuación obtenida en el examen final. Pero se supone que este coeficiente mide el efecto cuando $priGPA = 0$, lo cual no tiene ningún interés (el *priGPA* menor de la muestra es .86). También se debe tener cuidado de no considerar por separado las estimaciones de β_1 y β_6 y concluir que, como cada uno de los estadísticos t es insignificante, no se puede rechazar H_0: $\beta_1 = 0$, $\beta_6 = 0$. En realidad, el valor-p de la prueba F de esta hipótesis conjunta es .014, de manera que con toda seguridad se rechaza H_0 al nivel de 5%. Este es un buen ejemplo de que considerar por separado estadísticos t al probar una hipótesis conjunta puede conducir a conclusiones equivocadas.

¿Cómo debe estimarse el efecto parcial de *atndrte* sobre *stndfnl*? Para obtener el efecto parcial, en la ecuación deben sustituirse valores útiles de *priGPA*. En la muestra, el valor medio de *priGPA* es 2.59, de manera que en la media de *priGPA*, el efecto de *atndrte* sobre *stndfnl* es $-.0067 + .0056(2.59) \approx .0078$. ¿Qué significa esto? Como *atndrte* se mide como un porcentaje, esto significa que si *atndrte* aumenta 10 puntos porcentuales $\widehat{stndfnl}$ aumentará .078 desviaciones estándar a partir de la puntuación media en el examen final.

¿Cómo puede decirse si la estimación .0078 es estadísticamente distinta de cero? Se necesita volver a correr la regresión, sustituyendo $priGPA \cdot atndrte$ por $(priGPA - 2.59) \cdot atndrte$. Esto da, como nuevo coeficiente de *atndrte*, el efecto estimado cuando $priGPA = 2.59$, junto con su error estándar; en la regresión no cambia nada más. (En la sección 4.4 se describe este mecanismo.) Corriendo esta regresión se obtiene que el error estándar de $\hat{\beta}_1 + \hat{\beta}_6(2.59) = .0078$ es .0026, con lo que $t = .0078/.0026 = 3$. Por tanto, en el *priGPA*, promedio, se concluye que la asistencia a clases tiene un efecto positivo, estadísticamente significativo sobre la puntuación en el examen final.

Pregunta 6.3

Si a la ecuación (6.18) se le agrega el término $\beta_7 ACT \cdot atndrte$, ¿cuál será el efecto parcial de *atndrte* sobre *stndfnl*?

Cuando se trata de determinar el efecto de *priGPA* sobre *stndfnl* las cosas se vuelven un poco más complicadas debido al término cuadrático $priGPA^2$. Para determinar el efecto en el valor medio de *priGPA* y en la tasa media de asistencia, 82, se sustituye $priGPA^2$ por $(priGPA - 2.59)^2$ y $priGPA \cdot atndrte$ por $priGPA \cdot (atndrte - 82)$. El coeficiente de *priGPA* será el efecto parcial en los valores medios y se tendrá su error estándar. (Vea el ejercicio para computadora C6.7.)

6.3 Más sobre bondad de ajuste y selección de los regresores

Hasta ahora, al evaluar los modelos de regresión no se ha dado mucha atención a la magnitud de R^2 esto se debe a que los estudiantes principiantes tienden a darle demasiada importancia a *R*-cuadrada. Como se verá en breve, elegir un conjunto de variables explicativas con base en la magnitud de la *R*-cuadrada puede conducir a modelos sin sentido. En el capítulo 10 se verá que las *R*-cuadradas obtenidas de regresiones de series de tiempo pueden ser artificialmente altas y, como resultado, llevar a conclusiones equivocadas.

En los supuestos del modelo lineal clásico no hay nada que requiera que R^2 sea mayor que algún valor particular; R^2 es simplemente una buena estimación de cuánto de la variación de y es explicada por $x_1, x_2, \ldots, x_k$ en la población. Aquí se han visto varias regresiones que han tenido *R*-cuadradas bastante pequeñas. Aunque esto significa que hay diversos factores que afectan a y, que no han sido tomados en cuenta, esto no significa que los factores en u estén correlacionados con las variables independientes. El supuesto RLM.4 de media condicional cero es lo que determina que se obtengan estimadores insesgados de los efectos *ceteris paribus* de las variables independientes, y la magnitud de la *R*-cuadrada no tiene relación directa con esto.

Que *R*-cuadrada sea pequeña implica que la varianza del error es relativamente grande en relación con la varianza de y, lo cual significa que es posible que sea difícil estimar con precisión las β_j. Pero recuerde que en la sección 3.4 se vio que una varianza grande del error puede ser contrarrestada con un tamaño de muestra grande: si se tienen suficientes datos, es posible que se puedan estimar con precisión los efectos parciales, aun cuando no se hayan controlado muchos de los factores no observados. El que se puedan o no obtener estimaciones suficientemente precisas depende de la aplicación. Por ejemplo, suponga que a algunos estudiantes nuevos de una universidad grande se les otorga financiamiento, de forma aleatoria, para comprar equipos de cómputo. Si los montos del financiamiento en realidad son determinados de forma aleatoria, se puede determinar el efecto *ceteris paribus* del monto del financiamiento sobre el subsecuente promedio general de calificaciones en la universidad empleando el análisis de regresión simple. [Debido a la asignación aleatoria, todos los demás factores que afectan el promedio general de calificaciones (GPA) no estarán correlacionados con el monto del financiamiento.] Es probable que el monto del financiamiento explique poco de la variación del GPA, de manera que la *R*-cuadrada de esta regresión tal vez será muy pequeña. Sin embargo, si se obtiene una muestra de tamaño grande, será posible obtener un estimador razonablemente preciso del efecto del financiamiento.

Otro buen ejemplo en el que el poder explicativo no tiene nada que ver con la estimación insesgada de las β_j es el que se obtiene mediante el análisis de la base de datos del archivo APPLE.RAW. A diferencia de otras bases de datos empleadas aquí, las variables explicativas clave del archivo APPLE.RAW se obtuvieron de manera experimental —es decir, sin atender a otros factores que pudieran afectar a la variable dependiente—. La variable que se desea explicar, *ecolbs*, es la cantidad (hipotética) de libras de manzanas “no dañinas para la ecología” (“ecoetiquetadas”) que demanda una familia. A cada familia (en realidad, a cada cabeza de familia) se le presentó una descripción de las manzanas ecoetiquetadas, junto con los precios de las manzanas normales (o regulares) (*regprc*) y los precios de las manzanas ecoetiquetadas hipotéticas (*ecoprc*). Dado que a las familias se les asignaron los pares de precios de forma aleatoria, éstos no están relacionados con otros factores observados (tales como el ingreso familiar) ni con factores no observados (tales como el interés por un ambiente limpio). Por tanto, la regresión de *ecolbs* sobre *ecoprc*, *regprc* (a través de las muestras generadas de esta manera) produce estimadores insesgados de los efectos del precio. Sin embargo, la R-cuadrada de esta regresión es sólo .0364: las variables precio sólo explican cerca de 3.6% de la variación total de *ecolbs*, de manera que aquí hay un caso en el que se explica muy poco de la variación de y aun cuando se está en la rara situación en que se sabe que los datos han sido generados de manera que se puede obtener una estimación insesgada de las β_j. (Dicho sea de paso, la adición de características observadas en la familia tiene un efecto muy pequeño sobre el poder explicativo. Vea el ejercicio para computadora C6.11.)

Recuerde, sin embargo, que cuando a una ecuación se le agregan variables la *variación* relativa de R-cuadrada es muy útil: en (4.41) el estadístico F para probar la significancia conjunta depende de manera crucial de la diferencia entre las R-cuadradas del modelo no restringido y el restringido.

R-cuadrada ajustada

La mayoría de los paquetes para regresión dan, junto con la R-cuadrada, un estadístico llamado ***R*-cuadrada ajustada**. Dado que esta última se reporta en muchas de las aplicaciones y tiene algunas características útiles, se verá en esta subsección.

Para ver cómo puede ajustarse la R-cuadrada usual, es útil escribirla de la manera siguiente:

$$R^2 = 1 - (\mathrm{SRC}/n)/(\mathrm{STC}/n), \tag{6.20}$$

donde SRC es la suma de los residuales cuadrados y STC es la suma total de cuadrados; comparando con la ecuación (3.28), lo único que se ha hecho es dividir entre n tanto la SRC como la STC. Esta expresión muestral lo que en realidad estima la R^2. Defínase σ_y^2 como la varianza poblacional de y y sea σ_u^2 la varianza poblacional del término del error, u. (Hasta ahora, σ^2 se ha empleado para denotar σ_u^2, pero aquí es útil ser más específico.) La ***R*-cuadrada poblacional** se define como $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$; que es la proporción de la variación de y en la población explicada por las variables independientes. Esto es lo que se supone que estima R^2.

R^2 estima σ_u^2 mediante SRC/n, que se sabe es sesgado, de manera que, ¿por qué no sustituir SRC/n por SRC/$(n - k - 1)$? Además, también se puede usar STC/$(n - 1)$ en lugar de STC/n, ya que el primero es un estimador insesgado de σ_y^2. Empleando estos estimadores, se llega a la R-cuadrada ajustada:

$$\begin{aligned} \bar{R}^2 &= 1 - [\mathrm{SRC}/(n - k - 1)]/[\mathrm{STC}/(n - 1)] \\ &= 1 - \hat{\sigma}^2/[\mathrm{STC}/(n - 1)], \end{aligned} \tag{6.21}$$

ya que $\hat{\sigma}^2 = \mathrm{SRC}/(n - k - 1)$. Debido a la notación empleada para denotar la R-cuadrada ajustada, también se le suele llamar *R-barra cuadrada*.

A la R-cuadrada ajustada se le llama también *R-cuadrada corregida*, pero este nombre no resulta muy adecuado porque implicaría que $\bar{R}^2$ de alguna manera es un mejor estimador de la R-cuadrada poblacional que R^2. Por desgracia, en general $\bar{R}^2$ *no* se sabe que sea un mejor estimador. Se podría pensar que $\bar{R}^2$ corrigiera el sesgo de R^2 al estimar la R-cuadrada poblacional, ρ^2, pero no lo hace: el cociente de dos estimadores insesgados no es un estimador insesgado.

La característica más atractiva de $\bar{R}^2$ es que impone una sanción a la adición de más variables independientes a un modelo. Se sabe que R^2 nunca disminuye cuando se agrega una variable independiente a la ecuación de regresión: esto se debe a que SRC nunca aumenta (y en general disminuye) a medida que se agregan más variables independientes. Pero la fórmula de $\bar{R}^2$ muestra que ésta depende de manera explícita de k, la cantidad de variables independientes. Si se agrega una variable independiente a la regresión, SRC disminuye, pero lo mismo ocurre con los *gl* en la regresión, $n - k - 1$. SRC/$(n - k - 1)$ puede aumentar o disminuir cuando se agrega una nueva variable independiente a una regresión.

Un hecho algebraico interesante es el siguiente: si se agrega una variable independiente a una ecuación de regresión, $\bar{R}^2$ aumenta si, y sólo si, el valor absoluto del estadístico t de la nueva variable es mayor que uno. (Una extensión de esto es que cuando a una regresión se agrega un grupo de variables $\bar{R}^2$ aumenta si, y sólo si, el estadístico F de la significancia conjunta de las nuevas variables es mayor que la unidad.) Por tanto, de inmediato se ve que emplear $\bar{R}^2$ para decidir si una determinada variable independiente (o un conjunto de variables) pertenece a un modelo proporciona una respuesta distinta a la de las pruebas estándar t o F (debido a que a los niveles de significancia tradicionales un estadístico t o F igual a uno no es estadísticamente significativo).

Algunas veces es útil tener una fórmula de $\bar{R}^2$ en términos de R^2. Mediante álgebra sencilla se obtiene

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1). \qquad \boxed{\textbf{6.22}}$$

Por ejemplo, si $R^2 = .30$, $n = 51$ y $k = 10$, entonces $\bar{R}^2 = 1 - .70(50)/40 = .125$. Por tanto, cuando n es pequeña y k es grande, $\bar{R}^2$ puede ser sustancialmente menor a R^2. En realidad, si la R-cuadrada usual es pequeña y $n - k - 1$ es pequeño, ¡$\bar{R}^2$ puede ser negativa! Por ejemplo, se puede sustituir $R^2 = .10$, $n = 51$ y $k = 10$ para comprobar que $\bar{R}^2 = -.125$. Una $\bar{R}^2$ negativa indica un ajuste muy pobre del modelo en relación con los grados de libertad.

En las regresiones, algunas veces se reporta la R-cuadrada ajustada junto con la R-cuadrada usual, y algunas veces se reporta $\bar{R}^2$ en lugar de R^2. Es importante recordar que la que aparece en el estadístico F en (4.41) es R^2 y no $\bar{R}^2$. La misma fórmula con $\bar{R}_r^2$ o $\bar{R}_{ur}^2$ *no* es válida.

Uso de la *R*-cuadrada ajustada para elegir entre modelos no anidados

En la sección 4.5 se vio cómo calcular un estadístico F para probar la significancia conjunta de un grupo de variables; esto permite decidir, a un determinado nivel de significancia, si por lo menos una de las variables del grupo afecta a la variable dependiente. Esta prueba no permite decidir *cuál* de las variables es la que tiene algún efecto. En algunos casos, se desea elegir un modelo que no tenga variables independientes redundantes, y para esto puede servir la R-cuadrada ajustada.

En el ejemplo del sueldo en la liga mayor de béisbol en la sección 4.5 se vio que ni *hrunsyr* ni *rbisyr* eran significativas individualmente. Estas dos variables están muy correlacionadas, de manera que habrá que elegir entre los modelos

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$$

y

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

Estas dos ecuaciones son **modelos no anidados**, ya que ninguna de ellas es un caso especial de la otra. El estadístico F visto en el capítulo 4 sólo permite probar modelos *anidados*: un modelo (el restringido) es un caso especial del otro (el no restringido). Como ejemplos de modelos restringidos y no restringidos vea las ecuaciones (4.32) y (4.28). Una posibilidad es crear un modelo compuesto que contenga *todas* las variables explicativas de los modelos originales y después probar cada modelo contra el general empleando la prueba F. El problema de este proceso es que puede que ambos se rechacen o que ninguno se rechace (como ocurre en el ejemplo del sueldo en la liga mayor de béisbol en la sección 4.5). De manera que este proceso no siempre proporciona una manera de distinguir entre modelos con regresores no anidados.

En la regresión del sueldo de los jugadores de béisbol, $\bar{R}^2$ en la regresión que contiene *hrunsyr* es .6211 y $\bar{R}^2$ en la regresión que contiene *rbisyr* es .6226. Por tanto, con base en la R-cuadrada ajustada, hay una muy ligera preferencia por el modelo que tiene *rbisyr*. Pero, prácticamente la diferencia es muy pequeña y es posible que se obtengan respuestas diferentes controlando algunas de las variables del ejercicio para computadora C4.5. (Como los dos modelos no anidados contienen cinco parámetros, puede emplearse la R-cuadrada usual para llegar a la misma conclusión.)

Comparar $\bar{R}^2$ para elegir entre diversos conjuntos no anidados de variables independientes puede ser valioso cuando las variables representan formas funcionales diferentes. Considere dos modelos que relacionan la intensidad de la Investigación y Desarrollo (I&D) con las ventas de una empresa:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u. \tag{6.23}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \tag{6.24}$$

El primer modelo capta el rendimiento decreciente mediante la inclusión de *sales* en forma logarítmica; el segundo modelo hace lo mismo empleando un término cuadrático y, por tanto, contiene un parámetro más que el primero.

Cuando se estima la ecuación (6.23) empleando las 32 observaciones de empresas químicas del archivo RDCHEM.RAW, R^2 es .061 y para la ecuación (6.24) R^2 es .148. Por tanto, parece que la ecuación cuadrática tiene un ajuste mucho mejor. Pero la comparación de las R-cuadradas usuales no es justa para el primer modelo porque éste contiene un parámetro menos que la ecuación (6.24). Es decir, (6.23) es un modelo más parsimonioso que (6.24).

Permaneciendo todo lo demás igual, los modelos más sencillos son mejores. Dado que la R-cuadrada usual no sanciona modelos más complicados, es mejor emplear $\bar{R}^2$. Para (6.23) $\bar{R}^2$ es .030, mientras que para (6.24) $\bar{R}^2$ es .090. Por tanto, aun después del ajuste para la diferencia en los grados de libertad, el modelo cuadrático gana. Tal modelo también se prefiere cuando a cada regresión se agregan márgenes de ganancia.

Existe una limitación importante al usar $\bar{R}^2$ para elegir entre modelos no anidados: no se puede usar para escoger entre distintas formas funcionales de la variable dependiente. Esto es lamentable, porque con frecuencia se desea decidir entre si usar y o $\log(y)$ (o alguna otra transformación) como variable dependiente con base en la bondad de ajuste. Pero para este propósito no se pueden usar ni R^2 ni $\bar{R}^2$. La razón es sencilla: estas R-cuadradas miden la proporción explicada de la variación total en la variable dependiente que se emplee en la regresión, y funciones diferentes de la variable dependiente tendrán cantidades distintas de variación a explicar. Por ejemplo, la variación total de y y de $\log(y)$ no son las mismas, y con frecuencia son muy distintas. Comparar las R-cuadradas ajustadas de regresiones que tienen estas distintas formas de las variables dependientes no dice nada acerca de qué modelo se ajusta mejor; los modelos se ajustan a dos variables dependientes distintas.

Pregunta 6.4

Explique por qué es lo mismo elegir un modelo maximizando $\bar{R}^2$ que minimizando $\hat{\sigma}$ (el error estándar de la regresión).

Ejemplo 6.4

[Compensación de los directores generales y desempeño de una empresa]

Considere dos modelos estimados que relacionan la compensación de los directores generales (CEO) con el desempeño de la empresa:

$$\begin{aligned}\widehat{salary} &= 830.63 + .0163\ sales + 19.63\ roe \\ &\quad\ (223.90)\quad (.0089)\qquad\quad (11.08) \\ n &= 209,\ R^2 = .029,\ \bar{R}^2 = .020\end{aligned} \tag{6.25}$$

y

$$\begin{aligned}\widehat{lsalary} &= 4.36 + .275\ lsales + .0179\ roe \\ &\quad\ (0.29)\quad (.033)\qquad\quad (.0040) \\ n &= 209,\ R^2 = .282,\ \bar{R}^2 = .275,\end{aligned} \tag{6.26}$$

donde *roe* es la rentabilidad de la inversión vista en el capítulo 2. Para simplificar, *lsalary* y *lsales* denotan los logaritmos naturales de *salary* y de *sales*. Se sabe, ya, cómo interpretar estas distintas ecuaciones estimadas. Pero, ¿puede decirse que un modelo se ajuste mejor que el otro?

La R-cuadrada de la ecuación (6.25) indica que *sales* y *roe* explican sólo cerca de 2.9% de la variación del sueldo de los CEO en la muestra. Tanto *sales* como *roe* tienen una significación estadística marginal.

La ecuación (6.26) indica que el log(*sales*) y *roe* explican aproximadamente 28.2% de la variación de log(*salary*). En términos de bondad de ajuste, esta R-cuadrada mayor parecería implicar que el modelo (6.26) es mucho mejor, pero este no es necesariamente el caso. La suma total de cuadrados de *salary* en la muestra es 391,732,982, mientras que la suma total de cuadrados de log(*salary*) es sólo de 66.72. Por tanto, en log(*salary*) hay mucho menos variación que debe ser explicada.

En este punto pueden emplearse otras características distintas de R^2 y $\bar{R}^2$ para decidir entre estos modelos. Por ejemplo, en (6.26) log(*sales*) y *roe* son estadísticamente mucho más significativas que *sales* y *roe* en (6.25), y los coeficientes en (6.26) son quizá de mayor interés. Sin embargo, para estar seguro, se requerirá hacer una comparación válida de la bondad de ajuste.

En la sección 6.4, se presentará una medida de la bondad de ajuste que permite comparar modelos en los que *y* aparece tanto de forma lineal como de forma logarítmica.

Control de demasiados factores en un análisis de regresión

En muchos de los ejemplos vistos y seguramente en el análisis sobre el sesgo de variables omitidas en el capítulo 3, la preocupación ha sido omitir, de un modelo, factores importantes que puedan estar correlacionados con las variables independientes. También puede ocurrir que en un análisis de regresión se controlen *demasiadas* variables.

Si se le da demasiada importancia a la bondad de ajuste, puede ocurrir que en un modelo de regresión se controlen factores que no deberían ser controlados. Para evitar este error, hay que recordar la interpretación *ceteris paribus* de los modelos de regresión múltiple.

Para ilustrar esto, suponga que se hace un estudio para evaluar el impacto de los impuestos estatales a la cerveza sobre los accidentes de tránsito fatales. La idea es que con impuestos más altos a la cerveza se reducirá el consumo de alcohol y, por consiguiente, la posibilidad de conducir bajo los efectos del alcohol, dando como resultado una reducción de los accidentes de tránsito fatales. Para medir el efecto *ceteris paribus* de los impuestos sobre los accidentes fatales, los *accidentes fatales* (*fatalities*) pueden modelarse como función de diversos factores, entre los que se encuentra el *impuesto* (*tax*) a la cerveza:

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 miles + \beta_3 percmale + \beta_4 perc16\_21 + \ldots,$$

donde

$miles$ = cantidad total de millas recorridas conduciendo.
$percmale$ = porcentaje de hombres en la población de un estado.
$perc16\_21$ = porcentaje de personas entre 16 y 21 años de edad en la población.

Observe que no se ha incluido una variable que mida el consumo de cerveza *per cápita*. ¿Se está cometiendo un error de omisión de variables? La respuesta es no. Si en esta ecuación se controla el consumo de cerveza, entonces ¿cómo afectarían los impuestos a la cerveza los accidentes de tránsito fatales? En la ecuación

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \ldots,$$

β_1 mide la diferencia en los accidentes fatales debido a un aumento de un punto porcentual en los *impuestos*, manteniendo constante el consumo de cerveza (*beercons*). No es fácil entender por qué esto podría ser de interés. No se deben controlar las diferencias en *beercons* entre los estados, a menos que se desee probar algún tipo de efecto indirecto de los impuestos a la cerveza. Otros factores, tales como distribución del género y de la edad sí deben ser controlados.

Para ver otro ejemplo, suponga que, en un país en desarrollo, se desea estimar el efecto del uso de pesticidas por los agricultores sobre el gasto en salud familiar. Además de las cantidades de pesticidas usadas, ¿debe incluirse, como variable explicativa, la cantidad de visitas al médico? No. Los gastos en salud comprenden las visitas al médico y lo que se desea es captar todos los efectos del uso de pesticidas sobre los gastos en salud. Si se incluye, como variable explicativa, la cantidad de visitas al médico, entonces sólo se estarán midiendo los efectos del uso de los pesticidas sobre gastos en la salud distintos a las visitas al médico. Tiene más sentido usar la cantidad de visitas al médico como variable dependiente en otra regresión contra las cantidades de pesticidas.

Los anteriores son ejemplos de lo que podría llamarse un **sobrecontrol** de factores en la regresión múltiple. Esto suele ocurrir como resultado de la preocupación por los sesgos potenciales que pueden surgir al dejar fuera variables explicativas importantes. Pero es importante recordar la naturaleza *ceteris paribus* de la regresión múltiple. En algunas situaciones no tiene caso mantener ciertos factores fijos debido precisamente a que variarán cuando cambie una variable de política.

Por desgracia, el asunto de si controlar o no ciertos factores no siempre es claro. Por ejemplo, Betts (1995) estudió el efecto de la calidad del bachillerato sobre los ingresos subsecuentes. Él indica que, si una mejor calidad de la escuela da como resultado más educación, entonces controlar en la regresión la educación junto con medidas de la calidad subestima el rendimiento de la calidad. Betts realiza el análisis con y sin años de educación en la ecuación para obtener un intervalo de los efectos estimados de la calidad de la educación.

Para ver de manera explícita cómo enfocarse en *R*-cuadradas altas puede conducir a problemas, considere el ejemplo del precio de la vivienda visto en la sección 4.5 que ilustra la prueba de hipótesis múltiples. En ese caso se deseaba probar lo razonable de los avalúos de los precios de las viviendas. Se regresó log(*price*) sobre log(*assess*), log(*lotsize*), log(*sqrft*) y *bdrms* y se probó si las últimas tres variables tenían coeficientes poblacionales iguales a cero mientras que log(*assess*) tenía coeficiente igual a uno. Pero, ¿qué pasa si se modifica el objetivo del análisis y se estima un *modelo de precios hedónicos*, que permita obtener los valores marginales de distintos atributos de las viviendas? ¿Se debe incluir log(*assess*) en la ecuación? La *R*-cuadrada ajustada de la regresión con log(*assess*) es .762, mientras que la *R*-cuadrada ajustada sin él es .630. Basándose sólo en la bondad de ajuste debería incluirse log(*assess*). Pero esto es incorrecto si lo que se quiere determinar son los efectos del tamaño del terreno, de la superficie de la vivienda en pies cuadrados y de la cantidad de recámaras sobre el valor de la vivienda. Incluir log(*assess*) en la ecuación equivale a mantener fija una medida del valor y preguntar después cuánto modificará una recámara más otra medida del valor. Esto no tiene sentido para la evaluación de los atributos de una vivienda.

Si se recuerda que modelos distintos tienen propósitos diferentes y se atiende a la interpretación *ceteris paribus* de la regresión, no se incluirán factores inadecuados en un modelo de regresión.

Adición de regresores para reducir la varianza del error

Se acaban de ver algunos ejemplos en los que ciertas variables independientes no deben ser incluidas en un modelo de regresión, aun cuando estas variables estén correlacionadas con la variable independiente. De acuerdo con lo visto en el capítulo 3, se sabe que agregar una variable independiente a la regresión puede exacerbar el problema de multicolinealidad. Por otro lado, como se está sacando algo del término de error, adicionar una variable reduce, por lo general, la varianza del error. Por lo común no se puede saber cuál será el efecto que domine.

Sin embargo, hay un caso que es claro: siempre deben incluirse las variables independientes que afecten a *y* y que *no estén correlacionadas* con todas las variables independientes de interés. ¿Por qué? Porque agregar esas variables no induce multicolinealidad en la población (y por tanto la multicolinealidad en la muestra será despreciable), pero sí reduce la varianza del error. En muestras grandes, los errores estándar de los estimadores de MCO se reducirán.

Como ejemplo, considérese la estimación de la demanda individual de cerveza como función del precio promedio de la cerveza en un condado o municipio. Será razonable suponer que las características individuales no están correlacionadas con el nivel de precios en el condado y de esta manera una regresión simple del consumo de cerveza sobre el precio en el condado será suficiente para estimar el efecto del precio sobre la demanda individual. Pero se puede obtener un estimador más preciso de la elasticidad precio de la demanda de cerveza mediante la inclusión de características individuales, tales como edad y cantidad de educación. Si estos factores que afectan la demanda no están correlacionados con el precio, entonces el error estándar del coeficiente del precio será menor, por lo menos cuando se tengan muestras grandes.

Para ver otro ejemplo, considere el financiamiento para equipo de cómputo mencionado al inicio de la sección 6.3. Si, además de la variable financiamiento, se controlan otros factores que puedan explicar el promedio general de calificaciones (GPA), tal vez se podrá obtener un estimado más preciso del efecto del financiamiento. Medidas del promedio general de calificaciones en el bachillerato y del ranking, puntuaciones de los exámenes de admisión a la universidad SAT y ACT (por sus siglas en inglés) y variables de antecedentes familiares son buenos candidatos. Dado que los montos del financiamiento son asignados de forma aleatoria, todas las demás variables de control no están correlacionadas con el monto del financiamiento; en la muestra, la multicolinealidad entre el monto del financiamiento y otras variables independientes deberá ser mínima. Pero la adición de los controles extra puede reducir de manera significativa la varianza del error, conduciendo a una estimación más precisa del efecto del financiamiento. Recuerde que aquí el problema no es insesgamiento: se agregue o no el desempeño en el bachillerato y las variables de antecedentes familiares se obtiene un estimador insesgado y consistente.

Por desgracia, en las ciencias sociales los casos en los que se tiene información adicional sobre variables explicativas que no estén correlacionadas con las variables explicativas de interés son raros. Pero vale la pena recordar que cuando existan estas variables, pueden incluirse en el modelo para reducir la varianza del error sin inducir multicolinealidad.

6.4 Predicción y análisis de residuales

En el capítulo 3 se definieron los valores predichos o ajustados de MCO, así como los residuales de MCO. Las **predicciones** son seguramente útiles, pero están sujetas a variaciones de muestreo, debido a que se obtienen empleando los estimadores de MCO. Así, en esta sección se muestra cómo obtener intervalos de confianza para una predicción a partir de la línea de regresión de MCO.

De acuerdo con lo visto en los capítulos 3 y 4, se sabe que los residuales se usan para obtener la suma de los residuales cuadrados y la *R*-cuadrada, de manera que son importantes para la bondad de ajuste y para las pruebas. Algunas veces los economistas estudian los residuales de determinadas observaciones para obtener información acerca de individuos (empresas, viviendas, etc.) de la muestra.

Intervalos de confianza para predicciones

Suponga que se tiene la ecuación estimada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{6.27}$$

Cuando se sustituyen valores específicos de las variables independientes, se obtiene una predicción para y, la cual es una estimación del *valor esperado* de y dados los valores específicos de las variables explicativas. Haciendo hincapié, sean $c_1, c_2, \ldots, c_k$ valores específicos de cada una de las k variables independientes, que pueden corresponder o no a datos reales de la muestra. El parámetro que se desea estimar es

$$\begin{aligned}\theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k \\ &= \mathrm{E}(y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k).\end{aligned} \tag{6.28}$$

El estimador de θ_0 es

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k. \tag{6.29}$$

En la práctica, esto es fácil de calcular. Pero, ¿y si se desea tener alguna medida de la incertidumbre de este valor predicho? Lo natural es construir un intervalo de confianza para θ_0, que esté centrado en $\hat{\theta}_0$.

Para obtener un intervalo de confianza para θ_0, se necesita un error estándar de $\hat{\theta}_0$. Entonces, con un número de *gl* grande, se puede construir un intervalo de confianza de 95% empleando la regla $\hat{\theta}_0 \pm 2 \cdot \mathrm{ee}(\hat{\theta}_0)$. (Como siempre, pueden emplearse los percentiles exactos de una distribución *t*.)

¿Cómo se obtiene el error estándar de $\hat{\theta}_0$? Este es el mismo problema encontrado en la sección 4.4: se necesita obtener un error estándar de una combinación lineal de los estimadores de MCO. Aquí, el problema es incluso más complicado, porque en general todos los estimadores de MCO aparecen en $\hat{\theta}_0$ (a menos que algunos de los c_j sean cero). No obstante, el mismo truco empleado en la sección 4.4 funciona aquí. Se escribe $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$ y esto se sustituye en la ecuación

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

con lo que se obtiene

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \dots + \beta_k(x_k - c_k) + u. \tag{6.30}$$

En otras palabras, a cada observación c_j se le resta x_j, y se vuelve a correr la regresión de

$$y_i \text{ sobre } (x_{i1} - c_1), \dots, (x_{ik} - c_k), i = 1, 2, \dots, n. \tag{6.31}$$

El valor predicho en (6.29) y, lo que es más importante, su error estándar, se obtienen del *intercepto* (o constante) de la regresión (6.31).

Para dar un ejemplo, se obtendrá un intervalo de confianza para la predicción de una regresión del promedio general de calificaciones (GPA) en la universidad, en la que se usa información del bachillerato.

Ejemplo 6.5

[Intervalo de confianza para el promedio general de calificaciones (GPA) en la universidad]

Empleando los datos del archivo GPA2.RAW se obtiene la ecuación siguiente para predecir el promedio general de calificaciones (GPA) en la universidad:

$$\begin{aligned}\widehat{colgpa} &= \underset{(0.075)}{1.493} + \underset{(.00007)}{.00149}\, sat - \underset{(.00056)}{.01386}\, hsperc \\ &\quad - \underset{(.01650)}{.06088}\, hsize + \underset{(.00227)}{.00546}\, hsize^2 \\ n &= 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,\end{aligned} \tag{6.32}$$

donde las estimaciones se han dado con varios dígitos para reducir el error de redondeo. ¿Cuál es el promedio general de calificaciones (GPA) en la universidad, si $sat = 1{,}200$, $hsperc = 30$ y $hsize = 5$ (que significa 500)? Esto se obtiene fácilmente sustituyendo estos valores en la ecuación (6.32): $\widehat{colgpa} = 2.70$ (redondeado a dos cifras decimales). Por desgracia, la ecuación (6.32) no puede emplearse para obtener directamente un intervalo de confianza para *colgpa* en los valores dados de las variables independientes. Una manera sencilla de obtener un intervalo de confianza es definir un nuevo conjunto de variables independientes: $sat0 = sat - 1{,}200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$ y $hsizesq0 = hsize^2 - 25$. Al regresar *colgpa* sobre estas nuevas variables independientes, se obtiene

$$\begin{aligned}\widehat{colgpa} &= \underset{(0.020)}{2.700} + \underset{(.00007)}{.00149}\, sat0 - \underset{(.00056)}{.01386}\, hsperc0 \\ &\quad - \underset{(.01650)}{.06088}\, hsize0 + \underset{(.00227)}{.00546}\, hsizesq0 \\ n &= 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560.\end{aligned}$$

La única diferencia entre esta regresión y la de (6.32) es el intercepto, que es la predicción que se busca, junto con su error estándar, .020. No es una coincidencia que los coeficientes de pendiente, sus errores estándar, la *R*-cuadrada, etc., sean las mismas que antes; esto proporciona una manera de verificar que se hicieron las transformaciones correctas. Construir un intervalo de confianza de 95% para el GPA esperado es fácil: $2.70 \pm 1.96(.020)$ es decir aproximadamente de 2.66 a 2.74. Este intervalo de confianza es bastante estrecho debido al tamaño tan grande de la muestra.

Como la varianza del estimador del intercepto es la menor posible cuando todas las variables explicativas tienen media muestral cero (vea la pregunta 2.5 para el caso de la regresión simple), se sigue, de acuerdo con la regresión (6.31), que la varianza de la predicción es la menor posible en los valores medios de las x_j. (Es decir, $c_j = \bar{x}_j$ para toda j.) Este resultado no debe extrañar, ya que cerca del centro de los datos se tiene la mayor confianza en la línea de regresión. A medida que los valores de las c_j se alejan de las $\bar{x}_j$, $\text{Var}(\hat{y})$ se vuelve cada vez mayor.

El método anterior permite colocar un intervalo de confianza alrededor de la estimación de MCO de $\text{E}(y|x_1, \ldots, x_k)$, para cualesquiera valores de las variables explicativas. En otras palabras, se obtiene un intervalo de confianza para el valor *promedio* de y en la subpoblación de un conjunto dado de variables independientes. Pero un intervalo de confianza para la persona promedio de la subpoblación no es lo mismo que un intervalo de confianza para una unidad específica (individuo, familia, empresa, etc.) de la población. Al formar un intervalo de confianza para un resultado desconocido de y, debe tomarse en cuenta otra fuente importante de variación: la varianza del error no observado, el cual mide la ignorancia de los factores no observados que afectan a y.

Sea y^0 el valor para el cual se desea construir un intervalo de confianza, al cual suele llamársele **intervalo de predicción**. Por ejemplo, y^0 puede representar, una persona o una empresa que no esté en la muestra original. Sean $x_1^0, \ldots, x_k^0$ los nuevos valores de las variables independientes, que se supone se observan, y sea u^0 el error no observado. Por tanto, se tiene

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0 + u^0. \tag{6.33}$$

Como antes, la mejor predicción para y^0 es el valor esperado de y^0 dadas las variables explicativas, el cual se estima a partir de la línea de regresión de MCO: $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$. El **error de predicción** al emplear $\hat{y}^0$ para predecir y^0 es

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}^0. \tag{6.34}$$

Ahora, $\text{E}(\hat{y}^0) = \text{E}(\hat{\beta}_0) + \text{E}(\hat{\beta}_1)x_1^0 + \text{E}(\hat{\beta}_2)x_2^0 + \ldots + \text{E}(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$, debido a que las $\hat{\beta}_j$ son insesgadas. (Como siempre, todos estos valores esperados son condicionales en los valores muestrales de las variables independientes.) Como u^0 tiene media cero, $\text{E}(\hat{e}^0) = 0$. Se ha demostrado que el error de predicción esperado es cero.

Al determinar la varianza de $\hat{e}^0$, observe que u^0 no está correlacionada con ninguna de las $\hat{\beta}_j$, porque tampoco lo está con los errores de la muestra empleada para obtener las $\hat{\beta}_j$. De acuerdo con las propiedades básicas de la covarianza (ver apéndice B), u^0 y $\hat{y}^0$ no están correlacionados. Por tanto **varianza del error de predicción** (condicional en todos los valores muestrales de las variables independientes) es la suma de las varianzas:

$$\text{Var}(\hat{e}^0) = \text{Var}(\hat{y}^0) + \text{Var}(u^0) = \text{Var}(\hat{y}^0) + \sigma^2, \tag{6.35}$$

donde $\sigma^2 = \text{Var}(u^0)$ es la varianza del error. En $\hat{e}^0$ hay dos fuentes de variación. La primera es el error de muestreo en $\hat{y}^0$, que surge debido a que se han estimado los β_j. Como cada $\hat{\beta}_j$ tiene una varianza proporcional a $1/n$, donde n es el tamaño de la muestra, $\text{Var}(\hat{y}^0)$ es proporcional a $1/n$. Esto significa que, con muestras grandes, $\text{Var}(\hat{y}^0)$ puede ser muy pequeña. En cambio, σ^2 es la varianza del error en la población: ésta no cambia con el tamaño de la muestra. En muchos ejemplos, σ^2 será el término dominante en (6.35).

Bajo los supuestos del modelo lineal clásico, las $\hat{\beta}_j$ y u^0 están distribuidas normalmente y entonces $\hat{e}^0$ tambien está distribuida normalmente (condicional a todos los valores muestrales de las variables explicativas). Antes, se describió cómo obtener un estimador insesgado de Var($\hat{y}^0$), y en el capítulo 3 se obtuvo el estimador insesgado de σ^2. Empleando estos estimadores, se puede definir el error estándar de $\hat{e}^0$ como

$$\mathrm{ee}(\hat{e}^0) = \{[\mathrm{ee}(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}. \tag{6.36}$$

Empleando el mismo razonamiento para los estadísticos t de los $\hat{\beta}_j$, $\hat{e}^0/\mathrm{ee}(\hat{e}^0)$ tiene una distribución t con $n - (k + 1)$ grados de libertad. Por tanto,

$$\mathrm{P}[-t_{.025} \leq \hat{e}^0/\mathrm{ee}(\hat{e}^0) \leq t_{.025}] = .95,$$

donde $t_{.025}$ es el percentil 97.5 de la distribución t_{n-k-1}. Si $n - k - 1$ es grande, recuerde que $t_{.025} \approx 1.96$. Al sustituir en $\hat{e}^0 = y^0 - \hat{y}^0$ y reordenar se obtiene un intervalo de predicción de 95% para y^0:

$$\hat{y}^0 \pm t_{.025}\cdot \mathrm{ee}(\hat{e}^0); \tag{6.37}$$

como de costumbre, excepto cuando el número de *gl* sea pequeño, una buena regla práctica es $\hat{y}^0 \pm 2\mathrm{ee}(\hat{e}^0)$. Este intervalo es más amplio que el de $\hat{y}^0$ debido a la presencia de $\hat{\sigma}^2$ en (6.36); este intervalo suele ser mucho más amplio para reflejar los factores en u^0 que no se han controlado.

Ejemplo 6.6

[Intervalo de confianza para un GPA futuro]

Suponga que se desea un IC de 95% para el GPA futuro de un estudiante de bachillerato para el que *sat* = 1,200, *hsperc* = 30 y *hsize* = 5. En el ejemplo 6.5, se obtuvo un intervalo de confianza de 95% para el *promedio* de GPA de todos los estudiantes con las características particulares *sat* = 1,200, *hsperc* = 30 y *hsize* = 5. Ahora se quiere un intervalo de confianza de 95% para un estudiante *específico* con estas características. El intervalo de predicción de 95% debe tomar en cuenta la variación en el individuo, características no observadas que afectan el desempeño universitario. Ya se tiene todo lo que se necesita para obtener un IC para *colgpa*. ee($\hat{y}^0$) = .020 y $\hat{\sigma}$ = .560 y, de esta manera, de acuerdo con (6.36), $\mathrm{ee}(\hat{e}^0) = [(.020)^2 + (.560)^2]^{1/2} \approx .560$. Observe lo pequeño que es ee($\hat{y}^0$) con relación a $\hat{\sigma}$: prácticamente toda la variación en $\hat{e}^0$ proviene de la variación en u^0. El IC de 95% es 2.70 ± 1.96(.560) es decir, aproximadamente de 1.60 a 3.80. Este es un intervalo de confianza amplio y muestra que, con base en los factores incluidos en la regresión, no es posible señalar con exactitud el promedio general de calificaciones futuras de un individuo. [En cierto sentido, esta es una buena noticia, pues significa que el desempeño en el bachillerato y en el examen de admisión (SAT) no predeterminan el desempeño en la universidad.] Evidentemente, las características no observadas varían de manera amplia entre los individuos que tienen una misma puntuación observada en el SAT y un mismo promedio observado en el bachillerato.

Análisis de residuales

Algunas veces es útil examinar las observaciones individuales para ver si el verdadero valor de la variable dependiente es superior o inferior al valor predicho; es decir, examinan los residuales de las observaciones individuales. A este proceso se le llama **análisis de residuales**.

Los economistas examinan los residuales de una regresión con objeto de ayudar en la compra de una vivienda. El ejemplo siguiente sobre precios de la vivienda ilustra el análisis de residuales. El precio de una vivienda está relacionado con diversas características observables de la vivienda. Se puede hacer una lista con todas las características que se consideran importantes, tales como tamaño, cantidad de recámaras y de baños, etc. Se puede emplear una muestra de viviendas para estimar la relación entre el precio y los atributos, con lo que se obtiene un valor predicho y un valor real de cada vivienda. Después, pueden calcularse los residuales, $\hat{u}_i = y_i - \hat{y}_i$. La vivienda con el residual más negativo es, al menos con base en los factores controlados, la más subvalorada con relación a las características *observadas*. Por supuesto, un precio de venta sustancialmente menor a su precio predicho indicará que existen algunas características indeseables de la vivienda que no han sido tomadas en cuenta y que están contenidas en el error no observado. Además de obtener la predicción y el residual, también es útil calcular un intervalo de confianza para el que puede ser el precio de venta de la vivienda, empleando el método descrito en la ecuación (6.37).

Empleando los datos del archivo HPRICE1.RAW, se corre la regresión de *price* sobre *lotsize*, *sqrft* y *bdrms*. En la muestra de 88 viviendas, el residual más negativo es -120.206, que es el de la casa número 81 de la muestra. Por tanto, el precio solicitado por esta vivienda es $120,206 inferior a su precio predicho.

El análisis de residuales tiene otros muchos usos. Una manera de jerarquizar las escuelas de leyes es regresando el sueldo inicial promedio sobre diversas características de los estudiantes [por ejemplo, la puntuación promedio en el examen de admisión a una escuela de leyes (LSAT, por sus siglas en inglés), al GPA promedio, etc.] y obtener un valor predicho y un residual para cada escuela de leyes. La escuela de leyes con el mayor residual tiene el valor agregado predicho más alto. (Por supuesto, todavía es incierta la relación entre el sueldo inicial de un individuo y la media general de una escuela de leyes.) Estos residuales pueden usarse junto con los costos de asistir a cada escuela de leyes para determinar el mejor valor; esto requerirá un descuento adecuado de las ganancias futuras.

El análisis de residuales también es importante en decisiones legales. Un artículo de la revista *New York Times* titulado "Judge Says Pupil's Poverty, Not Segregation, Hurts Scores" ("El juez dice que la pobreza de los alumnos, no la segregación, afecta las calificaciones") (6/28/95) describe un importante caso legal. El problema era si el mal desempeño en exámenes estandarizados del distrito escolar de Hartford, con relación al desempeño de suburbios vecinos se debía a la mala calidad de las escuelas más segregadas. El juez concluyó que "la disparidad en las puntuaciones de los exámenes no indica que Hartford esté trabajando mal o de manera inadecuada con relación a la educación de sus estudiantes o que sus escuelas estén fallando, ya que las puntuaciones predichas de acuerdo con los factores socioeconómicos relevantes se encuentran en los niveles que eran de esperarse". Esta conclusión está basada en un análisis de regresión de las puntuaciones promedio o medias sobre las características socioeconómicas de varios distritos escolares en Connecticut. La conclusión del juez indica que dados los niveles de pobreza de los estudiantes en las escuelas de Hartford, las puntuaciones reales en los exámenes son similares a las predichas con un análisis de regresión: el residual correspondiente a Hartford no era suficientemente negativo para concluir que las escuelas mismas fueron la causa de las bajas puntuaciones en el examen.

Pregunta 6.5

¿Cómo emplearía usted el análisis de residuales para determinar qué atletas profesionales son remunerados de manera excesiva o insuficiente con relación a su desempeño?

Predicción de *y* cuando log(*y*) es la variable dependiente

Como en la economía empírica se emplea con tanta frecuencia la transformación que usa el logaritmo natural, esta subsección se dedica a la predicción de y cuando la variable dependiente es log(y). Como subproducto se obtiene una medida de la bondad de ajuste para el modelo logarítmico, la cual puede compararse con la R-cuadrada del modelo lineal.

Para obtener una predicción es útil definir $logy = \log(y)$; con lo que se hace hincapié en que lo que predice el modelo es el logaritmo de y

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u. \tag{6.38}$$

En esta ecuación, las x_j pueden ser transformaciones de otras variables; así, en el ejemplo de los sueldos de los CEO puede tenerse $x_1 = \log(sales)$, $x_2 = \log(mktval)$, $x_3 = ceoten$.

Dados los estimadores de MCO, ya se sabe cómo predecir $logy$ para cualesquiera valores de las variables independientes:

$$\widehat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k. \tag{6.39}$$

Ahora, como la función exponencial elimina el logaritmo, el primer intento para predecir y es exponenciar el valor predicho para log(y): $\hat{y} = \exp(\widehat{logy})$. Esto no funciona; en realidad subestimará de manera *sistemática* el valor esperado de y. En efecto, si el modelo (6.38) sigue los supuestos RLM.1 a RLM.6 del MLC, puede demostrarse que

$$\mathrm{E}(y|\boldsymbol{x}) = \exp(\sigma^2/2)\cdot\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k),$$

donde $\boldsymbol{x}$ denota las variables independientes y σ^2 es la varianza de u. [Si $u \sim \text{Normal}(0,\sigma^2)$, entonces el valor esperado de $\exp(u)$ es $\exp(\sigma^2/2)$.] Esta ecuación muestra que para predecir y se necesita un ajuste sencillo:

$$\hat{y} = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}), \tag{6.40}$$

donde $\hat{\sigma}^2$ es simplemente el estimador insesgado de σ^2. Como siempre se reporta $\hat{\sigma}$, el error estándar de la regresión, obtener los valores predichos para y es sencillo. Como $\hat{\sigma}^2 > 0$, $\exp(\hat{\sigma}^2/2) > 1$. Cuando $\hat{\sigma}^2$ es grande, este factor de ajuste puede ser sustancialmente mayor que la unidad.

La predicción en (6.40) no es insesgada, pero es consistente. No hay predicciones insesgadas de y y, en muchos casos, la predicción de (6.40) funciona bien. Sin embargo, esta predicción se apoya en la normalidad del término del error, u. En el capítulo 5, se mostró que los MCO tienen propiedades deseables, aun cuando u no esté distribuido normalmente. Por tanto, es útil tener una predicción que no se apoye en la normalidad. Si simplemente se supone que u es independiente de las variables explicativas, entonces se tiene

$$\mathrm{E}(y|\boldsymbol{x}) = \alpha_0\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k), \tag{6.41}$$

donde α_0 es el valor esperado de $\exp(u)$, el cual debe ser mayor que la unidad.

Dada una estimación $\hat{\alpha}_0$, se puede predecir y de la manera siguiente

$$\hat{y} = \hat{\alpha}_0\exp(\widehat{logy}), \tag{6.42}$$

que una vez más simplemente debe exponenciar el valor predicho con el modelo logarítmico y multiplicar el resultado por $\hat{\alpha}_0$.

Se sugieren dos métodos para estimar α_0 sin el supuesto de normalidad. El primero se basa en $\alpha_0 = \mathrm{E}[\exp(u)]$. Para estimar α_0 se sustituye la esperanza poblacional por un promedio muestral y después se sustituyen los errores no observados, u_i, por los residuales de MCO, $\hat{u}_i = \log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$. Esto conduce al estimador del método de momentos (vea el apéndice C)

$$\hat{\alpha}_0 = n^{-1}\sum_{i=1}^{n}\exp(\hat{u}_i). \tag{6.43}$$

No debe extrañar que $\hat{\alpha}_0$ sea un estimador consistente de α_0, pero no es insesgado debido a que dentro de una función no lineal se ha sustituido u_i por $\hat{u}_i$. Esta versión de $\hat{\alpha}_0$ es un caso especial de lo que Duan (1983) llamó un estimador no paramétrico de retransformación **estimador smearing**. Como los residuales de MCO tienen promedio muestral cero, puede demostrarse que, para cualquier conjunto de datos, $\hat{\alpha}_0 > 1$. (Técnicamente, $\hat{\alpha}_0$ será igual a uno si todos los residuales de MCO son cero, pero esto no ocurre nunca en una aplicación interesante.) El que $\hat{\alpha}_0$ sea necesariamente mayor que uno es conveniente porque debe ser cierto que $\alpha_0 > 1$.

Otra estimación de α_0 se basa en una regresión simple a través del origen. Para ver cómo funciona, se define $m_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$, de manera que, de acuerdo con la ecuación (6.41), $\mathrm{E}(y_i|m_i) = \alpha_0 m_i$. Si las m_i pudieran observarse, mediante la regresión de y_i sobre m_i sin intercepto se obtendría un estimador insesgado de α_0. En lugar de esto, las β_j se sustituyen por sus estimaciones de MCO y se obtiene $\hat{m}_i = \exp(\widehat{logy}_i)$, donde, por supuesto, los $\widehat{logy}_i$ son los valores ajustados obtenidos de la regresión de $logy_i$ sobre $x_{i1}, \ldots, x_{ik}$ (con intercepto). Entonces $\check{\alpha}_0$ [para distinguirla de la $\hat{\alpha}_0$ de la ecuación (6.43)] es la pendiente estimada de MCO obtenida de la regresión simple de y_i sobre $\hat{m}_i$ (sin intercepto):

$$\check{\alpha}_0 = \left(\sum_{i=1}^{n} \hat{m}_i^2\right)^{-1}\left(\sum_{i=1}^{n} \hat{m}_i y_i\right). \tag{6.44}$$

A $\check{\alpha}_0$ se le llamará la estimación de la regresión de α_0. Al igual que $\hat{\alpha}_0$, $\check{\alpha}_0$ es consistente pero no insesgada. Curiosamente, no se garantiza que, $\check{\alpha}_0$ sea mayor que uno, aunque en la mayoría de las aplicaciones lo es. Si $\check{\alpha}_0$ es menor que uno, y en especial si es mucho menor que uno, es muy probable que se viole el supuesto de independencia entre u y x_j. Si $\check{\alpha}_0 < 1$, una posibilidad es emplear la estimación de (6.43), aunque esto puede sólo estar enmascarando algún problema con el modelo lineal de $\log(y)$. A continuación se resumen los pasos:

PREDICCIÓN DE *y* CUANDO LA VARIABLE DEPENDIENTE ES log(*y*):

1. Obtener los valores ajustados, $\widehat{logy}_i$, y los residuales, $\hat{u}_i$, mediante la regresión de $logy$ sobre $x_1, \ldots, x_k$.
2. Obtener $\hat{\alpha}_0$ de la ecuación (6.43) o $\check{\alpha}_0$ de la ecuación (6.44).
3. Para valores dados de $x_1, \ldots, x_k$, obtener $\widehat{logy}$ mediante (6.42).
4. Obtener la predicción $\hat{y}$ mediante (6.42) (con $\hat{\alpha}_0$ o $\check{\alpha}_0$).

A continuación se muestra cómo predecir el sueldo de los CEO empleando este procedimiento.

Ejemplo 6.7

[Predicción del sueldo de los directores generales]

El modelo de interés es

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 ceoten + u,$$

de manera que β_1 y β_2 son elasticidades y $100 \cdot \beta_3$ es una semielasticidad. La ecuación estimada empleando el archivo CEOSAL2.RAW es

$$\begin{aligned} \widehat{lsalary} &= \underset{(.257)}{4.504} + \underset{(.039)}{.163}\, lsales + \underset{(.050)}{.109}\, lmktval + \underset{(.0053)}{.0117}\, ceoten \\ n &= 177,\ R^2 = .318, \end{aligned} \tag{6.45}$$

donde, para mayor claridad *lsalary* denota el logaritmo de *salary*, y de manera similar *lsales* y *lmktval*. A continuación se obtiene $\hat{m}_i = \exp(\widehat{lsalary_i})$ para cada una de las observaciones de la muestra.

La estimación no paramétrica de retransformación de Duan de (6.43) es aproximadamente $\hat{\alpha}_0 = 1.136$ y la estimación de la regresión de acuerdo con (6.44) es $\check{\alpha}_0 = 1.117$. Se puede usar cualquier estimación para predecir *salary* para cualesquiera valores de *sales*, *mktval* y *ceoten*. Se va a determinar la predicción para *sales* = 5,000 (lo que significa \$5 mil millones porque *sales* está en millones), *mktval* = 10,000 (o \$10 mil millones) y *ceoten* = 10. De acuerdo con (6.45), la predicción para *lsalary* es $4.504 + .163 \cdot \log(5{,}000) + .109 \cdot \log(10{,}000) + .0117(10) \approx 7.013$ y $\exp(7.013) \approx 1{,}110.983$. Empleando la estimación de α_0 obtenido con (6.43), el sueldo predicho es aproximadamente 1,262.077, es decir, \$1,262.077. Usando el estimado que se obtiene con (6.44), el sueldo predicho es aproximadamente \$1,240,968. Éstos difieren uno de otro en mucho menos de lo que cada uno difiere de la ingenua predicción de \$1,110,983.

Los métodos anteriores pueden usarse en la obtención de predicciones para determinar qué tan bien puede explicar y el modelo que usa $\log(y)$ como variable dependiente. Ya se tienen mediciones para los modelos en los que y es la variable dependiente: R-cuadrada y R-cuadrada ajustada. El objetivo es hallar una medida de la bondad de ajuste para el modelo $\log(y)$ que pueda compararse con una R-cuadrada de un modelo en el que y sea la variable dependiente.

Existen varias maneras de definir una medida de la bondad de ajuste después de transformar un modelo con $\log(y)$ para predecir y. Aquí se presenta un método que es fácil de realizar y con el que se obtiene el mismo valor ya sea que α_0 se estime como en (6.40), (6.43) o (6.44). Para motivar esta medida, recuerde que en la ecuación de regresión lineal estimada mediante MCO,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \tag{6.46}$$

la R-cuadrada usual es simplemente el cuadrado de la correlación entre y_i y $\hat{y}_i$ (vea la sección 3.2). Ahora, si en lugar de esto se calculan valores ajustados de acuerdo con (6.42) —es decir, $\hat{y}_i = \hat{\alpha}_0 m_i$ para todas las observaciones i— entonces tiene sentido emplear, como una R-cuadrada, el cuadrado de la correlación entre y_i y estos valores ajustados. Como a la correlación no le afecta que se multiplique por una constante, no importa qué estimación de α_0 se emplee. En realidad, esta medida R-cuadrada para y [no para $\log(y)$] es precisamente el cuadrado de la correlación entre y_i y $\hat{m}_i$. Esto se puede comparar de manera directa con la R-cuadrada de la ecuación (6.46). [Como el cálculo de la R-cuadrada no depende de la estimación de α_0, no permite elegir entre (6.40), (6.43) y (6.44). Pero se sabe que (6.44) minimiza la suma de los residuales cuadrados entre y_i y $\hat{m}_i$ sin una constante. En otras palabras, dados los $\hat{m}_i$, $\check{\alpha}_0$ se elige de manera que produzca el mejor ajuste con base en la suma de los residuales cuadrados. En lo que se está interesado aquí es en elegir entre el modelo lineal para y y $\log(y)$, y de esta manera es adecuada una medida de R-cuadrada que no dependa de la manera en que se estima α_0.]

Ejemplo 6.8

[Predicción de los sueldos de los directores generales (CEO)]

Después de obtener los $\hat{m}_i$ simplemente se obtiene la correlación entre $salary_i$ y $\hat{m}_i$; ésta es .493. Su cuadrado es aproximadamente .243 y esta es una medida de qué tan bien explica el modelo logarítmico la variación de *salary* no de log(*salary*). [La R^2 obtenida con (6.45), .318, indica que el modelo logarítmico explica aproximadamente 31.8% de la variación en log(*salary*).]

Como modelo lineal alternativo, suponga que se estima un modelo con todas las variables en forma lineal:

$$salary = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u. \tag{6.47}$$

La clave es que la variable dependiente es *salary*. En el lado derecho se podrían usar los logaritmos de *sales* o *mktval*, pero, si (*salary*) aparece de forma lineal, es más razonable tener todos los valores en dólares en forma lineal. La *R*-cuadrada de la estimación de esta ecuación empleando las mismas 177 observaciones es .201. De manera que el modelo logarítmico explica más de la variación en *salary*, y por tanto se prefiere a (6.47) por razones de la bondad de ajuste. El modelo logarítmico se prefiere también debido a que parece más realista y sus parámetros son más fáciles de interpretar.

Si en el modelo (6.38) se conserva la base completa de supuestos del modelo lineal clásico, con facilidad se pueden obtener intervalos de predicción para $y^0 = \exp(\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0 + u^0)$ una vez que se haya estimado el modelo lineal para log(y). Recuerde que $x_1^0, x_2^0, \ldots, x_k^0$ son valores conocidos y que u^0 es el error no observado que determina de manera parcial y^0. De acuerdo con la ecuación (6.37), un intervalo de predicción de 95% para $logy^0 = \log(y^0)$ es simplemente $\widehat{logy}^0 \pm t_{.025} \cdot \mathrm{ee}(\hat{e}^0)$, donde $\mathrm{ee}(\hat{e}^0)$ se obtiene mediante la regresión de log(y) sobre $x_1, \ldots, x_k$ empleando las n observaciones originales. Sean $c_l = -t_{.025} \cdot \mathrm{ee}(\hat{e}^0)$ y $c_u = t_{.025} \cdot \mathrm{ee}(\hat{e}^0)$ los límites inferior y superior del intervalo de predicción para $logy^0$. Es decir, $\mathrm{P}(c_l \leq logy^0 \leq c_u) = .95$. Dado que la función exponencial es estrictamente creciente, también se tiene que $\mathrm{P}[\exp(c_l) \leq \exp(logy^0) \leq \exp(c_u)] = .95$, es decir, $\mathrm{P}[\exp(c_l) \leq y^0 \leq \exp(c_u)] = .95$. Por tanto, se pueden tomar $\exp(c_l)$ y $\exp(c_u)$ como los límites inferior y superior, respectivamente, de un intervalo de predicción de 95% para y^0. Para n grande, $t_{.025} = 1.96$, y de esta manera un intervalo de predicción de 95% para y^0 es $\exp[-1.96 \cdot \mathrm{ee}(\hat{e}^0)] \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ a $\exp[-1.96 \cdot \mathrm{ee}(\hat{e}^0)] \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$, donde $\mathbf{x}^0\hat{\boldsymbol{\beta}}$ es una abreviación de $\hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_k x_k^0$. Recuerde, los $\hat{\beta}_j$ y $\mathrm{ee}(\hat{e}^0)$ se obtienen mediante la regresión con log(y) como variable dependiente. Como en (6.38) se asume la normalidad de u, para obtener una predicción puntual de y^0. probablemente se emplee (6.40). A diferencia de la ecuación (6.37), esta predicción puntual no se encontrará a la mitad entre los límites inferior y superior $\exp(c_l)$ y $\exp(c_u)$. Se pueden obtener distintos intervalos de predicción de 95% eligiendo distintas cantidades en la distribución t_{n-k-1}. Si $q_{\alpha 1}$ y $q_{\alpha 2}$ son cuantiles con $\alpha_2 - \alpha_1 = .95$, entonces se pueden elegir $c_l = q_{\alpha 1}\mathrm{ee}(\hat{e}^0)$ y $c_u = q_{\alpha 2}\mathrm{ee}(\hat{e}^0)$.

Como ejemplo, considérese la regresión del sueldo de CEO, donde la predicción sobre los mismos valores de *sales*, *mktval* y *ceoten* que en el ejemplo 6.7. El error estándar de la regresión para (6.43) es aproximadamente .505 y el error estándar de $\widehat{logy}^0$ es .075. Por tanto, empleando la ecuación (6.36), $\mathrm{ee}(\hat{e}^0) \approx .511$; como en el ejemplo del promedio general de calificaciones (GPA), la varianza del error domina el error de estimación en los parámetros, aun cuando aquí el tamaño de la muestra sea de sólo 177. Un intervalo de predicción de 95% para $salary^0$ es $\exp[-1.96 \cdot (.511)] \exp(7.013)$ a $\exp[1.96 \cdot (.511)] \exp(7.013)$, es decir, aproximadamente 408.071 a 3,024.678, es decir \$408.071 a \$3,024.678. Este intervalo de predicción de 95%, tan amplio, para el sueldo de CEO, a los valores dados de ventas, valor de mercado y de antigüedad, indica que existen muchos más factores que no fueron incluidos en la regresión que determina el sueldo. Dicho sea de paso, la predicción puntual para el sueldo (*salary*), empleando (6.40), es aproximadamente \$1,262, 075 —superior a las predicciones empleando las otras estimaciones de α_0 y más cercana al límite inferior que al límite superior del intervalo de predicción de 95 por ciento.

RESUMEN

En este capítulo se vieron algunos temas importantes del análisis de regresión múltiple.

En la sección 6.1 se mostró que cualquier cambio en las unidades de medición de una variable independiente, modifica los coeficientes de MCO como era de esperarse: si x_j se multiplica por c, su coeficiente se dividirá entre c. Si la variable dependiente se multiplica por c, *todos* los coeficientes de MCO se multiplicarán por c. Modificar las unidades de medición de una variable no afecta ni a los estadísticos t ni a los estadísticos F.

Se analizaron los coeficientes beta, los cuales miden los efectos de las variables independientes en unidades de desviaciones estándar. Los coeficientes beta se obtienen de la regresión usual de MCO después de transformar la variable dependiente y las variables independientes en valores-z.

Como ya se ha visto en varios ejemplos, la forma funcional logarítmica proporciona coeficientes que se interpretan como efectos porcentuales. En la sección 6.2 se analizaron sus ventajas adicionales. Se vio también cómo calcular el efecto porcentual exacto en caso de que en un modelo logarítmico lineal un coeficiente sea grande. Modelos con términos cuadráticos permiten efectos marginales, ya sea crecientes o decrecientes. Modelos con interacciones permiten que el efecto marginal de una variable explicativa dependa del nivel de otra variable explicativa.

Se introdujo la R-cuadrada ajustada, $\bar{R}^2$, como una alternativa a la R-cuadrada usual en la medición de la bondad de ajuste. Mientras que R^2 nunca puede disminuir cuando se agrega una variable más a la regresión, $\bar{R}^2$ sí penaliza la cantidad de regresores y puede disminuir cuando se agrega una variable independiente. Esto hace que $\bar{R}^2$ se prefiera cuando se trata de elegir entre modelos no anidados con cantidades diferentes de variables explicativas. En la comparación de modelos con variables dependientes diferentes no se puede usar ni R^2 ni $\bar{R}^2$. Sin embargo, como se mostró en la sección 6.4, es bastante fácil obtener medidas de la bondad de ajuste para elegir entre y y $\log(y)$ como variable dependiente.

En la sección 6.3 se analizó el problema algo sutil de confiar demasiado en R^2 o en $\bar{R}^2$ para llegar a un modelo final: es posible que en un modelo de regresión se controlen demasiados factores. Debido a esto es importante pensar de antemano en las especificaciones del modelo, en especial en la naturaleza *ceteris paribus* de la ecuación de regresión. Las variables explicativas que afectan a y y que no están correlacionadas con todas las demás variables explicativas pueden emplearse para reducir la varianza del error sin inducir multicolinealidad.

En la sección 6.4 se demostró cómo obtener un intervalo de confianza para una predicción hecha a partir de la línea de regresión de MCO. También se mostró cómo construir un intervalo de confianza para un valor futuro, desconocido de y.

Ocasionalmente se desea predecir y cuando en un modelo de regresión se ha usado $\log(y)$ como variable dependiente. En la sección 6.4 se explica un sencillo método para esto. Por último, algunas veces se tiene interés en conocer el signo y la magnitud de los residuales de determinadas observaciones. Para determinar si los valores predichos para ciertos miembros de una muestra son mucho mayores o mucho menores que los valores reales puede emplearse el análisis de residuales.

TÉRMINOS CLAVE

Análisis de residuales
Bootstrap
Coeficientes beta
Coeficientes estandarizados
Efecto de interacción
Error de predicción
Error estándar bootstrap
Estimador no paramétrico de retransformación (estimador smearing)
Funciones cuadráticas
Intervalo de predicción
Métodos de remuestreo
Modelos no anidados
Predicciones
R-cuadrada ajustada
R-cuadrada poblacional
Sobrecontrol
Varianza del error de predicción

PROBLEMAS

6.1 La ecuación siguiente se estimó empleando los datos del archivo CEOSAL1.RAW:

$$\widehat{\log(salary)} = \underset{(.324)}{4.322} + \underset{(.033)}{.276} \log(sales) + \underset{(.0129)}{.0215}\, roe - \underset{(.00026)}{.00008}\, roe^2$$
$$n = 209,\ R^2 = .282.$$

Esta ecuación permite que *roe* tenga un efecto decreciente sobre log(*salary*). ¿Es esto, en general, necesario? Explique por qué.

6.2 Sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ las estimaciones de MCO de la regresión de y_i sobre $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. Dadas constantes distintas de cero $c_1, \ldots, c_k$, argumente que el intercepto y las pendientes de MCO de la regresión de c_0y_i sobre $c_1x_{i1}, \ldots, c_kx_{ik}$, $i = 1, 2, \ldots, n$, están dadas por $\tilde{\beta}_0 = c_0\hat{\beta}_0$, $\tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Sugerencia*: use el hecho de que las $\hat{\beta}_j$ resuelven las condiciones de primer orden en (3.13), y que las $\tilde{\beta}_j$ deben resolver las condiciones de primer orden relacionadas con la reescalación de las variables dependiente e independiente.]

6.3 Empleando los datos del archivo RDCHEM.RAW, mediante MCO se obtuvo la ecuación siguiente:

$$\widehat{rdintens} = \underset{(.429)}{2.613} + \underset{(.00014)}{.00030}\, sales - \underset{(.0000000037)}{.0000000070}\, sales^2$$
$$n = 32,\ R^2 = .1484.$$

i) ¿En qué punto se vuelve negativo el efecto de *sales* sobre *rdintens*?
ii) ¿Conservaría usted el término cuadrático del modelo? Explique.
iii) Defina *salesbil* como las ventas medidas en miles de millones de dólares: $salesbil = sales/1{,}000$. Escriba de nuevo la ecuación estimada con *salesbil* y $salesbil^2$ como variables independientes. No olvide dar los errores estándar y la R-cuadrada. [*Sugerencia*: observe que $salesbil^2 = sales^2/(1{,}000)^2$.]
iv) ¿Qué ecuación prefiere con objeto de dar los resultados?

6.4 El modelo siguiente permite que el rendimiento de la educación sobre el salario dependa de la cantidad total de educación de los dos padres, denominada *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

i) Muestre que, de forma decimal, el rendimiento de un año más de educación en este modelo es

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

¿Qué signo espera que tenga β_2? ¿Por qué?

ii) Empleando los datos del archivo WAGE2.RAW, la ecuación estimada es

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(.13)}{5.65} + \underset{(.010)}{.047}\ educ + \underset{(.00021)}{.00078}\ educ \cdot pareduc + \\ &\quad \underset{(.004)}{.019}\ exper + \underset{(.003)}{.010}\ tenure \\ n &= 722,\ R^2 = .169.\end{aligned}$$

(Sólo 722 observaciones contienen información completa sobre la educación de los padres.) Interprete el coeficiente del término de interacción. Puede ayudar elegir dos valores específicos para *pareduc* —por ejemplo, *pareduc* = 32 si ambos padres tienen educación universitaria o *pareduc* = 24 si los dos padres tienen bachillerato— y de esta manera estimar el rendimiento de *educ*.

iii) Si a esta ecuación se le agrega *pareduc* como una variable aparte, se obtienen:

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(.38)}{4.94} + \underset{(.027)}{.097}\ educ + \underset{(.017)}{.033}\ pareduc - \underset{(.0012)}{.0016}\ educ \cdot pareduc \\ &\quad + \underset{(.004)}{.020}\ exper + \underset{(.003)}{.010}\ tenure \\ n &= 722,\ R^2 = .174.\end{aligned}$$

¿Depende ahora el rendimiento estimado de la educación positivamente de la educación de los padres? Pruebe la hipótesis nula de que el rendimiento de la educación no depende de la educación de los padres.

6.5 En el ejemplo 4.2, en donde el porcentaje de estudiantes que obtienen una puntuación aprobatoria en el examen del décimo grado (*math*10) es la variable dependiente, ¿tiene sentido incluir *sci11* —porcentaje de estudiantes del grado undécimo que aprueban el examen de ciencias— como una variable explicativa más?

6.6 Cuando a la ecuación estimada en (6.19) se le agregan $atndrte^2$ y $ACT \cdot atndrte$ la R-cuadrada se vuelve .232. Al nivel de significancia de 10%, ¿son estos términos adicionales conjuntamente significativos? ¿Los incluiría usted en el modelo?

6.7 Las siguientes tres ecuaciones se estimaron empleando 1,534 observaciones del archivo 401K.RAW:

$$\begin{aligned}\widehat{prate} &= \underset{(.78)}{80.29} + \underset{(.52)}{5.44}\ mrate + \underset{(.045)}{.269}\ age - \underset{(.00004)}{.00013}\ totemp \\ R^2 &= .100,\ \bar{R}^2 = .098.\end{aligned}$$

$$\begin{aligned}\widehat{prate} &= \underset{(1.95)}{97.32} + \underset{(0.51)}{5.02}\ mrate + \underset{(.044)}{.314}\ age - \underset{(.28)}{2.66}\ \log(totemp) \\ R^2 &= .144,\ \bar{R}^2 = .142.\end{aligned}$$

$$\begin{aligned}\widehat{prate} &= \underset{(.78)}{80.62} + \underset{(.52)}{5.34}\ mrate + \underset{(.045)}{.290}\ age - \underset{(.00009)}{.00043}\ totemp \\ &\quad + \underset{(.0000000010)}{.0000000039}\ totemp^2 \\ R^2 &= .108,\ \bar{R}^2 = .106.\end{aligned}$$

¿Cuál de estos tres modelos prefiere usted? ¿Por qué?

6.8 Suponga que se desea estimar el efecto del alcohol (*alcohol*) sobre el promedio general de calificaciones en la universidad (*colGPA*). Además de la información acerca del consumo del alcohol y el promedio general de calificaciones, también se obtiene información sobre la asistencia a clases (porcentaje de asistencia a clases, que se denomina *attend*). Se cuenta también con la calificación en una prueba estandarizada *SAT*) y con el promedio general de calificaciones en el bachillerato (*hsGPA*).

i) En un modelo de regresión múltiple, ¿debe incluirse *attend* además de *alcohol* como variables explicativas en un modelo de regresión múltiple? (Reflexione sobre cómo se interpretaría $\beta_{alcohol}$.)

ii) ¿Hay que incluir *SAT* y *hsGPA* como variables explicativas? Explique.

6.9 Si se empieza en (6.38) bajo los supuestos del MLC, suponiendo que n es grande e ignorando el error de estimación en las $\hat{\beta}_j$, un intervalo de predicción de 95% para y^0 es $[\exp(-1.96\hat{\sigma})\exp(\widehat{logy}^0)$, $\exp(1.96\hat{\sigma})\exp(\widehat{logy}^0)]$. La predicción puntual para y^0 es $\hat{y}^0 = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}^0)$.

i) ¿Para qué valores de $\hat{\sigma}$ estará la predicción puntual en el intervalo de predicción de 95%? ¿Es posible que se satisfaga esta condición en la mayor parte de las aplicaciones?

ii) Compruebe que la condición del inciso i) se verifica en el ejemplo del sueldo de CEO.

EJERCICIOS EN COMPUTADORA

C6.1 Del archivo KIELMC.RAW emplee los datos de 1981 para responder las preguntas siguientes. Estos datos son de viviendas vendidas durante 1981 en North Andover, Massachusetts; 1981 fue el año en que se inició la construcción de un incinerador local de basura.

i) Para estudiar el efecto de la ubicación del incinerador sobre los precios de la vivienda, considere el modelo de regresión simple

$$\log(price) = \beta_0 + \beta_1\log(dist) + u,$$

donde *price* es el precio de la vivienda en dólares y *dist* es la distancia de la vivienda al incinerador, medida en pies. Interpretando esta ecuación de manera causal, ¿que signo espera usted que tenga β_1 si la presencia del incinerador hace decrecer el precio de la vivienda? Estime esta ecuación e interprete los resultados.

ii) Al modelo de regresión simple del inciso i) agregue las variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths* y *age*, donde *intst* es la distancia de la vivienda a la carretera interestatal, *area* es el área de la vivienda en pies cuadrados, *land* es el tamaño del terreno en pies cuadrados, *rooms* es la cantidad de habitaciones, *baths* es la cantidad de baños y *age* la antigüedad de la vivienda en años. Ahora, ¿qué concluye usted acerca del efecto del incinerador? Explique por qué los incisos i) y ii) dan resultados contradictorios.

iii) Al modelo del inciso ii) agregue la variable $[\log(intst)]^2$. ¿Qué pasa ahora? ¿Qué concluye usted acerca de la importancia de la forma funcional?

iv) ¿Es significativo el cuadrado de log(*dist*) cuando se agrega al modelo del inciso iii)?

C6.2 Para este ejercicio emplee los datos del archivo WAGE1.RAW.

i) Use MCO para estimar la ecuación

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

y dé los resultados empleando el formato acostumbrado.

ii) ¿Es $exper^2$ estadísticamente significativa al nivel de 1%?

iii) Empleando la aproximación

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

determine el rendimiento aproximado que tiene el quinto año de experiencia. ¿Cuál es el rendimiento aproximado del vigésimo año de experiencia?

iv) ¿Cuál es el valor de *exper* al que más experiencia disminuye el log(*wage*) predicho? ¿En esta muestra cuántas personas tienen una experiencia mayor que ese nivel?

C6.3 Considere un modelo en el que el rendimiento de la educación depende de la cantidad de experiencia de trabajo (y viceversa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$

i) Muestre que el rendimiento de un año más de educación (en forma decimal, manteniendo *exper* constante, es $\beta_1 + \beta_3 exper$.

ii) Establezca la hipótesis nula de que el rendimiento de la educación no depende del nivel de *exper*. ¿Cuál piensa que sea la alternativa adecuada?

iii) Emplee los datos del archivo WAGE2.RAWpara probar la hipótesis nula del inciso ii) contra la alternativa dada por usted.

iv) Sea θ_1 el rendimiento de la educación (en forma decimal) cuando $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Obtenga $\hat{\theta}_1$ y un intervalo de confianza de 95% para θ_1. (*Sugerencia*: escriba $\beta_1 = \theta_1 - 10\beta_3$ y sustitúyalo en la ecuación; después ordénelo. Esto da la regresión para obtener el intervalo de confianza para θ_1.)

C6.4 Para hacer este ejercicio, emplee los datos del archivo GPA2.RAW.

i) Estime el modelo

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

donde *sat* es el puntaje del examen de admisión a la universidad y *hsize* es el tamaño del grupo que termina sus estudios de bachillerato (en miles); dé los resultados en la forma habitual. ¿Es el término cuadrático estadísticamente significativo?

ii) Empleando la ecuación estimada en el inciso i), ¿cuál es el tamaño "óptimo" para un grupo de bachillerato? Justifique su respuesta.

iii) ¿Es representativo este análisis del desempeño de *todos* los estudiantes que terminan bachillerato? Explique.

iv) Encuentre el tamaño óptimo de una escuela de bachillerato, empleando log(*sat*) como variable dependiente. ¿Es esto muy distinto a lo que obtuvo en el inciso ii)?

C6.5 Para hacer este ejercicio, emplee los datos del archivo HPRICE1.RAW.

i) Estime el modelo

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

y dé los resultados en el formato habitual de MCO.

ii) Determine el valor que se predice para log(*price*), cuando *lotsize* = 20,000, *sqrft* = 2,500 y *bdrms* = 4. Empleando los métodos de la sección 6.4, halle el valor que se predice para *price* a estos mismos valores de la variable explicativa.

iii) Si se trata de explicar la variación de *price*, diga si prefiere el modelo del inciso i) o el modelo

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u.$$

C6.6 Para hacer este ejercicio, emplee los datos del archivo VOTE1.RAW.

i) Considere un modelo con una interacción entre los gastos:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

¿Cuál es el efecto parcial de *expendB* sobre *voteA*, cuando se mantienen constantes *prtystrA* y *expendA*? ¿Cuál es el efecto parcial de *expendA* sobre *voteA*? ¿Es obvio el signo que se espera para β_4?

ii) Estime la ecuación del inciso i) y dé los resultados de la manera habitual. ¿Es estadísticamente significativo el término de interacción?

iii) Determine el promedio de *expendA* en la muestra. Fije *expendA* en 300 (que significa \$300,000). ¿Cuál es el efecto estimado de \$100,000 más gastados por el candidato B sobre *voteA*? ¿Es grande este efecto?

iv) Ahora fije *expendB* en 100. ¿Cuál es el efecto estimado de $\Delta expendA = 100$ sobre *voteA*? ¿Es esto razonable?

v) Ahora, estime un modelo en el que la interacción se remplace por *shareA*, la participación porcentual del candidato A en los gastos de campaña. ¿Tiene sentido mantener constantes *expendA* y *expendB* y variar *shareA*?

vi) (Se requiere cálculo.) En el modelo del inciso v), determine el efecto parcial de *expendB* sobre *voteA*, manteniendo constantes *prtystrA* y *expendA*. Evalúe este modelo cuando *expendA* = 300 y *expendB* = 0 y comente los resultados.

C6.7 Para hacer este ejercicio, emplee los datos del archivo ATTEND.RAW.

i) En el modelo del ejemplo 6.3 argumente que

$$\Delta stndfnl/\Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Emplee la ecuación 6.19 para estimar el efecto parcial cuando *priGPA* = 2.59 y *atndrte* = 82. Interprete su estimación.

ii) Muestre que la ecuación puede escribirse como

$$\begin{aligned} stndfnl = \theta_0 &+ \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 \\ &+ \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u, \end{aligned}$$

donde $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. (Observe que el intercepto ha cambiado, pero esto no tiene importancia.) Use esto para obtener el error estándar de $\hat{\theta}_2$ de acuerdo con el inciso i).

iii) Suponga que en lugar de *priGPA*(*atndrte* − 82), emplea (*priGPA* − 2.59)· (*atndrte* − 82). ¿Cómo interpreta ahora los coeficientes de *atndrte* y de *priGPA*?

C6.8 Para hacer este ejercicio, emplee los datos del archivo HPRICE.RAW.

i) Estime el modelo

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

y dé los resultados de la manera habitual, dando también el error estándar de regresión. Obtenga el precio que se predice si se sustituye $lotsize = 10{,}000$, $sqrft = 2{,}300$ y $bdrms = 4$; redondee el precio obtenido al dólar más cercano.

ii) Corra una regresión que le permita colocar un intervalo de confianza de 95% en torno del valor predicho en el inciso i). Observe que su predicción diferirá un poco debido al error de redondeo.

iii) Sea $price^0$ el precio futuro no conocido al que se vende la vivienda con las características empleadas en los incisos i) y ii). Encuentre un IC de 95% para $price^0$ y analice la amplitud de este intervalo de confianza.

C6.9 La base de datos NBASAL.RAW contiene información de los sueldos y estadísticas sobre las carreras de 269 jugadores de la National Basketball Association (NBA).

i) Estime un modelo que relacione puntos por juego (*points*) con años en la liga (*exper*), edad (*age*), y años que se ha jugado en la universidad (*coll*). Incluya un término cuadrático en *exper*; las demás variables deben aparecer en forma lineal. Dé los resultados en la forma habitual.

ii) Manteniendo constantes los años en la universidad (*coll*) y la edad (*age*), ¿a qué valor de la experiencia un año adicional de ésta reduce los puntos por juego? ¿Es esto razonable?

iii) ¿Por qué piensa usted que *coll* tenga un coeficiente negativo y estadísticamente significativo? (*Sugerencia:* los jugadores de la NBA pueden ser reclutados antes de terminar sus carreras universitarias e incluso directamente al salir del bachillerato.)

iv) Agregue a la ecuación un término cuadrático en *age*. ¿Es necesario ese término? ¿Qué parece implicar este término acerca de los efectos de *age*, una vez controlado experiencia y educación?

v) Ahora regrese log(*wage*) sobre *points*, *exper*, $exper^2$, *age* y *coll*. Dé los resultados en el formato habitual.

vi) Pruebe si en la regresión del inciso v) son *age* y *coll* conjuntamente significativas. ¿Qué implica esto acerca de si la edad (*age*) y la educación tienen efectos separados sobre el salario (*wage*) una vez tomadas en cuenta la productividad y la antigüedad?

C6.10 Para hacer este ejercicio, emplee los datos del archivo BWGHT2.RAW.

i) Mediante MCO estime la ecuación

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

donde *bwght* es el peso de los niños al nacer y *npvis* es el número de visitas prenatales al médico; dé los resultados de la manera habitual. ¿Es significativo el término cuadrático?

ii) Muestre que, de acuerdo con la ecuación del inciso i), el número de visitas prenatales que maximiza log(*bwght*) se estima que es alrededor de 22. En la muestra, ¿cuántas mujeres tienen por lo menos 22 visitas prenatales?

iii) ¿Es razonable que el peso al nacer disminuya con más de 22 visitas prenatales? Explique.

iv) Agregue a esta ecuación la edad de la madre, empleando una forma funcional cuadrática. Manteniendo constante *npvis* ¿cuál es la edad de la madre en la que el peso del niño al nacer alcanza su máximo? ¿Qué proporción de las mujeres de la muestra tienen una edad mayor a la "óptima"?

v) ¿Considera usted que la edad de la madre y el número de visitas prenatales explican mucho de la variación en log(*bwght*)?

vi) Empleando términos cuadráticos tanto para *npvis* como para *age* (edad), diga si para predecir *bwght* es mejor emplear el logaritmo natural de *bwght* o *bwght* en nivel original.

C6.11 Para verificar algunas de las afirmaciones hechas en la sección 6.3 emplee el archivo APPLE.RAW.

i) Corra la regresión de *ecolbs* sobre *ecoprc*, *regprc* y dé los resultados de la manera habitual, dando también la R-cuadrada y la R-cuadrada ajustada. Interprete los coeficientes de las variables del precio y haga un comentario sobre sus signos y magnitudes.

ii) ¿Son estadísticamente significativas las variables del precio? Dé los valores-p de las pruebas t individuales.

iii) ¿Cuál es el rango de los valores ajustados para *ecolbs*? ¿En qué proporción de la muestra se tiene $ecolbs = 0$? Analice.

iv) ¿Considera usted que las variables del precio, juntas, explican suficiente la variación en *ecolbs*? Explique.

v) A la regresión del inciso i) agregue las variables *faminc*, *hhsize* (tamaño de la familia), *educ* y *age* (edad). Encuentre el valor-p para su significancia conjunta. ¿Qué concluye usted?

C6.12 Emplee el subconjunto del archivo 401KSUBS.RAW con $fsize = 1$; esto restringe el análisis a los hogares de una sola persona; vea también el ejercicio para computadora C4.8.

i) ¿Cuál es la edad de las personas más jóvenes en esta muestra? ¿Cuántas personas tienen esa edad?

ii) En el modelo

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

¿cuál es la interpretación literal de β_2? ¿Tiene mucho interés en sí misma?

iii) Estime el modelo del inciso ii) y dé los resultados de la manera habitual. ¿Le preocupa que el coeficiente de *age* (edad) sea negativo? Explique.

iv) Dado que las personas más jóvenes de la muestra tienen 25 años, es razonable pensar que, dado un determinado nivel de ingreso, la menor cantidad promedio de activo financiero neto es a la edad de 25 años. Recuerde que el efecto parcial de *age* sobre *nettfa* es $\beta_2 + 2\beta_3 age$, de manera que este efecto parcial a la edad de 25 años es $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; llámese a esto θ_2. Determine $\hat{\theta}_2$ y obtenga el valor-p de dos colas para probar H_0: $\theta_2 = 0$. Debe concluir que $\hat{\theta}_2$ es pequeño y estadísticamente muy poco significativo. [*Sugerencia:* una manera de hacer esto es estimar el modelo $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, donde el intercepto, α_0, es diferente de β_0. Existen también otras maneras de hacerlo.]

v) Dado que la evidencia contra H_0: $\theta_2 = 0$ es muy débil, iguálela a cero y estime el modelo

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

En términos de la bondad de ajuste, ¿es mejor este modelo que el del inciso ii)?

vi) Dada la ecuación estimada en el inciso v), haga $inc = 30$ (aproximadamente, el valor promedio) y grafique la relación entre *nettfa* y *age*, pero sólo para $age \geq 25$. Describa lo que ve.

vii) Verifique si es necesario incluir un término cuadrático para *inc*.

C6.13 Para este ejercicio emplee los datos del archivo MEAP00_01.

i) Estime mediante MCO el modelo

$$math4 = \beta_0 + \beta_2 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

donde *math*4 es el porcentaje de aprobación en matemáticas en 4o. grado, *lexppp* es el logaritmo del gasto por alumno, *lenroll* es el logaritmo del número de alumnos en la escuela, *lunch* es el porcentaje de estudiantes con desayuno gratuito o subsidiado; dé los resultados en la forma habitual. ¿Es cada una de las variables explicativas estadísticamente significativa al nivel de 5%?

ii) Obtenga los valores ajustados a partir de la regresión del inciso i). ¿Cuál es el rango de los valores ajustados? ¿Cómo es este rango en comparación con el rango de los datos reales en *math*4?

iii) Obtenga los residuales correspondientes a la regresión del inciso i). ¿Cuál es el código de la escuela (*bcode*) que tiene el residual (*positivo*) mayor? Interprete este residual.

iv) Agregue a la ecuación términos cuadráticos de todas las variables explicativas y pruebe su significancia conjunta. ¿Dejaría usted estos términos en el modelo?

v) Volviendo al modelo del inciso i), divida la variable independiente y cada una de las variables explicativas entre su desviación estándar muestral y vuelva a correr la regresión. (Incluya un intercepto a menos que primero también reste a cada variable su media.) En términos de unidades de desviaciones estándar, ¿cuál de las variables explicativas tiene el mayor efecto sobre la tasa de aprobación en matemáticas?

Apéndice 6A

6A. Breve introducción al bootstrapping

En muchos casos en los que se dificulta obtener matemáticamente las fórmulas para el error estándar, o cuando se cree que éstas no son aproximaciones muy buenas a la verdadera variación de muestreo de un estimador, puede uno apoyarse en un **método de remuestreo**. La idea general es tratar los datos observados como una población de donde se pueden sacar muestras. El método de remuestreo más común es el **bootstrap**. (En realidad existen varias versiones de bootstrap, pero el más general y de más fácil aplicación, al que se le llama *bootstrap no paramétrico* es el que se describe aquí.)

Suponga que se tiene una estimación $\hat{\theta}$ de un parámetro poblacional, θ. Esta estimación, que puede ser función de estimaciones de MCO (o de estimaciones que se verán en capítulos posteriores), fue obtenido de una muestra aleatoria de tamaño n. Se desea obtener un error estándar de $\hat{\theta}$ que pueda emplearse para calcular estadísticos t o intervalos de confianza. Naturalmente, un error estándar válido puede obtenerse calculando la estimación a partir de varias muestras aleatorias obtenidas de los datos originales.

La implementación es sencilla. Si se numeran las observaciones de la 1 a la n, y se obtienen de forma aleatoria n de estos números, con reposición. Esto produce un nuevo conjunto de datos (de tamaño n) que consta de los datos originales, pero en el que muchas de las observaciones aparecen repetidas (salvo en el raro caso que se obtenga la base original). Cada vez que se toma una muestra aleatoria de los datos originales, puede estimarse θ empleando el mismo procedimiento que se empleó con los datos originales. Sea $\hat{\theta}^{(b)}$ la estimación obtenida por bootstrap de la muestra b. Ahora, repitiendo el remuestreo y la estimación m veces, se obtienen m nuevas estimaciones, $\{\hat{\theta}^{(b)}: b = 1, 2, \dots, m\}$. El **error estándar bootstrap** de $\hat{\theta}$ es precisamente la desviación estándar muestral de los $\hat{\theta}^{(b)}$, a saber,

$$\text{eeb}(\hat{\theta}) = \left[(m-1)^{-1}\sum_{b=1}^{m}(\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2\right]^{1/2}, \tag{6.48}$$

donde $\bar{\hat{\theta}}$ es el promedio de las estimaciones bootstrap.

Si el tiempo de cálculo requerido para obtener una estimación de θ con una muestra de tamaño n es poco, como en el caso de MCO y de todos los demás estimadores encontrados en este libro, puede elegirse una m —cantidad de réplicas bootstrap— que sea grande. Un valor típico es $m = 1{,}000$, pero incluso $m = 500$ o valores un poco menores permiten obtener un error estándar confiable. Obsérvese que m —la cantidad de veces que se remuestrea de los datos originales— no tiene nada que ver con el tamaño n de la muestra. (En ciertos problemas de estimación, que quedan fuera del alcance de este libro, una n grande puede obligar a hacer menos réplicas bootstrap.) Muchos de los paquetes para estadística y econometría cuentan con comandos para bootstrap, lo que facilita el cálculo de errores estándar, en especial en comparación con el trabajo que se requiere para obtener una fórmula analítica para un error estándar asintótico.

En muchos casos, en lugar de obtener un error estándar bootstrap para emplearlo en la construcción de estadísticos t o de intervalos de confianza pueden obtenerse mejores resultados si se emplean las muestras bootstrap. Ver Horovitz (2001) para una explicación más completa.

CAPÍTULO 7

Análisis de regresión múltiple con información cualitativa: variables binarias (o *dummy*)

En los capítulos anteriores, la variable dependiente y las variables independientes en los modelos de regresión múltiple han tenido un significado *cuantitativo*. Algunos ejemplos son salario por hora, años de educación, promedio general de calificaciones en la universidad, cantidad de contaminación atmosférica, nivel de ventas de una empresa y número de arrestos. En cada caso, la magnitud de la variable proporciona información útil. En el trabajo empírico es necesario incluir también factores *cualitativos* en los modelos de regresión. El género o la raza de una persona, la industria de una empresa (manufactura, minorista, etc.) y la región de Estados Unidos (norte, sur, este, etc.) en la que se encuentra una ciudad son considerados factores cualitativos.

La mayor parte de este capítulo se dedica a las variables *independientes* cualitativas. Después de analizar, en la sección 7.1, la manera adecuada de describir variables cualitativas, en las secciones 7.2, 7.3 y 7.4 se describe cómo incorporar variables explicativas cualitativas a los modelos de regresión múltiple. En estas secciones se ven también casi todas las maneras más usuales de manejar variables independientes cualitativas en el análisis de regresión de corte transversal.

En la sección 7.5 se analiza una variable dependiente binaria, que es un tipo especial de variable dependiente cualitativa. En este caso, el modelo de regresión múltiple tiene una interpretación interesante y se le llama modelo de probabilidad lineal. Aunque ha sido muy difamado por muchos econometristas, la sencillez del modelo de probabilidad lineal lo hace útil en muchos contextos empíricos. Sus desventajas se describen en la sección 7.5, pero éstas suelen ser secundarias en el trabajo empírico.

7.1 Descripción de la información cualitativa

Los factores cualitativos surgen casi siempre en forma de información bivariada: una persona es mujer u hombre; una persona tiene o no computadora; una empresa ofrece o no un determinado tipo de plan de pensión a sus empleados; en un estado existe o no la pena de muerte. En todos estos ejemplos, la información que interesa puede ser captada empleando una **variable binaria** o una variable cero-uno. En econometría a las variables binarias se les suele llamar **variables binarias o *dummy***, aunque este nombre no es especialmente descriptivo.

Al definir una variable binaria hay que decidir a qué evento se le asigna el valor uno y a cuál el valor cero. Por ejemplo, en un estudio para determinar el salario de los individuos, puede definirse *female* como una variable binaria que tome el valor uno para mujer y el valor cero para hombre. En este caso, el nombre de la variable indica el evento que tiene valor uno. Esta misma información se capta definiendo *male* (hombre) igual a uno si la persona es hombre y cero si la persona es mujer. Cualquiera de éstas es mejor que emplear *gender* (género) porque este nombre

TABLA 7.1

Enumeración parcial de los datos del archivo WAGE1.RAW

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

no indica cuándo la variable binaria es uno: ¿*gender* = 1 corresponde a hombre o a mujer? Cómo se le llame a las variables no tiene importancia en la obtención de los resultados de la regresión, pero siempre ayuda elegir nombres que hagan más clara la ecuación y la exposición.

Pregunta 7.1

Suponga que, en un estudio para hacer una comparación entre los resultados de la elección para los demócratas y los republicanos, se desea indicar el partido de cada candidato. ¿Es un nombre como *party* (partido) una buena elección, en este caso, para una variable binaria? ¿Cuál sería un mejor nombre?

Suponga que en el ejemplo del salario se ha empleado *female* para indicar el género. Se define además una variable binaria *married* (casado) igual a uno si la persona está casada y cero si no es así. En la tabla 7.1 se muestra una enumeración parcial de los datos sobre salario que pueden obtenerse. Se ve que la persona 1 es mujer y que no está casada. La persona 2 es mujer y está casada. La persona 3 es hombre y no está casado y así sucesivamente.

¿Por qué se usan los valores cero y uno para describir información cualitativa? En cierto sentido, estos valores son arbitrarios: otros dos valores cualesquiera podrían servir igual. La verdadera ventaja de capturar la información cualitativa empleando variables cero-uno es que esto conduce a modelos de regresión en los que los parámetros tienen interpretaciones muy naturales, como se verá ahora.

7.2 Una sola variable binaria independiente

¿Cómo se incorpora la información binaria a los modelos de regresión? En el caso más sencillo en el que sólo hay una variable binaria explicativa, ésta simplemente se agrega a la ecuación como una variable independiente. Por ejemplo, considere el sencillo modelo siguiente para determinar el salario por hora:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \qquad \boxed{7.1}$$

Se emplea δ_0 como parámetro de *female* para resaltar la interpretación de los parámetros que multiplican a las variables binarias; más adelante se usará la notación que resulte más conveniente.

En el modelo 7.1, sólo hay dos factores que afectan al salario: el género y la educación. Como $female = 1$ si la persona es mujer y $female = 0$ si la persona es hombre, el parámetro δ_0 tiene la interpretación siguiente: δ_0 es la diferencia del salario por hora entre hombres y mujeres, *dada* una misma cantidad de educación (y un mismo término del error, u). De esta manera, el coeficiente δ_0 determina si hay discriminación en contra de las mujeres: si, para un mismo nivel de los demás factores, $\delta_0 < 0$, las mujeres ganan, en promedio, menos que los hombres.

En términos de expectativas, considerando el supuesto de media condicional cero $\mathrm{E}(u|female,educ) = 0$, entonces

$$\delta_0 = \mathrm{E}(wage|female = 1,educ) - \mathrm{E}(wage|female = 0,educ).$$

Como $female = 1$ corresponde a mujer y $female = 0$ corresponde a hombre, esto puede escribirse de manera más sencilla como

$$\delta_0 = \mathrm{E}(wage|female,educ) - \mathrm{E}(wage|male,educ). \qquad \boxed{\mathbf{7.2}}$$

Lo importante aquí es que el nivel de educación es el mismo para las dos expectativas; la diferencia, δ_0, se debe sólo al género.

Esta situación puede representarse gráficamente como un **desplazamiento del intercepto** entre hombres y mujeres. En la figura 7.1 se muestra el caso $\delta_0 < 0$, de manera que, por hora, los hombres ganan más, en una cantidad fija, que las mujeres. Esta diferencia no depende de la cantidad de educación, y esto explica por qué las líneas de salario-educación de hombres y mujeres son paralelas.

En este punto el lector se preguntará por qué no se incluye también en (7.1) una variable binaria, por ejemplo, *male* (hombres), que sea uno para hombres y cero para mujeres. Esto sería redundante. En (7.1) el intercepto para hombres es β_0 y el intercepto para mujeres es $\beta_0 + \delta_0$. Dado que hay dos grupos, sólo se necesitan dos interceptos. Esto significa que además de β_0 sólo se necesita usar *una* variable binaria; se eligió incluir una variable binaria para mujeres. Usar dos variables binarias introduciría colinealidad perfecta, ya que $female + male = 1$, lo que significa que *male* es una función lineal perfecta de *female*. Incluir variables binarias para los dos géneros es el ejemplo más sencillo de lo que se conoce como **trampa de las variables binarias**, que surge cuando demasiadas variables binarias describen una determinada cantidad de grupos. Este problema se analizará más adelante.

En (7.1) se eligió hombres como **grupo base** o **grupo de referencia** (*benchmark*), es decir, el grupo contra el que se hacen las comparaciones. A esto se debe que β_0 sea el intercepto para hombres y δ_0 sea la *diferencia* entre los interceptos para hombres y para mujeres. También podría haberse elegido mujeres como grupo base, expresando el modelo como

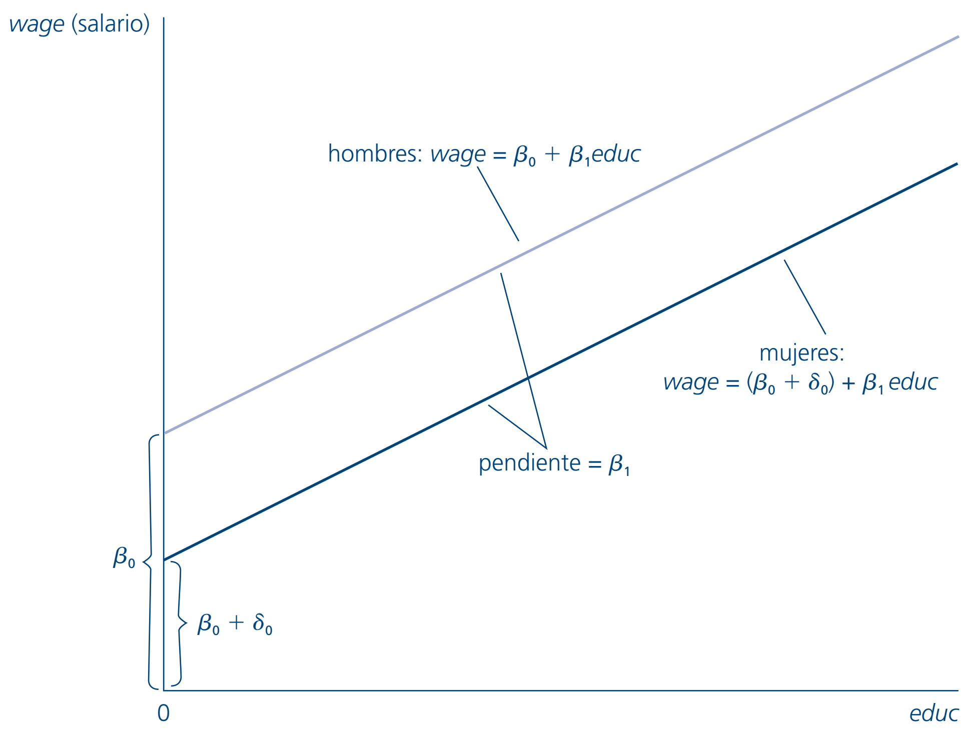
$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

donde el intercepto para mujeres es α_0 y el intercepto para hombres es $\alpha_0 + \gamma_0$; esto implica que $\alpha_0 = \beta_0 + \delta_0$ y $\alpha_0 + \gamma_0 = \beta_0$. En cualquier aplicación no tiene importancia qué grupo se elija como grupo base, pero no se debe olvidar qué grupo es el grupo base.

Algunos investigadores prefieren eliminar el intercepto general y emplear variables binarias para cada grupo. La ecuación será entonces $wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$, donde el intercepto para los hombres es β_0 y el intercepto para las mujeres es α_0. En este caso no hay trampa

FIGURA 7.1

Gráfica de $wage = \beta_0 + \delta_0 female + \beta_1 educ$ en la que $\delta_0 < 0$.



de las variables binarias, porque no se tiene un intercepto general. Sin embargo, esta formulación tiene poco que ofrecer, ya que probar la diferencia entre los interceptos es más complicado y, además, para regresiones sin intercepto, no existe un acuerdo general sobre cómo calcular la R-cuadrada. Por tanto, aquí siempre se incluirá un intercepto general para el grupo base.

Nada cambia mucho cuando intervienen más variables explicativas. Tomando hombres como grupo base, un modelo en el que, además de la educación se controle la experiencia y la antigüedad es

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u. \tag{7.3}$$

Si tanto *educ* como *exper* y *tenure* son características importantes para la productividad, la hipótesis nula de que *no* hay diferencia entre hombres y mujeres es H_0: $\delta_0 = 0$. La alternativa de que existe discriminación contra las mujeres es H_1: $\delta_0 < 0$.

¿Cómo se puede probar que existe discriminación en los salarios? La respuesta es sencilla: simplemente se estima el modelo mediante MCO, *exactamente* como antes, y se usa el estadístico t habitual. Cuando algunas de las variables independientes se definen como variables binarias no cambia nada de la mecánica de MCO ni de la teoría estadística. La única diferencia encontrada hasta ahora es la interpretación del coeficiente de la variable binaria.

Ejemplo 7.1

[Ecuación para el salario por hora]

Empleando los datos del archivo WAGE1.RAW, se estimará el modelo (7.3). Por ahora, como variable dependiente se usará *wage* y no log(*wage*):

$$\begin{aligned}\widehat{wage} = &-1.57 - 1.81\, female + .572\, educ \\ &(.72) \quad (.26) \qquad\qquad (.049) \\ &+ .025\, exper + .141\, tenure \\ &(.012) \qquad (.021) \\ n = &\ 526,\ R^2 = .364.\end{aligned} \tag{7.4}$$

El intercepto negativo —el intercepto para hombres, en este caso— no tiene mucho significado, porque en la muestra ninguna de las variables *educ*, *exper* o *tenure* (antigüedad) tiene valor cero. El coeficiente de *female* (mujer) es interesante porque mide la diferencia promedio entre el salario por hora de una mujer y de un hombre, dados los *mismos* niveles de *educ*, *exper* y *tenure*. Si se toman un hombre y una mujer con los mismos niveles de educación, experiencia y antigüedad, la mujer gana, en promedio $1.81 menos por hora que el hombre. (Recuerde que estos son salarios de 1976.)

Es importante recordar que, como se ha realizado una regresión múltiple en la que se han controlado *educ*, *exper* y *tenure*, la diferencia de $1.81 en el salario no puede ser explicada por diferencias en los niveles de educación, experiencia o antigüedad entre hombres y mujeres. Se puede concluir que tal diferencia se debe al género o a factores relacionados con el género que no han sido controlados en la regresión. [En dólares de 2003, la diferencia en el salario es aproximadamente $3.23(1.81) \approx 5.85$.]

Es interesante comparar el coeficiente de *female* en la ecuación (7.4) con la estimación que se obtiene cuando todas las demás variables explicativas se eliminan de la ecuación:

$$\begin{aligned}\widehat{wage} = &\ 7.10 - 2.51\, female \\ &\ (.21) \quad (.30) \\ n = &\ 526,\ R^2 = .116.\end{aligned} \tag{7.5}$$

Los coeficientes en (7.5) tienen una interpretación sencilla. El intercepto es el salario promedio de los hombres en la muestra (si $female = 0$), de manera que los hombres ganan, en promedio, $7.10 dólares por hora. El coeficiente de *female* es la diferencia entre el salario promedio de los hombres y el de las mujeres. Por tanto, en la muestra, el salario promedio de las mujeres es $7.10 - 2.51 = 4.59$, es decir, $4.59 por hora. (Dicho sea de paso, en la muestra hay 274 hombres y 252 mujeres.)

La ecuación (7.5) proporciona una manera sencilla de realizar una *prueba de comparación de las medias* entre dos grupos, que en este caso son hombres y mujeres. La diferencia estimada, -2.51, tiene un estadístico t de -8.37, que es estadísticamente muy significativo (y, por supuesto, que $2.51 también es económicamente una diferencia grande). En general, una regresión simple sobre una constante y una variable binaria es una manera sencilla de comparar las medias de dos grupos. Para que la prueba t habitual sea válida es necesario asumir que el supuesto de homocedasticidad se satisface, lo que significa que las varianzas poblacionales de los salarios de los hombres y de las mujeres son iguales.

La diferencia estimada entre los salarios de los hombres y de las mujeres es más grande en (7.5) que en (7.4) debido a que en (7.5) no se controlan las diferencias en educación, experiencia y antigüedad, y en promedio éstas tienen valores menores para las mujeres que para los hombres en la muestra. La ecuación (7.4) da una estimación más confiable de la brecha *ceteris paribus* entre los salarios según el género, e indica una diferencia aún muy grande.

En muchos casos, las variables binarias independientes reflejan elecciones de los individuos u otras unidades económicas (a diferencia de algo predeterminado como el género). En tales situaciones el asunto de la causalidad es de nuevo central. En el ejemplo siguiente se desea saber si tener una computadora personal es *causa* de un promedio de calificaciones superior en la universidad.

Ejemplo 7.2

[Efecto que tiene poseer una computadora sobre el promedio general de calificaciones (GPA) en la universidad]

Para determinar los efectos que poseer una computadora tiene sobre el promedio general de calificaciones se estima en el modelo

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

donde la variable binaria *PC* es uno si el estudiante posee una computadora personal y cero si no es así. Hay varias razones por las que poseer una computadora puede tener algún efecto sobre *colGPA*. Los trabajos de un estudiante pueden ser de mayor calidad si se realizan en una computadora y se puede ahorrar tiempo al no tener que esperar en un laboratorio de cómputo. Por supuesto, también un estudiante que posee una computadora puede que tienda más a jugar con los juegos de la computadora o a navegar por Internet, de manera que no es obvio que δ_0 sea positivo. Las variables *hsGPA* (promedio general de calificaciones en el bachillerato) y *ACT* (resultados en el examen de admisión) se emplean como controles: puede que sea más probable que los mejores estudiantes, de acuerdo con las puntuaciones en el *GPA* del bachillerato y en el *ACT*, tengan computadora. Estos factores se controlan porque se quiere conocer el efecto promedio sobre *colGPA* si se toma un estudiante al azar y se le da una computadora.

Empleando los datos en el archivo GPA1.RAW, se obtiene

$$\begin{aligned} \widehat{colGPA} &= \underset{(.33)}{1.26} + \underset{(.057)}{.157}\, PC + \underset{(.094)}{.447}\, hsGPA + \underset{(.0105)}{.0087}\, ACT \\ n &= 141,\ R^2 = .219. \end{aligned} \tag{7.6}$$

Esta ecuación implica que el GPA que se pronostica para un estudiante que posee una PC es aproximadamente .16 puntos más alto que el de un estudiante comparable que no tiene PC (recuerde que tanto *colGPA* como *hsGPA* se dan en una escala de cuatro puntos). Este efecto también es estadísticamente muy significativo, siendo $t_{PC} = .157/.057 \approx 2.75$.

¿Qué ocurre si de esta ecuación se eliminan *hsGPA* y *ACT*? Eliminar la última variable tendrá un efecto muy pequeño, ya que su coeficiente y su estadístico *t* son muy pequeños. Pero *hsGPA* es muy significativa y eliminarla puede afectar la estimación de β_{PC}. Regresando *colGPA* sobre *PC* se obtiene una estimación para *PC* igual aproximadamente a .170, con un error estándar de .063; en este caso, $\hat{\beta}_{PC}$ y su estadístico *t* no cambian mucho.

En los ejercicios al final del capítulo, se le pedirá que en la ecuación controle otros factores para ver si el efecto de la posesión de una computadora desaparece o si por lo menos se vuelve notablemente menor.

Cada uno de los ejemplos anteriores puede considerarse relevante en el **análisis de política**. En el primer ejemplo interesaba la discriminación de género en la fuerza de trabajo. En el segundo, el efecto de la posesión de una computadora sobre el desempeño en la universidad. Un caso especial del análisis de política es la **evaluación de programas**, en donde interesa conocer el efecto de programas económicos o sociales sobre las personas, las empresas, los vecindarios, las ciudades, etcétera.

En el caso más sencillo existen dos grupos de personas. El **grupo control** no participa en el programa; el **grupo experimental** o **grupo de tratamiento** sí. Estos nombres provienen de la literatura de las ciencias experimentales y no deben tomarse literalmente. Salvo en casos raros, tanto el grupo de control como el de tratamiento no son aleatorios. Sin embargo, en algunos casos, el análisis de regresión múltiple puede emplearse para controlar otros factores con objeto de estimar el efecto causal del programa.

Ejemplo 7.3

[Efecto de apoyos de capacitación sobre las horas de ésta]

Empleando los datos del archivo JTRAIN.RAW sobre empresas manufactureras de Michigan en 1988, se obtuvo la ecuación estimada siguiente:

$$
\begin{aligned}
\widehat{hrsemp} = {} & \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25}\, grant - \underset{(3.54)}{.98} \log(sales) \\
& - \underset{(3.88)}{6.07} \log(employ) \\
n = {} & 105,\ R^2 = .237.
\end{aligned}
\tag{7.7}
$$

La variable dependiente es horas de capacitación por empleado, al nivel de la empresa. La variable *grant* es una variable binaria igual a uno si en 1988 la empresa recibió una subvención para capacitación e igual a cero si no fue así. Las variables *sales* y *employ* corresponden a ventas anuales y a cantidad de empleados, respectivamente. No es posible emplear *hrsemp* en forma logarítmica, debido a que para 29 de las 105 empresas usadas en la regresión *hrsemp* es cero.

La variable *grant* es estadísticamente muy significativa, siendo $t_{grant} = 4.70$. Controlando ventas (*sales*) y empleados (*employ*), las empresas que recibieron una subvención dieron a cada trabajador, una capacitación de 26.25 horas más, en promedio. Dado que la cantidad promedio de horas de capacitación por empleado en la muestra es aproximadamente 17, siendo el valor máximo 164, *grant* (subvención) tiene un efecto importante sobre la capacitación, como era de esperarse.

El coeficiente de log(*sales*) es pequeño y muy poco significativo. El coeficiente de log(*employ*) significa que, si una empresa es 10% mayor, capacitará a sus empleados .61 hora menos. Su estadístico t es -1.56, que es sólo marginal en términos estadísticos significativos.

Como ocurre con cualquier otra variable independiente, es necesario preguntarse si el efecto de una variable cualitativa es causal. En la ecuación (7.7), ¿la diferencia en capacitación entre las empresas que recibieron subvención y las que no se debe a la subvención, o la recepción de la subvención es simplemente un indicador de algo más? Puede ser que las empresas que recibieron la subvención, de cualquier manera, en promedio, hubieran capacitado más a sus empleados aun sin subvención. No hay nada en este análisis que indique que se ha estimado un efecto causal; es necesario saber cómo se eligieron las empresas para que recibieran la subvención. Sólo se puede esperar que se hayan controlado tantos factores como sea posible relacionados con que la empresa haya recibido una subvención y con su nivel de capacitación.

En la sección 7.6, así como en capítulos posteriores, se volverá a ver el análisis de políticas con variables binarias.

Interpretación de los coeficientes de variables explicativas binarias cuando la variable dependiente es log(*y*)

En una especificación usual en el trabajo práctico, la variable dependiente aparece en forma logarítmica y una o más variables binarias aparecen como variables independientes. ¿Cómo se interpretan, en este caso, los coeficientes de las variables binarias? No sorprenderá que los coeficientes tengan una interpretación *porcentual*.

Ejemplo 7.4

[Regresión para el precio de la vivienda]

Empleando los datos del archivo HPRICE1.RAW, se obtiene la ecuación

$$\begin{aligned}\widehat{\log(price)} = \underset{(.65)}{-1.35} + \underset{(.038)}{.168} \log(lotsize) + \underset{(.093)}{.707} \log(sqrft) \\ + \underset{(.029)}{.027}\, bdrms + \underset{(.045)}{.054}\, colonial \\ n = 88, R^2 = .649.\end{aligned} \tag{7.8}$$

Todas las variables se explican por sí mismas, excepto *colonial*, que es una variable binaria igual a uno si la casa es de estilo colonial. ¿Qué significa el coeficiente de *colonial*? Para valores dados de *lotsize*, *sqrft* y *bdrms*, la diferencia en $\widehat{\log(price)}$ entre una casa de estilo colonial y una de otro estilo es .054. Esto significa que se predice que una casa de estilo colonial se venderá en aproximadamente 5.4% más, manteniendo constantes todos los demás factores.

Este ejemplo muestra que cuando en un modelo la variable dependiente es log(y) el coeficiente de una variable binaria, después de multiplicarlo por 100, se interpreta como la diferencia porcentual en y, manteniendo todos los demás factores constantes. Cuando el coeficiente de una variable binaria indica un cambio proporcional grande en y, la diferencia porcentual exacta puede obtenerse exactamente, como en el caso del cálculo de la semielasticidad en la sección 6.2.

Ejemplo 7.5

[Ecuación del logaritmo del salario por hora]

Se reestimará la ecuación para el salario, del ejemplo 7.1, empleando log(*wage*) como variable dependiente y agregando términos cuadráticos en *exper* y *tenure* (antigüedad):

$$\begin{aligned}\widehat{\log(wage)} = \underset{(.099)}{.417} - \underset{(.036)}{.297}\, female + \underset{(.007)}{.080}\, educ + \underset{(.005)}{.029}\, exper \\ - \underset{(.00010)}{.00058}\, exper^2 + \underset{(.007)}{.032}\, tenure - \underset{(.00023)}{.00059}\, tenure^2 \\ n = 526, R^2 = .441.\end{aligned} \tag{7.9}$$

Empleando el mismo método que en el ejemplo 7.4, el coeficiente de *female* (mujer) implica que dados los mismos valores de *educ*, *exper* y *tenure*, las mujeres ganan aproximadamente 100(.297) = 29.7% menos que los hombres. Este resultado se puede mejorar calculando la diferencia porcentual exacta entre los salarios predichos. Lo que se quiere es la diferencia proporcional entre los salarios de las mujeres y de los hombres, manteniendo todos los demás factores constantes: $(\widehat{wage}_F - \widehat{wage}_M)/\widehat{wage}_M$. Lo que se tiene, de acuerdo con (7.9), es

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -.297.$$

Exponenciando y restando uno se obtiene

$$(\widehat{wage}_F - \widehat{wage}_M)/\widehat{wage}_M = \exp(-.297) - 1 \approx -.257.$$

Esta estimación más exacta implica que el salario de una mujer es, en promedio, 25.7% inferior al salario comparable de un hombre.

Si en el ejemplo 7.4 se hace la misma corrección se obtiene exp(.054) − 1 ≈ .0555, es decir, aproximadamente 5.6%. Esta corrección tiene un efecto menor en el ejemplo 7.4 que en el ejemplo del salario, debido a que la magnitud del coeficiente de la variable binaria es mucho menor en (7.8) que en (7.9).

En general, si $\hat{\beta}_1$ es el coeficiente de una variable binaria, por ejemplo x_1, siendo log(y) la variable dependiente, la diferencia porcentual exacta en la y predicha para $x_1 = 1$ versus $x_1 = 0$ es

$$100 \cdot [\exp(\hat{\beta}_1) - 1]. \tag{7.10}$$

La estimación $\hat{\beta}_1$ puede ser positiva o negativa, y es importante preservar su signo al calcular (7.10).

El método logarítmico de aproximación tiene la ventaja de proporcionar una estimación entre las magnitudes obtenidas empleando cada grupo como grupo base. En particular, aunque la ecuación (7.10) da una estimación mejor que $100 \cdot \hat{\beta}_1$ del porcentaje en el que y para $x_1 = 1$ es mayor que y para $x_1 = 0$, (7.10) no es una buena estimación si se cambia el grupo base. En el ejemplo 7.5 se puede estimar el porcentaje en el que el salario de un hombre es superior al salario comparable de una mujer y esta estimación es $100 \cdot [\exp(-\hat{\beta}_1) - 1] = 100 \cdot [\exp(.297) - 1] \approx 34.6$. La aproximación, basada en $100 \cdot \hat{\beta}_1$, 29.7, se encuentra entre 25.7 y 34.6 (y cercano a la mitad). Por tanto, es razonable decir que "la diferencia que se predice entre los salarios de hombres y mujeres es aproximadamente 29.7%", sin tener que decir cuál es el grupo base.

7.3 Uso de variables binarias en categorías múltiples

En una misma ecuación pueden emplearse varias variables independientes binarias. Por ejemplo, a la ecuación (7.9) se le puede agregar la variable binaria *married* (casado). El coeficiente de *married* da la diferencia proporcional (aproximada) entre los salarios de los casados y de los solteros, manteniendo constantes género, *educ*, *exper* y *tenure* (antigüedad). Cuando se estima este modelo, el coeficiente de *married* (dando el error estándar entre paréntesis) es .053 (.041), y el coeficiente de *female* se convierte en −.290 (.036). Por tanto, se estima que la "prima de casado" es de aproximadamente 5.3%, pero no es estadísticamente distinta de cero ($t = 1.29$). Una limitación importante de este modelo es que se supone que la prima de casado es la misma para hombres que para mujeres; esto se soluciona en el ejemplo siguiente.

Ejemplo 7.6

[Ecuación para el logaritmo del salario por hora]

Ahora se estimará un modelo que toma en cuenta las diferencias entre cuatro grupos: hombres casados, mujeres casadas, hombres solteros y mujeres solteras. Para esto, se debe elegir un grupo base; se elige hombres solteros. Después se define una variable binaria para cada uno de los grupos restantes. Llámesele a estas variables *marrmale*, *marrfem* y *singfem*. Introduciendo estas tres variables en (7.9) (y, por supuesto, eliminando *female*, ya que ahora es redundante) se obtiene

$$
\begin{aligned}
\widehat{\log(wage)} = {} & \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem \\
& - \underset{(.056)}{.110}\, singfem + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054}\, exper^2 \\
& + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053}\, tenure^2 \\
& n = 526,\ R^2 = .461.
\end{aligned}
\tag{7.11}
$$

Todos los coeficientes, excepto *singfem*, tienen un estadístico t bastante mayor a dos, en valor absoluto. El estadístico t para *singfem* es aproximadamente -1.96, que apenas es significativo al nivel de 5% contra la alternativa de dos colas.

Para interpretar los coeficientes de las variables binarias, hay que recordar que el grupo base es hombres solteros. De esta manera, las estimaciones para las tres variables binarias miden la diferencia proporcional en el salario *con relación* a los hombres solteros. Por ejemplo, se estima que, manteniendo constantes los niveles de educación, experiencia y antigüedad, los hombres casados ganan aproximadamente 21.3% más que los solteros. [La estimación más precisa que se obtiene con (7.10) es aproximadamente 23.7%.] Una mujer casada, por otro lado, se predice que gana 19.8% menos que un hombre soltero siendo los niveles de otras variables los mismos.

Como en (7.11) el grupo base está representado por el intercepto, sólo se han incluido variables binarias para tres de los cuatro grupos. Si en (7.11) se introdujera una variable binaria para hombres casados se caería en la trampa de la variable binaria porque se introduciría colinealidad perfecta. Algunos paquetes para regresión corrigen, de manera automática, este error, mientras que otros simplemente indican que hay colinealidad perfecta. Lo mejor es especificar con cuidado las variables binarias, porque entonces se está forzando a interpretar de forma adecuada el modelo final.

Aunque en (7.11) el grupo base es hombres solteros, esta ecuación puede usarse para obtener la diferencia estimada entre cualesquiera dos de los grupos. Como el intercepto general es común a todos los grupos, ésta puede ignorarse al hallar las diferencias. Así, la diferencia proporcional estimada entre las mujeres solteras y casadas es $-.110 - (-.198) = .088$, lo que significa que las mujeres solteras ganan aproximadamente 8.8% más que las casadas. Por desgracia, la ecuación (7.11) no puede emplearse para probar si la diferencia estimada entre mujeres solteras y casadas es estadísticamente significativa. Conocer los errores estándar para *marrfem* y *singfem* no es suficiente para llevar a cabo esta prueba (vea la sección 4.4). Lo más fácil es elegir uno de estos grupos como grupo base y volver a estimar la ecuación. Con esto no cambia nada importante, pero la estimación buscada y su error estándar se obtienen de manera directa

$$
\widehat{\log(wage)} = \underset{(.106)}{.123} + \underset{(.056)}{.411}\, marrmale + \underset{(.058)}{.198}\, singmale + \underset{(.052)}{.088}\, singfem + \ldots,
$$

donde, por supuesto, ninguno de los coeficientes o errores estándar que no se reportan han cambiado. La estimación de *singfem* es, como se esperaba, .088. Ahora, se tiene un error estándar que corresponde a esta estimación. El estadístico t para la hipótesis nula de que en la población no hay diferencia entre las mujeres casadas y solteras es $t_{singfem} = .088/.052 \approx 1.69$. Esta es una evidencia marginal contra la hipótesis nula. Se observa también que la diferencia estimada entre hombres casados y mujeres casadas es estadísticamente muy significativa ($t_{marrmale} = 7.34$).

El ejemplo anterior ilustra un principio general para la inclusión de variables binarias para indicar grupos diferentes: si el modelo de regresión ha de tener interceptos diferentes para, por ejemplo, g grupos o categorías, en el modelo se deberán incluir $g - 1$ variables binarias y un intercepto. El intercepto correspondiente al grupo base es el intercepto general del modelo, y el coeficiente de la variable binaria de un determinado grupo representa la diferencia estimada entre el intercepto de ese grupo y el grupo base. Incluir g variables binarias y un intercepto dará como resultado la trampa de la variable binaria. Una alternativa es incluir g variables binarias y eliminar el intercepto general. Algunas veces es útil incluir g variables binarias sin un intercepto general, pero tiene dos desventajas prácticas. Primero, hace más complicado probar diferencias en relación con un grupo base. Segundo, cuando no se incluye un intercepto general, los paquetes para regresión suelen modificar la manera en que calculan R-cuadrada. En particular, en la fórmula $R^2 = 1 - \text{SRC/STC}$, la suma total de cuadrados, STC, es sustituida por una suma total de cuadrados que no centra las y_i en torno a su media, por ejemplo, $\text{STC}_0 = \sum_{i=1}^{n} y_i^2$. A la R-cuadrada que se obtiene, por ejemplo $R_0^2 = 1 - \text{SRC/STC}_0$, se le suele llamar ***R*-cuadrada descentrada**. Por desgracia, R_0^2 pocas veces es adecuada como una medida de la bondad de ajuste. Siempre se tiene $\text{STC}_0 \geq \text{STC}$, presentándose la igualdad sólo cuando $\bar{y} = 0$. Con frecuencia STC_0 es mucho más grande que STC, lo que hace que R_0^2 sea mucho más grande que R^2. Por ejemplo, si en el ejemplo anterior se regresa log(*wage*) sobre *marrmale*, *singmale*, *marrfem*, *singfem* y las demás variables explicativas —sin intercepto— la R-cuadrada que se obtiene con Stata, que es R_0^2, es .948. Esta R-cuadrada elevada es un error por no centrar la suma total de cuadrados en los cálculos. En la ecuación (7.11) se da la R-cuadrada correcta, que es .461. Algunos paquetes para regresión, como Stata, tienen una opción para hacer que se calcule la R-cuadrada centrada aun cuando no se haya incluido un intercepto general, y casi siempre es aconsejable usar esta opción. En la inmensa mayoría de los casos, en cualquier R-cuadrada que se base en la comparación de una SRC y una STC, la STC deberá haber sido calculada centrando las y_i en torno a $\bar{y}$. Esta STC puede entenderse como la suma de los cuadrados residuales que se obtiene si se usa la media muestral, $\bar{y}$, para predecir todas las y_i. Desde luego, se puede esperar muy poco de un modelo en el que lo único que se mide es su ajuste en relación con el uso de una constante como predictor. En un modelo sin intercepto que tenga un mal ajuste, es posible que $\text{SRC} > \text{STC}$, lo que significa que R^2 será negativa. La R-cuadrada descentrada estará siempre entre cero y uno, lo que posiblemente explica por qué suele ser la que se emplea si no se indica otra cosa cuando en los modelos de regresión no se estima un intercepto.

Pregunta 7.2

En la base de datos del archivo MLB1.RAW, sobre los salarios en el béisbol, a los jugadores se les asigna una de seis posiciones: *frstbase* (primera base), *scndbase* (segunda base), *thrdbase* (tercera base), *shrtstop* (parador en corto), *outfield* (jardinero) o *catcher* (receptor). ¿Cuáles son las variables binarias que deben incluirse como variables independientes en el modelo para considerar las diferencias de salario entre estas posiciones?

Incorporación de información ordinal mediante el uso de variables binarias

Suponga que se desea estimar el efecto de la calificación crediticia de la ciudad sobre las tasas de interés de los bonos municipales (*MBR*). Varias compañías financieras, por ejemplo Moody's Investors Service and Standard Poor's, califican la calidad de la deuda de los gobiernos locales, dependiendo de la calificación de cosas tales como la probabilidad de incumplimiento. (Los gobiernos locales prefieren tasas de interés bajas con objeto de reducir los costos de sus préstamos.) Para simplificar, suponga que las calificaciones van de cero a cuatro, siendo cero la peor calificación crediticia y cuatro la mejor. Este es un ejemplo de una **variable ordinal**. Llámesele a esta variable *CR*. La pregunta que hay que responder es: ¿Cómo se incorpora *CR* en un modelo que explique *MBR*?

Una posibilidad es incluir CR como se incorporaría cualquier otra variable explicativa:

$$MBR = \beta_0 + \beta_1 CR + otros\ factores,$$

donde no se muestra explícitamente qué otros factores están en el modelo. Entonces β_1 es la variación de MBR, en puntos porcentuales, cuando CR aumenta una unidad, permaneciendo todos los demás factores constantes. Por desgracia, es bastante difícil interpretar un aumento de CR de una unidad. Se conoce el significado cuantitativo de un año más de educación o de un dólar más que se gasta por estudiante, pero cosas como las calificaciones crediticias suelen tener sólo un significado ordinal. Se sabe que una CR de cuatro es mejor que una CR de tres, pero ¿es la diferencia entre cuatro y tres igual a la diferencia entre uno y cero? Si no es así, entonces no tiene sentido suponer que un aumento de una unidad en CR tiene un efecto constante sobre MBR.

Una mejor idea, que se puede emplear dado que CR toma relativamente pocos valores, es definir variables binarias para cada valor de CR. Así, sea $CR_1 = 1$ si $CR = 1$ y $CR_1 = 0$ si no es así; $CR_2 = 1$ si $CR = 2$ y $CR_2 = 0$ si no es así; etc. En efecto, se toma la calificación crediticia y se convierte en cinco categorías. Así se puede estimar el modelo

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + otros\ factores. \tag{7.12}$$

Siguiendo la regla para la inclusión de variables binarias en el modelo, se incluyen cuatro de estas variables, porque se tienen cinco categorías. La categoría omitida aquí es la calificación crediticia de cero, que de esta manera es el grupo base. (Esta es la razón por la que no se tiene que definir una variable binaria para esta categoría.) Los coeficientes tienen una interpretación sencilla: δ_1 es la diferencia en MBR (permaneciendo los demás factores constantes) entre una municipalidad con una calificación crediticia de uno y una con una calificación crediticia de cero; δ_2 es la diferencia en MBR entre una municipalidad con una calificación crediticia de dos y una con una calificación crediticia de cero, etc. Se ha permitido que el movimiento entre cada dos calificaciones crediticias tenga efectos diferentes, con lo que usar (7.12) es mucho más flexible que incluir CR como una sola variable. Una vez que se han definido estas variables binarias, es sencillo estimar (7.12).

Pregunta 7.3

En el modelo (7.12), ¿cómo probaría la hipótesis nula de que la calificación crediticia no tiene efecto en *MBR*?

La ecuación (7.12) contiene, como caso especial, el modelo en el que el efecto parcial es constante. Una manera de expresar las tres restricciones para que impliquen un efecto parcial constante es $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$ y $\delta_4 = 4\delta_1$. Sustituyendo en la ecuación (7.12) y reordenando, se obtiene $MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + otros\ factores$. Ahora, el término que multiplica a δ_1 es simplemente la variable original para la calificación del crédito, CR. Para obtener el estadístico F para probar las restricciones de efecto parcial constante, se obtiene la R-cuadrada no restringida de (7.12) y la R-cuadrada restringida de la regresión de MBR sobre CR y los demás factores que se han controlado. El estadístico F se obtiene como en la ecuación (4.41) con $q = 3$.

Ejemplo 7.7

[Efectos del atractivo físico sobre el salario]

Hamermesh y Biddle (1944) emplearon mediciones del atractivo físico en una ecuación de salario. (El archivo BEAUTY.RAW contiene menos variables, pero más observaciones de las usadas por Hamermesh y Biddle.) Un entrevistador dio a cada una de las personas de la muestra una calificación de acuerdo con su atractivo físico, empleando cinco categorías (feo, poco atractivo, regular, bien parecido, hermoso o guapo). Dado que en los dos extremos hay muy pocas personas, para el análisis de regresión los autores colocaron a las personas en tres grupos: promedio, superior al promedio (*abvavg*) e inferior al promedio (*belavg*), donde el grupo base es promedio (*average*). Empleando los datos de la encuesta Quality of Employment Survey de 1977, después de controlar las características de productividad habituales, Hamermesh y Biddle estimaron una ecuación para hombres:

$$\begin{aligned} \widehat{\log(wage)} &= \hat{\beta}_0 - \underset{(.046)}{.164}\, belavg + \underset{(.033)}{.016}\, abvavg + otros\ factores \\ n &= 700, \bar{R}^2 = .403 \end{aligned}$$

y una para mujeres:

$$\begin{aligned} \widehat{\log(wage)} &= \hat{\beta}_0 - \underset{(.066)}{.124}\, belavg + \underset{(.049)}{.035}\, abvavg + otros\ factores \\ n &= 409, \bar{R}^2 = .330. \end{aligned}$$

Los otros factores controlados en la regresión son educación, experiencia, antigüedad, estado civil y raza; para una lista más completa vea la tabla 3 en el artículo de Hamermesh y Biddle. Para ahorrar espacio, en ese artículo no se dan los coeficientes de las otras variables ni tampoco el intercepto.

En el caso de los hombres, se estima que aquellos cuya apariencia es inferior al promedio ganan aproximadamente 16.4% menos que un hombre con una apariencia promedio e igual en los demás aspectos (educación, experiencia, antigüedad, estado civil y raza). El efecto es estadísticamente distinto de cero con $t = -3.57$. De manera similar, los hombres con un aspecto superior al promedio ganan un estimado de 1.6% más, aunque este efecto no es estadísticamente significativo ($t < .5$).

Una mujer con una apariencia inferior al promedio gana aproximadamente 12.4% menos que una mujer con una apariencia promedio y por lo demás comparable a ella, con $t = -1.88$. Como ocurrió con los hombres, la estimación para *abvavg* no es estadísticamente distinta de cero.

Hay casos en los que una variable ordinal toma demasiados valores, por lo que no puede incluirse una variable binaria para cada valor. Por ejemplo, el archivo LAWSCH85.RAW contiene datos sobre los salarios iniciales medianos de egresados de escuelas de leyes. Una de las variables explicativas clave es la calificación dada a la escuela de leyes. Como cada escuela tiene una calificación distinta, es claro que no puede incluirse una variable binaria para cada calificación. Si no se quiere incluir directamente la calificación en la ecuación, las calificaciones pueden dividirse en categorías. En el ejemplo siguiente se muestra cómo se hace esto.

Ejemplo 7.8

[Efectos del ranking de una escuela de leyes sobre el salario inicial]

Se definen las variables binarias *top10*, *r11_25*, *r26_40*, *r41_60*, *r61_100* que toman el valor uno cuando la variable *rank* (ranking) cae dentro del rango correspondiente. Como grupo base se consideran las escuelas con un ranking que no esté entre las 100 mejores. La ecuación estimada es

$$\begin{aligned}\widehat{\log(salary)} = {} & \underset{(.41)}{9.17} + \underset{(.053)}{.700}\, top10 + \underset{(.039)}{.594}\, r11\_25 + \underset{(.034)}{.375}\, r26\_40 \\ & + \underset{(.028)}{.263}\, r41\_60 + \underset{(.021)}{.132}\, r61\_100 + \underset{(.0031)}{.0057}\, LSAT \\ & + \underset{(.074)}{.014}\, GPA + \underset{(.026)}{.036}\log(libvol) + \underset{(.0251)}{.0008}\log(cost) \\ n = {} & 136,\ R^2 = .911,\ \bar{R}^2 = .905.\end{aligned} \tag{7.13}$$

Se ve inmediatamente que todas las variables binarias que definen los diferentes rangos son estadísticamente muy significativas. La estimación para *r61_100* significa que, manteniendo constantes *LSAT*, *GPA*, *libvol* y *cost*, el salario medio en una escuela de leyes clasificada entre 61 y 100 es aproximadamente 13.2% superior que en una escuela de leyes que no esté clasificada entre las 100 mejores. La diferencia entre una escuela de las 10 mejores y una escuela no clasificada entre las 100 mejores es bastante grande. Empleando el cálculo exacto dado en la ecuación (7.10) se obtiene $\exp(.700) - 1 \approx 1.014$, con lo que el salario medio que se predice es más de 100% superior en las 100 mejores escuelas que en una de las escuelas no clasificada entre las 100 mejores.

Como indicación de si el dividir el ranking en categorías representa una mejora, se puede comparar la R-cuadrada ajustada de (7.13) con la R-cuadrada ajustada cuando se incluye *rank* (ranking) como una sola variable: la primera es .905 y la última es .836, de manera que está justificada la mayor flexibilidad de (7.13).

Es interesante observar que una vez que las posiciones del ranking se colocan en las (algo arbitrarias) categorías dadas, todas las demás variables se vuelven poco significativas. Una prueba de significancia conjunta para *LSAT*, *GPA*, log(*libvol*) y log(*cost*) da un valor-p de .055, el cual es apenas significativo. Cuando se incluye *rank* en su forma original, el valor-p para la significancia conjunta es cero a cuatro posiciones decimales.

Un último comentario acerca de este ejemplo. Al obtener las propiedades de mínimos cuadrados ordinarios, se asumió que se tenía una muestra aleatoria. En esta aplicación se viola ese supuesto debido a la forma en que se define *rank*: la posición en el ranking de una escuela depende necesariamente de las posiciones de las demás escuelas de la muestra y siendo así, los datos no pueden representar muestreos independientes de la población de todas las escuelas de leyes. Esto no ocasiona ningún problema serio, siempre y cuando el término de error no esté correlacionado con las variables explicativas.

7.4 Interacciones en las que intervienen variables binarias

Interacciones entre variables binarias

Así como variables con un significado cuantitativo pueden estar relacionadas en los modelos de regresión, también pueden estarlo las variables binarias. En realidad ya vimos esto en el ejemplo 7.6, en donde se definieron cuatro categorías con base en el estado civil y en el género. En efecto, este modelo puede retomarse agregando un **término de interacción** entre *female* (mujer) y *married* (casado) y donde *female* y *married* aparezcan por separado. Esto permite que la prima de casado dependa del género, como lo hace en la ecuación (7.11). Para propósitos de comparación, el modelo estimado con el término de interacción *female-married* es

$$\begin{aligned}\widehat{\log(wage)} = {} & \underset{(.100)}{.321} - \underset{(.056)}{.110}\, female + \underset{(.055)}{.213}\, married \\ & - \underset{(.072)}{.301}\, female \cdot married + \ldots,\end{aligned} \tag{7.14}$$

donde el resto de la regresión es necesariamente idéntico a (7.11). La ecuación (7.14) muestra de manera explícita que hay una interacción estadísticamente significativa entre el género y el estado civil. Este modelo también permite obtener la diferencia estimada entre los salarios de los cuatro grupos, pero aquí hay que tener cuidado al sustituir las combinaciones correctas de ceros y unos.

La combinación $female = 0$ y $married = 0$ corresponde al grupo de los hombres solteros, que es el grupo base, ya que esto elimina *female*, *married* y *female·married*. El intercepto para hombres casados puede determinarse haciendo en (7.14) $female = 0$ y $married = 1$; esto da un intercepto de $.321 + .213 = .534$, y así sucesivamente.

La ecuación (7.14) es sólo otra manera de encontrar las diferencias de salario entre las distintas combinaciones de género y estado civil. Esta ecuación permite probar fácilmente la hipótesis nula de que las diferencias debidas al género no dependen del estado civil (equivalentemente, que las diferencias debidas al estado civil no dependen del género). Para probar las diferencias en los salarios entre cualquier grupo y el grupo base de los hombres solteros, la ecuación (7.11) es más conveniente.

Ejemplo 7.9

[Efectos del manejo de la computadora sobre el salario]

Krueger (1993) estima el efecto del manejo de la computadora sobre los salarios. Krueger define una variable binaria a la que se le llamará aquí *compwork*, que es igual a uno si el individuo usa la computadora en su trabajo. Otra variable binaria es *comphome*, que es igual a uno si la persona usa la computadora en casa. Empleando 13,379 personas del censo de población de 1989. Krueger (1993, tabla 4) obtuvo

$$\begin{aligned}\widehat{\log(wage)} = \hat{\beta}_0 &+ \underset{(.009)}{.177}\, compwork + \underset{(.019)}{.070}\, comphome \\ &+ \underset{(.023)}{.017}\, compwork \cdot comphome + otros\ factores.\end{aligned} \tag{7.15}$$

(En otros factores están comprendidos los factores estándar de las regresiones para el salario, tales como educación, experiencia, género y estado civil; la lista exacta se puede encontrar en el artículo de Krueger.) Krueger no reporta el intercepto debido a que éste no tiene ninguna importancia; todo lo que se necesita saber es que el grupo base consta de las personas que no usan la computadora en casa o en el trabajo. Vale la pena observar que el rendimiento estimado del uso de la computadora en el trabajo (pero no en casa) es aproximadamente 17.7%. (Una estimación más precisa es 19.4%.) De manera similar, las personas que usan la computadora en casa, pero no en el trabajo tienen un salario 7% mayor a las que en absoluto usan la computadora. La diferencia entre los que emplean la computadora en ambos lugares y los que no la usan en ningún lugar es aproximadamente 26.4% (que se obtiene sumando los tres coeficientes y multiplicando por 100), o la estimación más precisa que es 30.2%, que se obtiene empleando la ecuación (7.10).

El término de interacción en (7.15) no es estadísticamente muy significativo ni tampoco es muy grande económicamente. Pero causa poco daño que esté en la ecuación.

Considerar pendientes diferentes

Se han visto varios ejemplos de cómo considerar, en un modelo de regresión múltiple, diversos interceptos para cualquier cantidad de grupos. También hay ocasiones en las que variables binarias interactuando con variables explicativas, que no son variables binarias, permiten considerar

diferencias en las pendientes. Continuando con el ejemplo del salario, suponga que se desea saber si el rendimiento de la educación es el mismo para los hombres y para las mujeres, considerando una diferencia constante entre los salarios de hombres y de mujeres (una diferencia de la que ya previamente se han encontrado evidencias). Para simplificar, en este modelo sólo se incluyen educación y género. ¿Qué tipo de modelo permite rendimientos diferentes de la educación? Considere el modelo

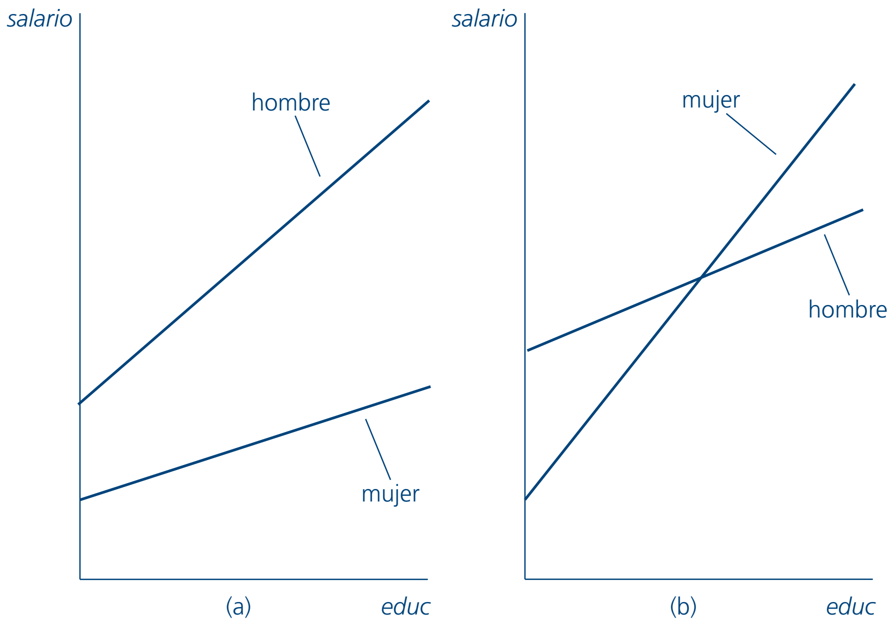
$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u. \quad \boxed{7.16}$$

Si en (7.16) se introduce $female = 0$ se encuentra que el intercepto para los hombres es β_0 y la pendiente para la educación de los hombres es β_1. Cuando se trata de las mujeres, se sustituye $female = 1$; entonces, el intercepto en el caso de las mujeres es $\beta_0 + \delta_0$ y la pendiente es $\beta_1 + \delta_1$. Por tanto, δ_0 mide la diferencia entre los interceptos para mujeres y hombres, y δ_1 mide la diferencia entre los rendimientos de la educación de mujeres y hombres. En la figura 7.2 se muestran dos de los cuatro casos posibles para los signos de δ_0 y δ_1.

En la gráfica a) se muestra el caso en que el intercepto correspondiente a las mujeres es menor que el correspondiente a los hombres y la pendiente de la recta es menor en el caso de las mujeres que en el caso de los hombres. Esto significa que en todos los niveles de educación las mujeres ganan menos que los hombres, y que esta diferencia aumenta a medida que *educ* es mayor. En la gráfica b), el intercepto correspondiente a las mujeres es menor que el correspondiente a los hombres, pero la pendiente de la educación es mayor en el caso de las mujeres. Esto significa que a niveles de educación bajos, las mujeres ganan menos que los hombres, pero a medida que la educación es mayor, esta brecha se reduce. En algún punto, dado un

FIGURA 7.2

Gráficas de la ecuación (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



mismo nivel de educación, una mujer gana más que un hombre (y dada la ecuación estimada, este punto se encuentra fácilmente).

¿Cómo se puede estimar el modelo (7.16)? Para aplicar MCO, el modelo debe expresarse con una interacción entre *female* y *educ*:

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u. \tag{7.17}$$

Ahora los parámetros pueden estimarse mediante la regresión de log(*wage*) sobre *female*, *educ* y *female·educ*. Con cualquier paquete para regresión es fácil obtener el término de interacción. No debe desalentar lo extraño de *female·educ*, que es cero para todos los hombres de la muestra e igual al nivel de educación para cada mujer de la muestra.

Una hipótesis importante es que el rendimiento de la educación es el mismo para hombres y mujeres. En términos del modelo (7.17), esto se expresa como H_0: $\delta_1 = 0$, que significa que la pendiente de log(*wage*) respecto a *educ* es igual para hombres y mujeres. Observe que esta hipótesis no pone ninguna restricción a la diferencia en los interceptos, δ_0. De acuerdo con esta hipótesis, puede haber diferencia entre los salarios de hombres y mujeres, pero esta diferencia debe ser la misma para todos los niveles de educación. Esta situación se describe en la figura 7.1.

Otra hipótesis interesante es que los salarios promedio son idénticos para hombres y mujeres con un mismo nivel de educación. Esto significa que, de acuerdo con la hipótesis nula *tanto* δ_0 *como* δ_1 deben de ser cero. En la ecuación (7.17) debe emplearse una prueba *F* para probar H_0: $\delta_0 = 0$, $\delta_1 = 0$. En el modelo donde sólo el intercepto es diferente, esta hipótesis se rechaza porque H_0: $\delta_0 = 0$ se rechaza claramente frente a H_1: $\delta_0 < 0$.

Ejemplo 7.10

[Ecuación para el logaritmo del salario por hora]

A (7.17) se le agregan ahora términos cuadráticos en experiencia y antigüedad (*tenure*):

$$\begin{aligned}
\widehat{\log(wage)} = {} & \underset{(.119)}{.389} - \underset{(.168)}{.227}\, female + \underset{(.008)}{.082}\, educ \\
& - \underset{(.0131)}{.0056}\, female \cdot educ + \underset{(.005)}{.029}\, exper - \underset{(.00011)}{.00058}\, exper^2 \\
& + \underset{(.007)}{.032}\, tenure - \underset{(.00024)}{.00059}\, tenure^2 \\
n = {} & 526,\ R^2 = .441.
\end{aligned} \tag{7.18}$$

En esta ecuación el rendimiento estimado de la educación, para los hombres, es .082, es decir, 8.2%. Para las mujeres, se tiene $.082 - .0056 = .0764$, es decir, aproximadamente 7.6%. La diferencia, $-.56\%$, es decir, poco más de medio punto porcentual para las mujeres, no es grande desde el punto de vista económico y tampoco es estadísticamente significativa: el estadístico *t* es $-.0056/.0131 \approx -.43$. De esta manera, se concluye que no hay evidencias contra la hipótesis de que el rendimiento de la educación sea el mismo para hombres y para mujeres.

El coeficiente de *female*, aunque sigue siendo grande desde el punto de vista económico, ya no es tan significativo a los niveles convencionales ($t = -1.35$). En la ecuación sin interacción, su coeficiente y su estadístico *t* eran $-.297$ y -8.25, respectivamente [vea la ecuación (7.9)]. Debe ahora concluirse que, dados los mismos niveles de *educ*, *exper* y *tenure*, ¿no hay evidencias estadísticamente significativas de que la remuneración de las mujeres sea inferior a la de los hombres? Este sería un serio error. Debido a que a la ecuación se le agregó la interacción *female·educ*, el coeficiente de *female* se estima ahora de una manera mucho menos precisa que en la ecuación (7.9), el error estándar ha aumentado y casi se ha quintuplicado (.168/.036 ≈ 4.67). Esto se debe a que, en la muestra *female* y *female·educ* están fuertemente correlacionadas. En este ejemplo hay una manera útil de ver la multicolinealidad: en la ecuación (7.17) y en la ecuación más general estimada (7.18), δ_0 mide la diferencia entre los salarios de hombres y mujeres cuando *educ* = 0. En la muestra muy pocas personas tienen niveles de educación muy bajos, de manera que no debe extrañar que resulte difícil estimar la diferencia en *educ* = 0 (tampoco es que la diferencia correspondiente a cero años de educación sea muy informativa). Más interesante será estimar la diferencia entre los géneros, por ejemplo, al nivel de educación promedio en la muestra (aproximadamente 12.5). Para esto hay que sustituir *female·educ* por *female*·(*educ* − 12.5) y volver a correr la regresión: esto sólo hace que cambie el coeficiente de *female* y su error estándar. (Vea el ejercicio para computadora C7.7.)

Si se calcula el estadístico F para H_0: $\delta_0 = 0$, $\delta_1 = 0$, se obtiene $F = 34.33$, que es un valor enorme para una variable aleatoria F con $gl = 2$ en el numerador y $gl = 518$: en el denominador: el valor-p es cero a cuatro cifras decimales. Al final, se prefiere el modelo (7.9), que considera una diferencia de salario constante entre hombres y mujeres.

Pregunta 7.4

¿Cómo ampliaría el modelo estimado en (7.18) para ver la diferencia en el retorno a *tenure (antigüedad) entre los géneros?*

Para ver un ejemplo más complicado con interacciones, se revisarán los efectos de la raza y de la composición racial de la ciudad sobre los salarios de los jugadores de la liga mayor de béisbol.

Ejemplo 7.11

[Efectos de la raza sobre los sueldos de los jugadores de béisbol]

Empleando el archivo MLB1.RAW, para los 330 jugadores de la liga mayor de béisbol, para los que se conoce la composición racial de la ciudad, se estima la ecuación a continuación. Las variables *black* e *hispan* son indicadores binarios para cada jugador. (El grupo base es jugadores blancos.) La variable *percblck* es el porcentaje de negros del equipo, en la ciudad y *perchisp* es el porcentaje de hispanos. Las demás variables miden aspectos de la productividad y de la longevidad del jugador. Aquí interesan los efectos de la raza una vez controlados estos otros factores.

Además de incluir *black* e *hispan* en la ecuación, se agregan las interacciones *black·percblck* e *hispan·perchisp*. La ecuación estimada es

$$
\begin{aligned}
\widehat{\log(salary)} = {} & \underset{(2.18)}{10.34} + \underset{(.0129)}{.0673}\, years + \underset{(.0034)}{.0089}\, gamesyr \\
& + \underset{(.00151)}{.00095}\, bavg + \underset{(.0164)}{.0146}\, hrunsyr + \underset{(.0076)}{.0045}\, rbisyr \\
& + \underset{(.0046)}{.0072}\, runsyr + \underset{(.0021)}{.0011}\, fldperc + \underset{(.0029)}{.0075}\, allstar \\
& - \underset{(.125)}{.198}\, black - \underset{(.153)}{.190}\, hispan + \underset{(.0050)}{.0125}\, black\cdot percblck \\
& + \underset{(.0098)}{.0201}\, hispan\cdot perchisp \\
n = 330,\ & R^2 = .638.
\end{aligned}
\tag{7.19}
$$

Primero, se debe probar si las cuatro variables raciales, *black*, *hispan*, *black·percblck* e *hispan·perchisp*, son conjuntamente significativas. Usando los mismos 330 jugadores, la *R*-cuadrada cuando se eliminan las cuatro variables raciales es .626. Como hay cuatro restricciones y en el modelo no restringido $gl = 330 - 13$ el estadístico F es aproximadamente 2.63, con lo que se obtiene un valor-p de .034. Por tanto, estas variables son conjuntamente significativas al nivel de 5% (aunque no al nivel de 1%).

¿Cómo se interpretan los coeficientes de las variables raciales? En el análisis siguiente, todos los factores de productividad se mantienen constantes. Primero, se considera lo que ocurre con los jugadores negros, cuando *perchisp* se mantiene constante. El coeficiente de *black* −.198, significa literalmente que si un jugador negro está en una ciudad en la que no haya ningún negro ($percblck = 0$), entonces el jugador negro ganará aproximadamente 19.8% menos que un jugador blanco igual a él. A medida que *percblck* aumenta —lo que significa que la población blanca disminuye, ya que *perchisp* se mantiene constante— el sueldo de los negros aumenta con relación al de los blancos. En una ciudad con 10% de negros, log(*salary*) para los negros en comparación con el de los blancos es $-.198 + .0125(10) = -.073$, de manera que en una ciudad así, el salario de los negros es aproximadamente 7.3% menor que el de los blancos. Cuando $percblck = 20$, los negros ganan aproximadamente 5.2% más que los blancos. El porcentaje más alto de negros en una ciudad es aproximadamente 74% (Detroit).

De manera similar, en ciudades en las que el porcentaje de hispanos es bajo, éstos ganan menos que los blancos. Pero el valor de *perchisp* que hace que la diferencia entre blancos e hispanos sea igual a cero puede calcularse fácilmente: este valor debe hacer $-.190 + .0201\, perchisp = 0$, de donde se obtiene $perchisp \approx 9.45$. En las ciudades en las que el porcentaje de hispanos sea menor a 9.45%, se predice que los hispanos ganarán menos que los blancos (con una población negra dada) y lo contrario ocurre si el porcentaje de hispanos es superior a 9.45%. En 12 de las 22 ciudades de la muestra la población hispana es menor al 9.45% de la población total. El mayor porcentaje de hispanos es aproximadamente 31%.

¿Cómo se interpretan estos hallazgos? No puede simplemente decirse que exista discriminación contra los negros y contra los hispanos, ya que las estimaciones implican que los blancos ganan menos que los negros y que los hispanos en las ciudades fuertemente pobladas por minorías. La importancia de la composición de la ciudad sobre el salario puede deberse a preferencias por los jugadores: quizá los mejores jugadores negros viven especialmente en ciudades con más negros y los mejores jugadores hispanos tienden a encontrarse en ciudades con más hispanos. Las estimaciones de (7.19) permiten decir que existe alguna relación, pero no se puede distinguir entre estas dos hipótesis.

Prueba para diferencias en las funciones de regresión a través de los grupos

Los ejemplos anteriores ilustran que las variables binarias que interactúen con otras variables independientes pueden ser una poderosa herramienta. Algunas veces se desea probar la hipótesis nula de que dos poblaciones o grupos siguen una misma función de regresión, contra la alternativa de que una o más de las pendientes difiere entre los grupos. En el capítulo 13, cuando se analice la combinación de diversos cortes transversales a lo largo del tiempo, se verán también otros ejemplos.

Suponga que se desea probar si el mismo modelo de regresión describe los promedios generales de calificaciones (*cumgpa*) de los atletas universitarios hombres y de las atletas universitarias mujeres. La ecuación es

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

donde *sat* es la puntuación obtenida en el examen de admisión a la universidad, *hsperc* es el percentil alcanzado en el bachillerato y *tothrs* es el total de horas de cursos en la universidad. Se sabe que para la diferencia entre los interceptos puede incluirse una variable binaria, ya sea para hombres o para mujeres. Si se desea que cualquiera de las pendientes dependa del género, simplemente se hace que la variable adecuada interactúe con, por ejemplo, *female*, y se incluye en la ecuación.

Si lo que interesa es *cualquier* diferencia que exista entre hombres y mujeres, entonces se debe buscar un modelo en el que el intercepto y todas las pendientes puedan ser diferentes entre los dos grupos:

$$\begin{aligned} cumgpa = {} & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\ & + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u. \end{aligned} \tag{7.20}$$

El parámetro δ_0 es la diferencia en el intercepto entre mujeres y hombres, δ_1 es la diferencia entre hombres y mujeres en la pendiente respecto a *sat*, etc. La hipótesis nula de que *cumgpa* sigue el mismo modelo para hombres y mujeres se expresa como sigue

$$H_0\text{: } \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \tag{7.21}$$

Si uno de los δ_j es distinto de cero, entonces el modelo es diferente para hombres y mujeres.

Con ayuda de los datos del semestre de primavera del archivo GPA3.RAW, el modelo completo estimado es

$$\begin{aligned} \widehat{cumgpa} = {} & \underset{(0.21)}{1.48} - \underset{(.411)}{.353}\, female + \underset{(.0002)}{.0011}\, sat + \underset{(.00039)}{.00075}\, female \cdot sat \\ & - \underset{(.0014)}{.0085}\, hsperc - \underset{(.00316)}{.00055}\, female \cdot hsperc + \underset{(.0009)}{.0023}\, tothrs \\ & - \underset{(.00163)}{.00012}\, female \cdot tothrs \\ n = {} & 366, R^2 = .406, \bar{R}^2 = .394. \end{aligned} \tag{7.22}$$

Ninguno de los cuatro términos en los que interviene la variable binaria para mujeres (*female*) es estadísticamente muy significativo; sólo la interacción *female·sat* tiene un estadístico t cercano a dos. Pero se conoce algo mejor que emplear estadísticos t individuales para probar una hipótesis conjunta como la (7.21). Para calcular el estadístico F se debe estimar el modelo restringido que se obtiene eliminando *female* y todas las interacciones; con esto se obtiene una R^2 (la R^2 restringida) de aproximadamente .352, con lo que el estadístico F es aproximadamente 8.14; el valor-p es cero a cinco cifras decimales, lo que hace que (7.21) claramente se rechace. Por tanto, los atletas, hombres y mujeres, siguen modelos diferentes para el GPA, aun los términos de (7.22) que toman en cuenta que hombres y mujeres sean diferentes no sean individualmente significativos al nivel de 5%.

Los grandes errores estándar de *female* y de los términos de interacción dificultan decir exactamente cómo difieren hombres y mujeres. Al interpretar la ecuación (7.22) se debe ser muy cuidadoso porque, al obtener las diferencias entre hombres y mujeres, deben tomarse en cuenta los términos de interacción. Si sólo se ve la variable *female* puede concluirse erróneamente que, manteniendo todo lo demás constante *cumgpa* es aproximadamente .353 menor para las mujeres que para los hombres. Esta es sólo la diferencia estimada cuando *sat*, *hsperc* y *tothrs* se igualan todas a cero, lo cual no se parece a un escenario posible. Cuando $sat = 1{,}100$, $hsperc = 10$ y $tothrs = 50$, la *diferencia* que se predice entre hombres y mujeres es $-.353 + .00075(1{,}100) - .00055(10) - .00012(50) \approx .461$. Es decir, se predice que las mujeres tengan un GPA casi medio punto superior al de los hombres atletas iguales a ellas.

En un modelo con tres variables, *sat*, *hsperc* y *tothrs*, es bastante fácil agregar todas las interacciones para probar las diferencias entre los grupos. En algunos casos, sin embargo, hay muchas más variables explicativas, y entonces es conveniente tener otro modo de calcular el estadístico. Resulta que la forma de suma de residuales cuadrados del estadístico F puede calcularse fácilmente aun cuando intervengan muchas variables independientes.

En el modelo general con k variables explicativas y un intercepto, suponga que se tienen dos grupos: $g = 1$ y $g = 2$. Se desearía probar si los interceptos y las pendientes son iguales en los dos grupos. Este modelo se expresa como

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u, \tag{7.23}$$

para $g = 1$ y $g = 2$. La hipótesis de que cada beta en (7.23) es igual en los dos grupos incluye $k + 1$ restricciones (en el ejemplo del GPA, $k + 1 = 4$). El modelo no restringido, que se puede entender como un modelo con una variable binaria de grupo y k términos de interacción, además del intercepto y las variables mismas, tiene $n - 2(k + 1)$ grados de libertad. [En el ejemplo del GPA, $n - 2(k + 1) = 366 - 2(4) = 358$.] Hasta ahora no hay nada nuevo. La idea clave es que la suma de los residuales cuadrados del modelo no restringido puede obtenerse de dos regresiones *separadas*, una para cada grupo. Sea SRC_1 la suma de residuales cuadrados obtenida estimando (7.23) para el primer grupo; aquí intervienen n_1 observaciones. Sea SRC_2 la suma de residuales cuadrados obtenida de la estimación del modelo empleando el segundo grupo (n_2 observaciones). En el ejemplo anterior, si el grupo 1 es mujeres, entonces $n_1 = 90$ y $n_2 = 276$. Ahora, la suma de los residuales cuadrados para el modelo no restringido es simplemente $\text{SRC}_{nr} = \text{SRC}_1 + \text{SRC}_2$. La suma restringida de residuales cuadrados es sencillamente la SRC de juntar los grupos y estimar una sola ecuación, SRC_P. Una vez que se tienen éstas, se calcula el estadístico F como de costumbre:

$$F = \frac{[\text{SRC}_P - (\text{SRC}_1 + \text{SRC}_2)]}{\text{SRC}_1 + \text{SRC}_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}, \tag{7.24}$$

donde n es la cantidad *total* de observaciones. A este estadístico F particular se le conoce en econometría como **estadístico de Chow**. Dado que la prueba de Chow es simplemente una prueba F, sólo es válida bajo homocedasticidad. En particular, bajo la hipótesis nula, las varianzas del error de los dos grupos deben ser iguales. Como siempre, la normalidad no es necesaria para el análisis asintótico.

Para aplicar el estadístico de Chow al ejemplo del GPA, se necesita la SRC de la regresión en la que se juntan los dos grupos: esto es $\text{SRC}_P = 85.515$. La SRC de las 90 mujeres de la muestra es $\text{SRC}_1 = 19.603$ y la SRC de los hombres es $\text{SRC}_2 = 58.752$. Por tanto, $\text{SRC}_{ur} = 19.603 + 58.752 = 78.355$. El estadístico F es $[(85.515 - 78.355)/78.355](358/4) \approx 8.18$; sujeto, por supuesto, al error de redondeo, esto es lo que se obtiene empleando la forma R-cuadrada de la prueba en los modelos con y sin los términos de interacción. [Una advertencia: no hay forma R-cuadrada sencilla de la prueba si para cada grupo se han estimado regresiones por separado; la forma R-cuadrada de la prueba sólo puede emplearse si se han introducido interacciones para crear el modelo no restringido.]

Una limitación importante de la prueba de Chow, al margen del método que se emplee para realizarla, es que la hipótesis nula no toma en consideración ninguna diferencia entre los grupos. En muchos casos es más interesante tomar en consideración diferencias entre los interceptos de los grupos y probar, después, si hay diferencias entre las pendientes. Un ejemplo de esto se vio en la ecuación del salario en el ejemplo 7.10. Hay dos maneras de considerar diferencias entre los interceptos bajo la hipótesis nula. Una es incluir la binaria de grupo y todos los términos de interacción como en la ecuación (7.22), y después probar sólo la significancia conjunta de los términos de interacción. La segunda es formar un estadístico F como en la ecuación (7.24), pero en el que la suma restringida de cuadrados, que en la ecuación (7.24) se designa como "SRC_P"

se obtenga mediante la regresión que sólo tome en cuenta un desplazamiento del intercepto. En otras palabras, se corre una regresión conjunta y sólo se incluye la variable binaria que distingue a los dos grupos. En el ejemplo del promedio de las calificaciones (GPA) se regresa *cumgpa* sobre *female*, *sat*, *hsperc* y *tothrs* usando los datos de los estudiantes atletas, hombres y mujeres. En el ejemplo del GPA se usa el primer método y, por tanto, en la ecuación (7.20) la hipótesis nula es H_0: $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$. (Bajo la hipótesis nula δ_0 no está restringida.) El estadístico F para estas tres restricciones es aproximadamente 1.53, lo que da un valor-p igual a .205. Por tanto, no se rechaza la hipótesis nula.

No rechazar la hipótesis de que los parámetros que multiplican a los términos de interacción son todos cero, sugiere que el mejor modelo considere sólo una intersección diferente de cero:

$$\begin{aligned}\widehat{cumgpa} = \underset{(.18)}{1.39} &+ \underset{(.059)}{.310}\, female + \underset{(.0002)}{.0012}\, sat - \underset{(.0012)}{.0084}\, hsperc \\ &+ \underset{(.0007)}{.0025}\, tothrs \\ n = 366,&\ R^2 = .398,\ \bar{R}^2 = .392.\end{aligned} \tag{7.25}$$

Los coeficientes de pendiente en (7.25) son muy parecidos a los del grupo base (hombres) en (7.22); eliminar las interacciones cambia muy poco. Sin embargo, en (7.25) *female* es muy significativa: su estadístico t es mayor a 5 y la estimación implica que para valores dados de *sat*, *hsperc* y *tothrs*, el GPA que se pronostica para una atleta (mujer) es .31 puntos superior que para un atleta (hombre). Esta diferencia es importante en la práctica.

7.5 Una variable dependiente binaria: el modelo de probabilidad lineal

Hasta ahora se ha estudiado mucho acerca de las propiedades y de la aplicabilidad del modelo de regresión lineal múltiple. En las últimas secciones se vio cómo puede incorporarse información cualitativa al modelo de regresión múltiple en forma de variables explicativas mediante el uso de variables binarias. En todos los modelos vistos hasta ahora, la variable dependiente y ha tenido un significado *cuantitativo* (por ejemplo, y ha sido una cantidad en dólares, la puntuación en un examen, un porcentaje, o el logaritmo de éstos). ¿Qué ocurre cuando se desea usar la regresión múltiple para *explicar* eventos cualitativos?

En el caso más sencillo, y que en la práctica se encuentra con frecuencia, el evento que se desea explicar tiene un resultado binario. En otras palabras, la variable dependiente y sólo toma dos valores: cero o uno. Por ejemplo, y puede indicar si un adulto tiene o no educación universitaria, o si un estudiante universitario consume o no drogas durante un año escolar, o si una empresa ha sido absorbida por otra durante un determinado año. En cada uno de estos ejemplos se puede hacer que $y = 1$ denote uno de los resultados y $y = 0$ denote el otro.

¿Qué significa escribir un modelo de regresión múltiple como el siguiente,

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u, \tag{7.26}$$

cuando y es una variable bivariada? Como y sólo puede tomar dos valores, las β_j no pueden interpretarse como el cambio en y para un aumento dado de x_j, permaneciendo constantes todos los demás factores: y cambia de uno a cero o de cero a uno (o no cambia). A pesar de esto, las β_j siguen teniendo una interpretación útil. Si se admite que el supuesto RLM.4 de media condicional cero se satisface, es decir que $E(u|x_1, \dots, x_k) = 0$, entonces se tiene, como siempre,

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

donde **x** es una manera abreviada de escribir todas las variables explicativas.

El punto clave es que cuando y es una variable binaria que toma los valores cero o uno, entonces se tiene siempre que $\mathrm{P}(y = 1|\mathbf{x}) = \mathrm{E}(y|\mathbf{x})$: la probabilidad de "éxito" —es decir, la probabilidad de que $y = 1$— es lo mismo que el valor esperado de y. Por tanto, se tiene la importante ecuación

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{7.27}$$

que dice que la probabilidad de éxito, es decir, $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$, es función lineal de las variables x_j. La ecuación (7.27) es un ejemplo de un modelo de respuesta bivariada y a $\mathrm{P}(y = 1|\mathbf{x})$ también se le conoce como la **probabilidad de respuesta**. (En el capítulo 17 se verán otros modelos de respuesta bivariada.) Dado que las probabilidades deben sumar uno, $\mathrm{P}(y = 0|\mathbf{x}) = 1 - \mathrm{P}(y = 1|\mathbf{x})$ es también una función lineal de las x_j.

A un modelo de regresión lineal múltiple en el que la variable dependiente es una variable binaria se le conoce como **modelo de probabilidad lineal (MPL)** debido a que la probabilidad de respuesta es lineal en los parámetros β_j. En el MPL, β_j mide la variación de la probabilidad de éxito al variar x_j permaneciendo los demás factores constantes:

$$\Delta \mathrm{P}(y = 1|\mathbf{x}) = \beta_j \Delta x_j. \tag{7.28}$$

Con esto en mente, el modelo de regresión múltiple permite estimar el efecto de diversas variables explicativas sobre un evento cualitativo. La mecánica de MCO es la misma que antes.

Si la ecuación estimada se escribe como

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k,$$

hay que recordar que $\hat{y}$ es la probabilidad de éxito predicha. Por tanto, $\hat{\beta}_0$ es la probabilidad de éxito predicha cuando cada una de las x_j es igual a cero, lo cual puede o no ser de interés. El coeficiente de pendiente $\hat{\beta}_1$ mide la variación de la probabilidad de éxito predicha cuando x_1 aumenta una unidad.

Para interpretar correctamente un modelo de probabilidad lineal, debe saberse qué es lo que constituye un "éxito". De manera que es buena idea dar a la variable dependiente un nombre que describa el evento $y = 1$. Por ejemplo, sea *inlf* ("en la fuerza de trabajo") una variable bivariada que indique que una mujer casada participó en la fuerza de trabajo durante 1975: $inlf = 1$ si una mujer informa haber trabajado, fuera de la casa, por un salario, durante ese año y cero si no es así. Se supone que la participación en la fuerza de trabajo depende de otras fuentes de ingresos, entre las que están los ingresos del esposo (*nwifeinc*, en miles de dólares), los años de educación (*educ*), los años de experiencia en el mercado laboral (*exper*), la edad (*age*), la cantidad de hijos menores de seis años (*kidslt6*) y la cantidad de hijos entre 6 y 18 años (*kidsge6*). Empleando los datos del archivo MROZ.RAW de Mroz (1987), se estima el siguiente modelo de probabilidad lineal, en el que 428 de las 753 mujeres que conforman la muestra indican haber pertenecido a la fuerza de trabajo en algún momento durante 1975:

$$\begin{aligned}
\widehat{inlf} = \underset{(.154)}{.586} &- \underset{(.0014)}{.0034}\, nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper \\
&- \underset{(.00018)}{.00060}\, exper^2 - \underset{(.002)}{.016}\, age - \underset{(.034)}{.262}\, kidslt6 + \underset{(.0132)}{.0130}\, kidsge6 \\
n &= 753,\ R^2 = .264.
\end{aligned} \tag{7.29}$$

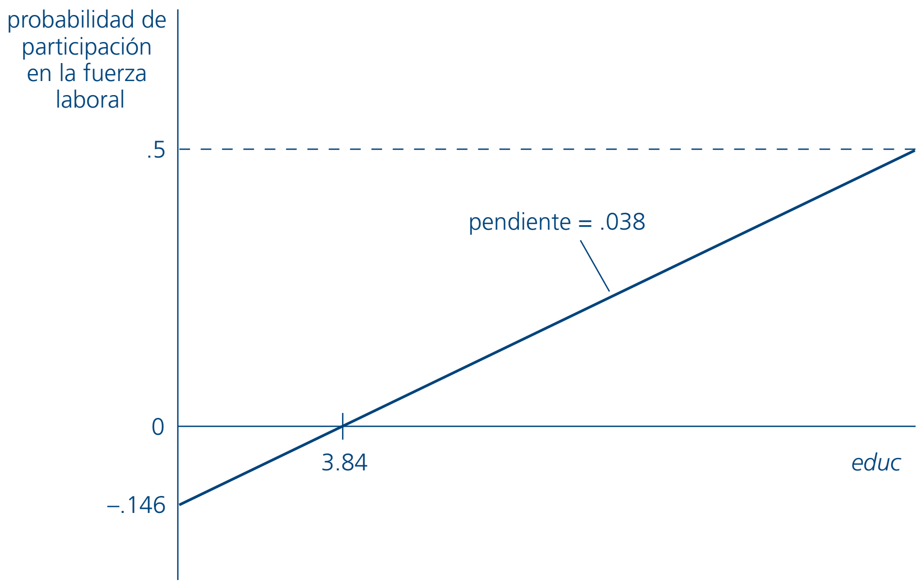
Empleando el estadístico t usual en (7.29) todas las variables, excepto *kidsge6* son estadísticamente significativas y todas las variables significativas tienen el efecto que, con base en la teoría económica (o en el sentido común), era de esperarse.

Para interpretar las estimaciones, se debe recordar que una variación en la variable independiente modifica la probabilidad de que $inlf = 1$. Por ejemplo, el coeficiente de *educ* significa que, permaneciendo en (7.29) todo lo demás constante, un año más de educación hace que la probabilidad de participación en la fuerza de trabajo aumente .038. Si esta ecuación se considera literalmente, 10 años más de educación incrementan la probabilidad de pertenecer a la fuerza de trabajo en $.038(10) = .38$, lo cual es un incremento bastante grande en una probabilidad. En la figura 7.3 se grafica la relación entre la probabilidad de participación en la fuerza de trabajo y *educ*. Para propósitos de ilustración, las demás variables independientes se mantienen constantes en los valores $nwifeinc = 50$, $exper = 5$, $age = 30$, $kidslt6 = 1$ y $kidsge6 = 0$. La probabilidad predicha es negativa hasta que la educación es igual a 3.84 años. Esto no debe preocupar mucho, debido a que en esta muestra ninguna mujer tiene menos de cinco años de educación. La mayor cantidad de años de educación encontrada es 17 años y lleva a predecir una probabilidad de .5. Si las demás variables independientes se igualan a otros valores, el rango de predicción para las probabilidades será diferente. Pero el efecto marginal de un año más de educación sobre la probabilidad de participación en la fuerza de trabajo es siempre .038.

El coeficiente de *nwifeinc* implica que, si $\Delta nwifeinc = 10$ (lo que significa un incremento de \$10,000), la probabilidad de que una mujer esté en la fuerza de trabajo disminuye .034. Este efecto no es especialmente grande, dado que un aumento de \$10,000 en el ingreso es significativo en términos de dólares de 1975. La experiencia se ha introducido como una función cuadrática para que el efecto de la experiencia tenga un efecto decreciente sobre la probabilidad de participación en la fuerza de trabajo. Manteniendo los demás factores constantes, la variación

FIGURA 7.3

Relación estimada entre la probabilidad de pertenecer a la fuerza laboral y los años de educación, permaneciendo las demás variables explicativas constantes.



de la probabilidad se aproxima como $.039 - 2(.0006)exper = .039 - .0012\ exper$. El punto en el que la experiencia transcurrida no tiene efecto sobre la probabilidad de participación en la fuerza de trabajo es $.039/.0012 = 32.5$, que es un valor grande para experiencia: sólo 13 de las 753 mujeres de la muestra tiene más de 32 años de experiencia.

A diferencia de la cantidad de hijos mayores, la cantidad de hijos pequeños tiene un enorme impacto sobre la participación en la fuerza de trabajo. Tener un hijo, menor de seis años adicional, reduce la probabilidad de participación $-.262$, a los valores dados de las demás variables. En la muestra, poco menos de 20% de las mujeres tienen al menos un hijo pequeño.

Este ejemplo muestra lo fácil que es estimar e interpretar los modelos de probabilidad lineal, pero también pone de manifiesto algunas desventajas del MPL. Primero, es fácil ver que si en (7.29) se sustituyen ciertas combinaciones de valores de las variables independientes, se pueden obtener predicciones, ya sea menores a cero o mayores a uno. Dado que se trata de predicciones de probabilidad, y que las probabilidades sólo toman valores desde cero hasta uno, esto puede ser un poco confuso. Por ejemplo, ¿qué significará decir que la probabilidad de que una mujer pertenezca a la fuerza de trabajo es $-.10$? De las 753 mujeres que conforman la muestra, para 16 los valores ajustados obtenidos con (7.29) son menores a cero y para 17 los valores ajustados son mayores a uno.

Otro problema es que la probabilidad no puede estar relacionada en forma lineal con las variables independientes para todos sus valores posibles. Por ejemplo, (7.29) predice que el efecto de pasar de cero hijos a un hijo pequeño hace que la probabilidad de trabajar se reduzca .262. Esta es también la disminución que se predice cuando una mujer pasa de un hijo a dos. Parece más razonable pensar que el primer hijo reduzca esta probabilidad en una cantidad grande, pero que los hijos subsecuentes tengan un efecto marginal. En realidad, llevada al extremo, (7.29) implica que pasar de cero a cuatro hijos pequeños hace que la probabilidad de trabajar se reduzca $\Delta\widehat{inlf} = .262(\Delta kidslt6) = .262(4) = 1.048$, lo cual es imposible.

A pesar de estos problemas, el modelo de probabilidad lineal es útil y suele emplearse con frecuencia en economía. Por lo general, este modelo funciona bien con valores de las variables independientes que estén cerca de los promedios en la muestra. En el ejemplo de participación en la fuerza de trabajo, ninguna de las mujeres de la muestra tiene cuatro hijos; sólo tres de ellas tienen tres hijos. Más de 96% de las mujeres de la muestra o no tiene ningún hijo pequeño o sólo uno y, por tanto, quizá al interpretar la ecuación estimada deba restringirse la atención a este caso.

Las probabilidades predichas fuera del intervalo unitario son un poco problemáticas cuando se quiere hacer predicciones. Sin embargo, existen maneras de usar las probabilidades estimadas (aun cuando algunas sean negativas o mayores a uno) para predecir un resultado cero o uno. Como antes, sean las $\hat{y}_i$ los valores ajustados —que pueden no quedar acotados entre cero y uno—. Defínase un valor predicho como $\tilde{y}_i = 1$ si $\hat{y}_i \geq .5$ y $\tilde{y}_i = 0$ si $\hat{y}_i < .5$. Entonces se tiene un conjunto de valores predichos, $\tilde{y}_i$, $i = 1, \ldots, n$, que son, como las y_i, cero o uno. Empleando los datos de las y_I y de las $\tilde{y}_i$ se puede obtener la frecuencia con la que se predice correctamente $y_i = 1$ y $y_i = 0$, así como la proporción de predicciones correctas en general. Esta última medida, convertida en porcentaje, es una medida muy usada de la bondad de ajuste para variables dependientes binarias: el **porcentaje predicho correctamente**. En el ejercicio para computadora C7.9(v) se da un ejemplo, y en la sección 17.1 se presenta un análisis más completo en el contexto de modelos más avanzados.

Debido al carácter binario de y, el modelo de probabilidad lineal vulnera uno de los supuestos de Gauss-Markov. Cuando y es una variable binaria, su varianza, condicional en $\mathbf{x}$, es

$$\mathrm{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \tag{7.30}$$

donde $p(\mathbf{x})$ es una abreviación de la probabilidad de éxito: $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. Esto significa que, salvo en el caso en que la probabilidad no depende de ninguna de las variables independientes, en un modelo de probabilidad lineal *habrá* heterocedasticidad. De acuerdo con lo visto en el capítulo 3, se sabe que esto no ocasiona sesgo en los estimadores de MCO de las β_j. Pero se sabe también, de acuerdo con lo visto en los capítulos 4 y 5, que la homocedasticidad es crucial para justificar los estadísticos t y F usuales, aun con muestras grandes. Como en (7.29) los errores estándar no son válidos en general, deben usarse con cuidado. En el capítulo 8 se mostrará cómo corregir los errores estándar cuando hay heterocedasticidad. Resulta que, en muchas aplicaciones, los estadísticos usuales de MCO no están lejos y en el trabajo práctico es aceptable presentar un análisis estándar de MCO de un modelo de probabilidad lineal.

Ejemplo 7.12

[Un modelo de probabilidad lineal para arrestos]

Sea *arr86* una variable binaria igual a uno si un hombre fue arrestado en 1986 e igual a cero si no fue así. La población es un grupo de hombres de California nacidos en 1960 o en 1961, que habían sido detenidos al menos una vez antes de 1986. Un modelo de probabilidad lineal para describir *arr86* es

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

donde

$pcnv$ = proporción de arrestos previos que condujeron a una condena.
$avgsen$ = sentencia promedio cumplida en condenas previas (en meses).
$tottime$ = meses en prisión y desde los 18 años de edad anteriores a 1986.
$ptime86$ = meses en prisión en 1986.
$qemp86$ = cantidad de trimestres (0 a 4) que el hombre estuvo empleado legalmente en 1986.

Los datos empleados se encuentran en el archivo CRIME1.RAW, la misma base de datos que se empleó en el ejemplo 3.5. Aquí se usa una variable dependiente bivariada, porque sólo 7.2% de los hombres de la muestra fueron detenidos más de una vez. Alrededor de 27.7% de los hombres de la muestra fueron detenidos al menos una vez durante 1986. La ecuación estimada es

$$\begin{aligned}\widehat{arr86} = &\ \underset{(.017)}{.441} - \underset{(.021)}{.162}\, pcnv + \underset{(.0065)}{.0061}\, avgsen - \underset{(.0050)}{.0023}\, tottime \\ &- \underset{(.005)}{.022}\, ptime86 - \underset{(.005)}{.043}\, qemp86 \\ n = &\ 2{,}725,\ R^2 = .0474.\end{aligned} \tag{7.31}$$

El intercepto, .441, es la probabilidad de ser arrestado que se predice para un hombre que no ha sido condenado (y, en consecuencia, tanto *pcnv* como *avgsen* son cero), no ha estado en prisión después de los 18 años, no ha estado en prisión en 1986 y ha estado desempleado todo el año. Las variables *avgsen* y *tottime* no son individual ni conjuntamente significativas (el valor-p en la prueba F es .347) y *avgsen* tiene un valor contraintuitivo si se supone que con sentencias más largas se desalienta la criminalidad. Grogger (1991), empleando un conjunto mayor, de este tipo de datos, y diversos métodos econométricos, halló que *tottime* tenía un efecto significativo *positivo* sobre las detenciones y concluyó que *tottime* es una medida del capital humano acumulado en la actividad delictiva.

El aumento de la probabilidad de ser condenado disminuye la probabilidad de ser arrestado, pero hay que tener cuidado al interpretar la magnitud del coeficiente. La variable *pcnv* es una proporción entre cero y uno; de manera que la variación de *pcnv* de cero a uno significa esencialmente el paso de ninguna probabilidad de ser sentenciado a ser sentenciado con seguridad. Aun este cambio tan grande reduce sólo .162 la probabilidad de ser detenido; un aumento de *pcnv* de .5 disminuye la probabilidad de detención .081.

El efecto del encarcelamiento está dado por el coeficiente de *ptime86*. Si un hombre está en prisión no puede ser arrestado. Como *ptime86* se mide en meses, seis meses más en prisión reduce la probabilidad de detención $.022(6) = .132$. La ecuación (7.31) proporciona otro ejemplo en el que el modelo de probabilidad lineal no puede ser cierto en todos los rangos de las variables independientes. Si un hombre está en prisión durante los 12 meses de 1986, no puede ser detenido durante 1986. Igualando a cero todas las demás variables, cuando $ptime86 = 12$ la probabilidad de detención predicha es $.441 - .022(12) = .177$, que es distinta de cero. Sin embargo, si se parte de la probabilidad incondicional de detención, .277, 12 meses en prisión reduce la probabilidad esencialmente a cero: $.277 - .022(12) = .013$.

Por último, tener un empleo reduce la probabilidad de detención de manera significativa. Permaneciendo todos los demás factores constantes, para un hombre empleado durante los cuatro trimestres, la probabilidad de ser detenido es .172 menor que para un hombre que en absoluto estuvo empleado.

En modelos con una variable dependiente binaria, también pueden incluirse variables independientes binarias. El coeficiente mide la diferencia que se predice para la probabilidad en relación con el grupo base. Por ejemplo, si a la ecuación para los arrestos se le agregan dos binarias para raza, *black* e *hispan*, se obtiene

$$
\begin{aligned}
\widehat{arr86} = {} & \underset{(.019)}{.380} - \underset{(.021)}{.152}\, pcnv + \underset{(.0064)}{.0046}\, avgsen - \underset{(.0049)}{.0026}\, tottime \\
& - \underset{(.005)}{.024}\, ptime86 - \underset{(.005)}{.038}\, qemp86 + \underset{(.024)}{.170}\, black + \underset{(.021)}{.096}\, hispan \\
& n = 2{,}725,\ R^2 = .0682.
\end{aligned}
\tag{7.32}
$$

El coeficiente de *black* significa que, permaneciendo todos los demás factores constantes, un hombre negro tiene una probabilidad .17 mayor de ser detenido que un hombre blanco (el grupo base). Otra manera de expresar esto es que la probabilidad de ser detenido es 17 puntos porcentuales mayor para negros que para blancos. Esta diferencia también es estadísticamente significativa. De manera similar, un hombre hispano tiene una probabilidad .096 mayor de ser arrestado que un hombre blanco.

Pregunta 7.5

¿Cuál es la probabilidad de ser arrestado que se predice para un hombre negro que no ha sido sentenciado antes —de manera que *pcnv*, *avgsen*, *tottime* y *ptime86* son todas cero— y que ha estado empleado durante los cuatro trimestres de 1986? ¿Parece esto razonable?

7.6 Más acerca del análisis de políticas y evaluación de programas

Se han visto algunos ejemplos de modelos que contienen variables binarias que pueden ser útiles para la evaluación de políticas. En el ejemplo 7.3 se presenta un ejemplo de la evaluación de un programa consistente en que algunas empresas reciban subsidios para capacitación para el trabajo mientras que otras no.

Como se dijo antes, cuando se evalúan programas debe tenerse cuidado, ya que en las ciencias sociales, en la mayoría de los ejemplos, los grupos de tratamiento y de control no se asignan aleatoriamente. Considere otra vez el estudio de Holzer *et al*. (1993); en este estudio interesa, ahora, el efecto de la subvención para la capacitación en el trabajo sobre la productividad de los trabajadores (y no sobre la cantidad de capacitación). La ecuación que interesa es

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u,$$

donde *scrap* es la tasa de desperdicio de la empresa y donde las dos últimas variables ventas (*sales*) y número de empleados (*employ*), se incluyen como controles. La variable binaria *grant* indica si en 1988 la empresa recibió subsidio para capacitación.

Antes de ver las estimaciones, puede que preocupe que los factores no observados, que puedan afectar la productividad de los trabajadores —tales como nivel promedio de educación, habilidad, experiencia y antigüedad— estén correlacionados con que la empresa reciba o no subsidio. Holzer *et al*., indican que los subsidios fueron otorgados a las empresas conforme los solicitaron. Pero esto no es lo mismo que otorgar subsidios aleatoriamente. Puede ser que las empresas con empleados menos productivos hayan visto en el subsidio una oportunidad de aumentar la productividad y por eso hayan sido las más diligentes para solicitarlo.

Empleando los datos del archivo JTRAIN.RAW para 1988 —cuando las empresas ya recibían el subsidio— se obtuvo

$$\begin{aligned} \widehat{\log(scrap)} = {} & \underset{(4.66)}{4.99} - \underset{(.431)}{.052}\, grant - \underset{(.373)}{.455} \log(sales) \\ & + \underset{(.365)}{.639} \log(employ) \\ n = {} & 50,\ R^2 = .072. \end{aligned} \tag{7.33}$$

(Diecisiete de las 50 empresas obtuvieron subsidio y la tasa promedio de desperdicio, considerando todas las empresas, es 3.47.) La estimación puntual obtenida para *grant* −.052, significa que, para valores dados de *sales* y *employ*, las empresas que obtienen un subsidio tienen tasas de desperdicio aproximadamente 5.2% menores que las empresas que no obtuvieron subsidio. Esta es la dirección del efecto esperado si los subsidios para capacitación son efectivos, pero el estadístico *t* es muy pequeño. Por tanto, de acuerdo con este análisis de corte transversal, se debe concluir que los subsidios no tuvieron ningún efecto sobre la productividad de las empresas. En el capítulo 9 se volverá a este ejemplo y se verá que añadiendo la información del año anterior se llega a una conclusión muy distinta.

Aun en casos en los que en el análisis de una política no haya que considerar la asignación de unidades al grupo control y al grupo de tratamiento, se debe tener cuidado de incluir factores que puedan estar sistemáticamente relacionados con la variable independiente binaria de interés. Un buen ejemplo de esto es una prueba de discriminación racial. La raza es algo que no está determinado por un individuo o por un administrador público. De hecho, la raza parece ser un ejemplo perfecto de una variable explicativa exógena, dado que está determinada por nacimiento. Sin embargo, debido a razones históricas, la raza suele estar relacionada con otros factores importantes: las razas tienen antecedentes de diferencias sistemáticas, los cuales pueden ser importantes en una prueba sobre la discriminación *actual*.

Para dar un ejemplo, considere una prueba de discriminación para aprobaciones de créditos hipotecarios. Si, por ejemplo, se recolectan datos sobre solicitudes personales de créditos hipotecarios, puede definirse una variable dependiente binaria *approved* que sea igual a uno si la solicitud es aprobada e igual a cero si no lo es. Una diferencia sistemática entre las razas, respecto a estas tasas de aprobación indica discriminación. Sin embargo como la aprobación depende de otros muchos factores, como ingreso, riqueza, calificación de crédito y la capacidad general para pagar el préstamo, estas variables deben controlarse *si* entre las razas existen diferencias sistemáticas en estos factores. Un modelo de probabilidad lineal para probar si existe discriminación puede ser como el siguiente:

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + other\ factors.$$

La discriminación contra las minorías estará indicada por el rechazo de H_0: $\beta_1 = 0$ en favor de H_0: $\beta_1 < 0$, ya que β_1 es la magnitud en que la probabilidad de que la solicitud de una persona que no sea de raza blanca (*nonwhite*) sea aprobada difiere de la probabilidad de que la solicitud de una persona de esa raza sea aprobada, dados los mismos valores de las demás variables en la ecuación. Si, en un análisis de regresión múltiple *income* (ingreso), *wealth* (riqueza), *credrate* (calificación crediticia), etc., difieren sistemáticamente entre las razas será importante controlar estos factores.

Otro problema que suele presentarse en la evaluación de políticas y programas es que las personas (las empresas o las ciudades) eligen tomar parte o no en ciertos comportamientos o programas. Por ejemplo, las personas deciden utilizar drogas o beber alcohol. Si se desea examinar el efecto que tales comportamientos tienen sobre la condición de desempleo, los ingresos o el comportamiento delictivo, debe tomarse en cuenta que el uso de drogas puede estar correlacionado con otros factores que pueden tener efectos sobre el empleo y la delincuencia. Los niños que reúnen las características para participar en programas como Head Start (un programa de inserción social) participan con base en la decisión de los padres. Dado que los antecedentes familiares son importantes en la decisión de participar en Head Start y afectan también los resultados obtenidos por el estudiante, al examinar el efecto de Head Start deben controlarse estos factores [vea, por ejemplo, Currie y Thomas (1995)]. Las personas elegidas por los patrones o por instituciones del gobierno para participar en los programas de capacitación para el trabajo pueden decidir participar o no, y esta decisión es poco probable que sea aleatoria [vea, por ejemplo, Lynch (1992)]. Las ciudades y los estados deciden si ponen en marcha ciertas leyes para el control de armas de fuego, y es muy probable que esta decisión esté sistemáticamente relacionada con otros factores que afectan la delincuencia con violencia [vea, por ejemplo, Kleck y Patterson (1993)].

En el párrafo anterior se dan ejemplos de lo que en economía generalmente se conoce como problemas de **autoselección**. Literalmente, el término proviene del hecho de que las personas se autoseleccionan para participar en ciertas conductas o en ciertos programas: esta participación no es determinada de forma aleatoria. Este término suele usarse cuando un indicador binario de participación puede estar sistemáticamente relacionado con factores no observados. Así, si se da el sencillo modelo

$$y = \beta_0 + \beta_1 partic + u, \tag{7.34}$$

donde y es una variable de resultado y *partic* es una variable binaria igual a uno si la persona, empresa o ciudad participa en un comportamiento o en un programa o adopta cierto tipo de ley; entonces preocupará que el valor promedio de u dependa de la participación: $E(u|partic = 1) \neq E(u|partic = 0)$. Como se sabe, esto hace que el estimador β_1 en la regresión simple sea sesgado y de esta manera no se podrá hallar el verdadero efecto de la participación. Por tanto, el problema de la autoselección es otra manera en que la variable explicativa (*partic* en este caso) puede ser endógena.

Se sabe, por lo visto hasta ahora, que el análisis de regresión múltiple puede, hasta cierto grado, mitigar el problema de la autoselección. Los factores en el término del error de (7.34) que estén correlacionados con *partic* pueden incluirse en una ecuación de regresión múltiple suponiendo, por supuesto, que puedan ser recolectados datos sobre estos factores. Por desgracia, en muchos casos, preocupa que los factores no observados estén relacionados con la participación, en cuyo caso la regresión múltiple produce estimadores sesgados.

En el análisis usual de regresión múltiple usando datos de corte transversal, hay que estar atentos a efectos espurios de los programas sobre las variables de resultado, causados por el problema de la autoselección. Un buen ejemplo de esto se encuentra en Currie y Cole (1993). Estos autores examinan el efecto de la participación en el programa Ayuda a Familias con Hijos Dependientes (AFDC, por sus siglas en inglés) sobre el peso del niño al nacer. Aun después de controlar diversas características familiares y de educación, los autores obtuvieron estimaciones de MCO que implican que la participación en el programa AFDC *reduce* el peso del niño al nacer. Como indican los autores, es difícil creer que la participación en el programa AFDC *ocasione* esa reducción del peso al nacer. [Vea más ejemplos en Currie, 1995.] Empleando otro método econométrico que se verá en el capítulo 15, Currie y Cole encontraron evidencias de ningún efecto o de un efecto *positivo* de la participación en el programa AFDC sobre el peso de un niño al nacer.

Cuando el problema de la autoselección ocasiona que el análisis de regresión múltiple sea sesgado debido a la falta de suficientes variables de control, puede hacerse uso de los métodos más avanzados que se ven en los capítulos 13, 14 y 15.

RESUMEN

En este capítulo se vio cómo usar información cualitativa en el análisis de regresión. En el caso más sencillo, se define una variable binaria para distinguir entre dos grupos y el coeficiente estimado de la variable binaria estima la diferencia *ceteris paribus* entre los dos grupos. Definiendo un conjunto de variables binarias se pueden considerar más de dos grupos: si hay g grupos, entonces en el modelo se incluyen $g - 1$ variables binarias. Todas las estimaciones de las variables binarias se interpretan con relación al grupo base o al grupo de referencia (el grupo para el que en el modelo no se incluye una variable binaria).

Las variables binarias también son útiles para incorporar información ordinal, como calificación de un crédito o de la belleza, a los modelos de regresión. Simplemente se define un conjunto de variables binarias que representan los diferentes valores de la variable ordinal, dejando una de las categorías como grupo base.

Las variables binarias pueden interactuar con las variables cuantitativas para poder considerar diferencias de pendientes entre los diferentes grupos. En caso extremo, cada grupo puede tener su propia pendiente en cada variable, así como su propio intercepto. Para detectar si hay diferencias entre los grupos puede emplearse la prueba de Chow. En muchos casos es más interesante probar si, después de considerar diferencia en el intercepto, las pendientes de dos grupos diferentes son iguales. Para este propósito puede usarse una prueba F estándar con un modelo no restringido en el que se incluyan las interacciones entre la binaria de grupo y todas las variables.

El modelo de probabilidad lineal, que se estima simplemente mediante MCO, permite explicar una respuesta binaria empleando el análisis de regresión. En este caso las estimaciones de MCO se interpretan como variaciones en la probabilidad de "éxito" ($y = 1$), dado un aumento de una unidad en la variable explicativa correspondiente. El MPL tiene algunas desventajas: se pueden obtener probabilidades predichas que sean menores a cero o mayores a uno, implica un efecto marginal constante de cada variable explicativa que aparece en su forma original, y contiene heterocedasticidad. Los dos primeros problemas suelen no ser serios cuando se obtienen estimaciones de las variables explicativas en el rango central de los datos. La heterocedasticidad invalida los errores estándar usuales de MCO y los estadísticos de prueba, pero como se verá en el capítulo siguiente, esto se resuelve fácilmente cuando las muestras son suficientemente grandes.

Este capítulo terminó con un análisis de cómo usar variables binarias para evaluar políticas y programas. Como en todo el análisis de regresión, debe recordarse que la participación en programas, o algún otro regresor binario con implicaciones de políticas, puede estar correlacionado con factores no observados que afecten a la variable dependiente, dando como resultado el usual sesgo de variable omitida.

TÉRMINOS CLAVE

Análisis de políticas
Autoselección
Desplazamiento del intercepto
Diferencia de pendientes
Estadístico de Chow
Evaluación de programa
Grupo base
Grupo control
Grupo de referencia
Grupo de tratamiento
Grupo experimental
Modelo de probabilidad lineal (MPL)
Porcentaje predicho correctamente
Probabilidad de respuesta
R-cuadrada descentrada
Término de interacción
Trampa de la variable binaria
Variable binaria
Variable ordinal
Variables binarias o *dummy*

PROBLEMAS

7.1 Con ayuda de los datos del archivo SLEEP75.RAW (vea también el problema 3.3) se obtiene la ecuación estimada

$$\begin{aligned}
\widehat{sleep} = {} & \underset{(235.11)}{3{,}840.83} - \underset{(.018)}{.163}\, totwrk - \underset{(5.86)}{11.71}\, educ - \underset{(11.21)}{8.70}\, age \\
& + \underset{(.134)}{.128}\, age^2 + \underset{(34.33)}{87.75}\, male \\
& n = 706,\ R^2 = .123,\ \bar{R}^2 = .117.
\end{aligned}$$

La variable *sleep* es la cantidad total de minutos, por semana, dormidos durante la noche, *totwrk* es la cantidad total de minutos que se trabajó por semana, *educ* y *age* (edad) están dadas en años y *male* (hombre) es una binaria para género.

i) Permaneciendo todo lo demás constante, ¿hay alguna evidencia de que los hombres duerman más que las mujeres? ¿Qué tan fuerte es esa evidencia?
ii) ¿Hay un costo de oportunidad estadísticamente significativo entre trabajar y dormir? ¿Cuál es el costo de oportunidad estimado?
iii) ¿Qué otra regresión necesita correr para probar la hipótesis nula de que, manteniendo todos los demás factores constantes, la edad no afecta el dormir?

7.2 Las ecuaciones siguientes se estimaron empleando los datos del archivo BWGHT.RAW:

$$\begin{aligned}
\widehat{\log(bwght)} = {} & \underset{(.22)}{4.66} - \underset{(.0009)}{.0044}\, cigs + \underset{(.0059)}{.0093} \log(faminc) + \underset{(.006)}{.016}\, parity \\
& + \underset{(.010)}{.027}\, male + \underset{(.013)}{.055}\, white \\
& n = 1{,}388,\ R^2 = .0472
\end{aligned}$$

y

$$\begin{aligned}
\widehat{\log(bwght)} = {} & \underset{(.38)}{4.65} - \underset{(.0010)}{.0052}\, cigs + \underset{(.0085)}{.0110} \log(faminc) + \underset{(.006)}{.017}\, parity \\
& + \underset{(.011)}{.034}\, male + \underset{(.015)}{.045}\, white - \underset{(.0030)}{.0030}\, motheduc + \underset{(.0026)}{.0032}\, fatheduc \\
& n = 1{,}191,\ R^2 = .0493.
\end{aligned}$$

Las variables se han definido como en el ejemplo 4.9, pero se ha agregado una variable binaria que indica si el niño es varón (*male*) y que indica si el niño se clasifica como blanco (*white*).

i) Interprete el coeficiente de la variable *cigs* en la primera ecuación. En particular, ¿qué efecto tiene sobre el peso al nacer que la madre haya fumado 10 cigarros más por día?

ii) En la primera ecuación, permaneciendo todos los demás factores constantes, ¿cuánto más se predice que pese un niño blanco en comparación con uno no blanco? ¿Es esta diferencia significativa?

iii) Analice el efecto estimado y la significancia estadística de *motheduc*.

iv) Con base en la información dada, ¿por qué no puede usted calcular el estadístico F para la significancia conjunta de *motheduc* y *fatheduc*? ¿Qué tendría que hacer para calcular este estadístico F?

7.3 Con ayuda de los datos del archivo GPA2.RAW se obtuvo la ecuación estimada siguiente:

$$\begin{aligned}
\widehat{sat} = {} & \underset{(6.29)}{1{,}028.10} + \underset{(3.83)}{19.30}\, hsize - \underset{(.53)}{2.19}\, hsize^2 - \underset{(4.29)}{45.09}\, female \\
& - \underset{(12.71)}{169.81}\, black + \underset{(18.15)}{62.31}\, female \cdot black \\
n = {} & 4{,}137,\ R^2 = .0858.
\end{aligned}$$

La variable *sat* es la puntuación combinada en el examen de admisión (SAT), *hsize* es la cantidad de alumnos, dada en cientos, que en la escuela del estudiante terminaron con él el bachillerato, *female* es una variable binaria para el género femenino y *black* es una variable binaria para raza, que es igual a uno para negros y cero para los no negros.

i) ¿Existe alguna evidencia fuerte de que $hsize^2$ deba ser incluida en el modelo? Con base en esta ecuación, ¿cuál es el tamaño óptimo de una escuela?

ii) Manteniendo *hsize* constante, ¿cuál es la diferencia estimada en la puntuación del SAT entre mujeres no negras y hombres no negros? ¿Qué tan estadísticamente significativa es esta diferencia?

iii) ¿Cuál es la diferencia estimada en la puntuación del SAT entre hombres no negros y hombres negros? Pruebe la hipótesis nula de que no hay diferencia entre estas puntuaciones contra la alternativa de que sí hay diferencia.

iv) ¿Cuál es la diferencia estimada en la puntuación del SAT entre mujeres negras y mujeres no negras? ¿Qué se necesita hacer para probar si esta diferencia es estadísticamente significativa?

7.4 Una ecuación que explica el sueldo de los presidentes de consejos de administración es

$$\begin{aligned}
\widehat{\log(salary)} = {} & \underset{(.30)}{4.59} + \underset{(.032)}{.257}\, \log(sales) + \underset{(.004)}{.011}\, roe + \underset{(.089)}{.158}\, finance \\
& + \underset{(.085)}{.181}\, consprod - \underset{(.099)}{.283}\, utility \\
n = {} & 209,\ R^2 = .357.
\end{aligned}$$

Los datos que se emplearon son los del archivo CEOSAL1.RAW, donde *finance*, *consprod* y *utility* son variables binarias que corresponden a las industrias financiera, de productos de consumo y de servicios. La industria que se ha omitido es la del transporte.

i) Calcule la diferencia porcentual aproximada entre los sueldos estimados de las industrias de servicios y de transporte, manteniendo *sales* y *roe* constantes. ¿Es esta diferencia estadísticamente significativa al nivel de significancia de 1%?

ii) Emplee la ecuación (7.10) para obtener una diferencia porcentual exacta entre los sueldos estimados de las industrias de servicios y de transporte, y compare el resultado con el obtenido en el inciso i).

iii) ¿Cuál es la diferencia porcentual aproximada entre los sueldos estimados de la industria de productos de consumo y en el sector financiero? Dé una ecuación que permita probar si esta diferencia es estadísticamente significativa.

7.5 En el ejemplo 7.2 sea *noPC* una variable binaria igual a uno si el estudiante no posee una PC e igual a cero en caso contrario.

i) Si en la ecuación (7.6) se usa *noPC* en lugar de *PC*, ¿qué pasa con el intercepto de la ecuación estimada? ¿Cuál será el coeficiente de *noPC*? (*Sugerencia:* escriba $PC = 1 - noPC$ y sustituya esto en la ecuación $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

ii) ¿Qué ocurre con la *R*-cuadrada si en lugar de *PC* se usa *noPC*?

iii) ¿Deben incluirse ambas, *PC* y *noPC* como variables independientes en el modelo? Explique.

7.6 Para probar la eficacia de un programa de capacitación laboral sobre los salarios de los trabajadores, se especifica el modelo

$$\log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u,$$

donde *train* es una variable binaria que es igual a uno si el trabajador participó en el programa. Considere que el término de error *u* comprende capacidades no observadas del trabajador. Si los trabajadores con menos capacidades tienen más probabilidades de ser elegidos para participar en el programa y se emplea un análisis por MCO, ¿qué se puede decir acerca del posible sesgo en β_1? (*Sugerencia:* como referencia vuelva al capítulo 3.)

7.7 En la ecuación (7.29), suponga que se define *outlf* igual a uno si la mujer se encuentra fuera de la fuerza de trabajo y cero si no es así.

i) Si se regresa *outlf* sobre todas las variables independientes de la ecuación (7.29), ¿qué pasará con el intercepto y con las pendientes estimadas? (*Sugerencia:* $inlf = 1 - outlf$. Sustituya esto en la ecuación poblacional $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \ldots$ y reordene.)

ii) ¿Qué pasará con los errores estándar del intercepto y de las pendientes estimadas?

iii) ¿Qué pasará con la *R*-cuadrada?

7.8 Suponga que mediante una encuesta recolecta usted datos sobre salarios, educación, experiencia y género. Solicita también información sobre uso de la mariguana. La pregunta original es "¿Cuántas veces fumó mariguana el mes pasado?".

i) Dé una ecuación que permita estimar el efecto de fumar mariguana sobre el salario, controlando los demás factores. La ecuación deberá permitir hacer afirmaciones como "se estima que fumar mariguana cinco veces más al mes hace que el salario varíe *x*%".

ii) Formule una ecuación que permita probar si el uso de drogas tiene efectos diferentes sobre los salarios de hombres y mujeres. ¿Cómo puede probarse que los efectos del uso de drogas no son diferentes entre hombres y mujeres?

iii) Suponga que considera que es mejor medir el uso de mariguana clasificando a las personas en cuatro categorías: no usuario, usuario suave (1 a 5 veces por mes), usuario moderado (6 a 10 veces por mes) y usuario fuerte (más de 10 veces por mes). Ahora, diseñe un modelo que permita estimar el efecto del uso de la mariguana sobre el salario.

iv) Usando el modelo del inciso iii), explique detalladamente cómo probar la hipótesis nula de que fumar mariguana no tiene ningún efecto sobre el salario. Sea muy específico y dé una lista detallada de los grados de libertad.

v) ¿Cuáles son algunos de los problemas potenciales de hacer inferencias causales a partir de los datos recolectados por usted?

7.9 Sea d una variable binaria y sea z una variable cuantitativa. Considere el modelo

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

esta es la versión general de un modelo con una interacción entre una variable binaria y una variable cuantitativa. [La ecuación (7.17) es un ejemplo.]

i) Como no se altera nada, haga el error igual a cero, $u = 0$. Entonces, cuando $d = 0$ la relación entre y y z puede expresarse mediante la función $f_0(z) = \beta_0 + \beta_1 z$. Escriba la misma relación para el caso en que $d = 1$, donde, en el lado izquierdo debe usar $f_1(z)$ para denotar la función lineal de z.

ii) Suponiendo que $\delta_1 \neq 0$ (lo que significa que las dos rectas no son paralelas), muestre que el valor z^* para el que $f_0(z^*) = f_1(z^*)$ es $z^* = -\delta_0/\delta_1$. Este es el punto en el que se interceptan las dos rectas [como en la figura 7.2(b)]. Argumente que z^* es positivo si y sólo si δ_0 y δ_1 tienen signos contrarios.

iii) Empleando los datos del archivo TWOYEAR.RAW, puede estimarse la ecuación siguiente:

$$\widehat{\log(wage)} = \underset{(0.011)}{2.289} - \underset{(.015)}{.357}\, female + \underset{(.003)}{.50}\, totcoll + \underset{(.005)}{.030}\, female \cdot totcoll$$

$$n = 6{,}763,\ R^2 = .202,$$

donde todos los coeficientes y desviaciones estándar se han redondeado a dos cifras decimales. Empleando esta ecuación, encuentre un valor de *totcoll*, tal que los valores que se predicen para log(*wage*) sean iguales para hombres y para mujeres.

iv) Con base en la ecuación del inciso iii), ¿es realmente posible que las mujeres logren suficientes años de universidad de manera que sus ingresos estén al nivel de los de los hombres? Explique.

7.10 Dado un niño i que vive en un determinado distrito escolar, sea $voucher_i$ una variable binaria que sea igual a uno si el niño es elegido para participar en un programa de vales escolares, y sea $score_i$ la puntuación del niño en un examen estandarizado subsecuente. Suponga que la variable de participación $voucher_i$, es completamente aleatorizada en el sentido de que es independiente, tanto de los factores observados como de los no observados que pueden afectar la puntuación del examen.

i) Si corre una regresión simple de $score_i$ sobre $voucher_i$ empleando una muestra aleatoria de tamaño n, ¿proporciona el estimador de MCO un estimador insesgado del efecto del programa de vales?

ii) Suponga que logra obtener información adicional sobre antecedentes del niño, tales como ingreso familiar, estructura familiar (por ejemplo, si el niño vive con sus padres) y nivel de educación de los padres. ¿Necesita controlar estos factores para obtener un estimador insesgado de los efectos del programa de vales? Explique.

iii) ¿Cuál es la razón para incluir en la regresión las variables sobre los antecedentes familiares? ¿Hay alguna situación en la que usted no incluiría tales variables?

EJERCICIOS EN COMPUTADORA

C7.1 Para este ejercicio emplee los datos del archivo GPA1.RAW.

i) A la ecuación estimada en (7.6) agregue las variables *mothcoll* y *fathcoll* y dé los resultados en la forma habitual. ¿Qué pasa con el efecto estimado de la posesión de una PC? ¿Sigue siendo *PC* estadísticamente significativa?

ii) Pruebe la significancia conjunta de *mothcoll* y *fathcoll* en la ecuación del inciso i) y no olvide dar el valor-*p*.

iii) Al modelo del inciso i) agregue $hsGPA^2$ y decida si es necesaria esta generalización.

C7.2 Para este ejercicio emplee los datos del archivo WAGE2.RAW.

i) Estime el modelo

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

y dé el resultado en la forma habitual. Manteniendo todos los demás factores constantes, ¿cuál es la diferencia aproximada entre el salario mensual de negros y no negros? ¿Es esta diferencia estadísticamente significativa?

ii) Agregue a esta ecuación las variables $exper^2$ y $tenure^2$ y muestre que no son conjuntamente significativas al nivel de 20%.

iii) Amplíe el modelo original de manera que el rendimiento a la educación dependa de la raza y pruebe si en realidad el rendimiento de la educación depende de la raza.

iv) Partiendo nuevamente del modelo original, permita que los salarios difieran entre cuatro grupos: casados negros, casados no negros, solteros negros y solteros no negros. ¿Cuál es la diferencia de salario estimada entre negros casados y no negros casados?

C7.3 Un modelo en el que se consideran variaciones en los sueldos de los jugadores de la liga mayor de béisbol de acuerdo con su posición es

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} catcher + u,$$

donde el grupo base es jardineros (*outfield*).

i) Dé la hipótesis nula que afirma que, controlando los demás factores, receptores (*catchers*) y jardineros ganan, en promedio lo mismo. Pruebe esta hipótesis empleando los datos del archivo MLB1.RAW y analice la magnitud de la diferencia salarial estimada.

ii) Dé y pruebe la hipótesis nula que afirma que, una vez controlados otros factores, no hay diferencia entre los sueldos de acuerdo con las posiciones.

iii) ¿Hay coherencia entre los resultados de los incisos i) y ii)? Si no es así, explique lo que sucede.

C7.4 Para este ejercicio emplee los datos del archivo GPA2.RAW.

i) Considere la ecuación

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

en donde *colgpa* es el promedio acumulado de calificaciones en la universidad, *hsize* es el tamaño del grupo de graduación del bachillerato dado en cientos, *hsperc* es el percentil académico dentro del grupo de graduados de bachillerato, *sat* es la puntuación en la prueba SAT combinada de admisión a la universidad, *female* es una variable binaria para género femenino y *athlete* es una variable binaria, que es uno para estudiantes atletas. ¿Cómo espera que sean los coeficientes de esta ecuación? ¿De cuáles no está seguro?

ii) Estime la ecuación del inciso i) y dé los resultados en la forma usual. ¿Cuál es la diferencia estimada entre las calificaciones de los atletas y de los no atletas? ¿Es esta diferencia estadísticamente significativa?

iii) Elimine *sat* de este modelo y vuelva a estimar la ecuación. ¿Cuál es ahora el efecto estimado de ser atleta? Analice por qué esta estimación es distinta de la obtenida en el inciso ii).

iv) En el modelo del inciso i), tome en consideración que el efecto de ser atleta difiera entre hombres y mujeres y pruebe la hipótesis nula de que no hay ninguna diferencia *ceteris paribus* entre mujeres atletas y mujeres no atletas.

v) ¿Difiere el efecto de *sat* sobre *colgpa* con base en el género? Justifique su respuesta.

C7.5 En el problema 4.2 al modelo que explica el sueldo de los CEO se agregó el rendimiento de las acciones de una empresa, *ros*, el cual resultó ser no significativo. Ahora, defina una variable binaria, *rosneg*, que sea igual a uno si $ros < 0$ e igual a cero si $ros \geq 0$. Utilice CEOSAL1.RAW para estimar el modelo

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 rosneg + u.$$

Analice la interpretación y la significancia estadística de $\hat{\beta}_3$.

C7.6 Para este ejercicio use la base de datos SLEEP75.RAW. La ecuación de interés es

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

i) Estime esta ecuación por separado para hombres y mujeres y dé los resultados de la manera habitual. ¿Hay diferencias importantes entre las dos ecuaciones estimadas?

ii) Realice la prueba de Chow para la igualdad entre los parámetros para hombres y mujeres en la ecuación del sueño. Utilice la forma de la prueba en la que se agrega *male* (hombre) y los términos de interacción *male*·*totwrk*, …, *male*·*yngkid* y se usa el conjunto completo de observaciones. ¿Cuáles son los *gl* pertinentes en esta prueba? ¿Debe rechazarse la hipótesis nula al nivel de significancia de 5%?

iii) Ahora tome en consideración interceptos diferentes entre hombres y mujeres y determine si los términos de interacción en los que parece *male* son conjuntamente significativos.

iv) Dados los resultados de los incisos ii) y iii), ¿cuál es el modelo final?

C7.7 Para este ejercicio utilice los datos del archivo WAGE1.RAW.

i) Utilice la ecuación (7.18) para estimar la diferencia entre los géneros cuando $educ = 12.5$. Compare con la diferencia estimada cuando $educ = 0$.

ii) Corra la regresión usada para obtener (7.18), pero con $female \cdot (educ - 12.5)$ en vez de $female \cdot educ$. ¿Cómo interpreta ahora el coeficiente de *female*?

iii) En el inciso ii), ¿es el coeficiente de *female* estadísticamente significativo? Compare con (7.18) y comente.

C7.8 Para este ejercicio utilice los datos del archivo LOANAPP.RAW. La variable binaria a explicar es *approve*, que es igual a uno si a un individuo se le aprueba el préstamo hipotecario. La variable explicativa clave es *white*, una variable binaria igual a uno si el solicitante es blanco. En esta base de datos, los demás solicitantes son negros e hispanos.

Para probar si hay discriminación en el mercado de préstamos hipotecarios, puede emplearse un modelo de probabilidad lineal:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

i) Si existe discriminación contra las minorías y se han controlado los factores adecuados, ¿cuál es el signo de β_1?

ii) Regrese *approve* sobre *white* y dé los resultados en la forma habitual. Interprete el coeficiente de *white* ¿es estadísticamente significativo? ¿Es grande en sentido práctico?

iii) Agregue, como controles, las variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2* y *vr*. ¿Qué ocurre con el coeficiente de *white*? ¿Sigue habiendo evidencias de discriminación contra las personas que no son blancas?

iv) Ahora permita que el efecto de la raza interactúe con la variable que mide otras obligaciones como porcentaje del ingreso (*obrat*). ¿Es significativo este término de interacción?

v) Utilizando el modelo del inciso (iv), ¿qué efecto tiene ser blanco sobre la probabilidad de aprobación cuando $obrat = 32$, que es aproximadamente el valor medio en la muestra? Obtenga un intervalo de confianza de 95% para este efecto.

C7.9 Ha existido un gran interés por determinar si la presencia del plan de pensiones 401(k), disponible para muchos trabajadores de Estados Unidos, incrementa los ahorros netos. La base de datos 401KSUBS.RAW contiene información sobre activos financieros (*nettfa*), ingreso familiar (*inc*), una variable binaria para elegibilidad en un plan 401(k) plan (*e401k*), y otras variables.

i) ¿Qué proporción de las familias de la muestra son elegibles para participar en un plan 401(k)?

ii) Estime un modelo de probabilidad lineal que explique la elegibilidad para un plan 401(k) en términos de ingreso, edad y género. Incluya ingreso y edad en forma cuadrática y dé los resultados en la forma habitual.

iii) ¿Diría usted que la elegibilidad para un plan 401(k) es independiente del ingreso y de la edad? ¿Y del género? Explique.

iv) Obtenga los valores ajustados para el modelo de probabilidad lineal estimado en el inciso ii). ¿Es negativo o mayor a uno alguno de los valores ajustados?

v) Utilizando los valores ajustados $\widehat{e401k}_i$ del inciso iv), defina $\widetilde{e401k}_i = 1$ si $\widehat{2e401k} \geq .5$ y $\widetilde{401k} = 0$ si $\widehat{2e401k} < .5$. De 9,275 familias, ¿cuántas se predice que sean elegibles para un plan 401(k)?

vi) De las 5,638 familias que no son elegibles para un plan 401(k), empleando el predictor $\widetilde{e401k}_i$ ¿qué porcentaje se predice que no tenga un plan 401(k)? De las 3,637 familias elegibles para un plan 401(k), ¿qué porcentaje se predice que tenga uno? (Es útil que su paquete para econometría tenga el comando “tabulate”.)

vii) El porcentaje general que se predice de manera correcta es aproximadamente 64.9% ¿Considera que esta sea una descripción completa de qué tan bien funciona el modelo, dada su respuesta del inciso vi)?

viii) A este modelo de probabilidad lineal agregue como variable explicativa la variable binaria *pira*. Permaneciendo todo lo demás constante, si en una familia hay alguien que tenga una cuenta de retiro individual ($pira = 1$), ¿qué tanto mayor es la probabilidad estimada de que la familia sea elegible para un plan 401(k)? ¿Es distinta de cero al nivel de 10%?

C7.10 Para este ejercicio emplee los datos del archivo NBASAL.RAW.

i) Estime el modelo de regresión lineal que relaciona puntos por juego (*points*) con experiencia en la liga (*exper*) y posición (defensa, delantero o centro —*guard*, *forward*, *center*—). Incluya experiencia en forma cuadrática y como grupo base use *center*. Dé los resultados en la forma habitual.

ii) En el inciso i), ¿por qué no incluye las tres variables binarias para las posiciones?

iii) Permaneciendo experiencia constante, ¿anota más un defensa que un centro? ¿Cuánto más? ¿Es estadísticamente significativa esta diferencia?

iv) Ahora a la ecuación agregue el estado civil ($marr = 1$ si es casado). Manteniendo constantes la posición y la experiencia, ¿son más productivos los jugadores casados (de acuerdo con los puntos por juego)?

v) Agregue interacciones del estado civil con las dos variables de experiencia. En este modelo ampliado, ¿hay una fuerte evidencia de que el estado civil tenga algún efecto sobre los puntos por juego?

vi) Estime el modelo del inciso iv) pero use las asistencias por juego (*assists*) como variable dependiente. ¿Hay alguna diferencia notable con el inciso iv)? Analice.

C7.11 Para este ejercicio emplee los datos del archivo 401KSUBS.RAW.

i) Calcule promedio, desviación estándar y los valores mínimo y máximo de *nettfa* en la muestra.

ii) Pruebe la hipótesis de que el *nettfa* promedio no difiere de acuerdo con la elegibilidad para un plan 401(k): emplee una alternativa de dos colas. ¿Cuál es el monto en dólares de la diferencia estimada?

iii) De acuerdo con el inciso ii) del ejercicio C7.9, es claro que en un modelo de regresión simple, *e401k* no es exógeno: como mínimo varía de acuerdo con el ingreso y con la edad. Estime un modelo de regresión lineal múltiple para *nettfa* que como variables explicativas incluya ingreso, edad y *e401k*. Las variables ingreso y edad deben aparecer en forma cuadrática. Ahora, ¿cuál es el efecto en dólares de la elegibilidad a un plan 401(k)?

iv) Al modelo estimado en el inciso iii), agregue las interacciones $e401k \cdot (age - 41)$ y $e401k \cdot (\text{age} - 41)^2$. Observe que en la muestra, la edad promedio es aproximadamente 41 años, de manera que en el nuevo modelo, el coeficiente de *e401k* es el efecto estimado de la elegibilidad para el plan 401(k) a la edad promedio. ¿Cuál de los términos de interacción es significativo?

v) Compare las estimaciones de los incisos iii) y iv), ¿difieren mucho los efectos estimados de la elegibilidad para el plan 401(k) a la edad de 41 años? Explique.

vi) Ahora elimine del modelo los términos de interacción, pero defina cinco variables binarias para tamaño de la familia: *fsize1, fsize2, fsize3, fsize4* y *fsize5*. La variable *fsize5* es uno para familias que tengan de cinco o más miembros. Incluya las variables binarias del tamaño de familia al modelo estimado en el inciso iii); no olvide elegir un grupo base. ¿Las variables binarias son significativas al nivel de 1%?

vii) Ahora, realice una prueba Chow para el modelo

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

para las cinco categorías de tamaños de familias, considerando diferencias en el intercepto. La suma restringida de residuales cuadrados, SRC_r, se obtiene del inciso vi) porque en esa regresión se supone que todas las pendientes son iguales. La suma no restringida de residuales cuadrados es $\text{SRC}_{ur} = \text{SRC}_1 + \text{SRC}_2 + ... + \text{SRC}_5$, donde SRC_f es la suma de residuales cuadrados para la ecuación estimada usando sólo el tamaño *f* de la familia. Debe convencerse de que en el modelo no restringido hay 30 parámetros (5 interceptos más 25 pendientes) y 10 parámetros en el modelo restringido (5 interceptos más 5 pendientes). Por tanto, la cantidad de restricciones que se prueban es $q = 20$, y los *gl* para el modelo no restringido son $9{,}275 - 30 = 9{,}245$.

C7.12 Use la base de datos BEAUTY.RAW, que contiene un subconjunto de las variables (pero más observaciones útiles que en la regresión) reportadas por Hamermesh y Biddle (1994).

i) Encuentre las proporciones de hombres y de mujeres clasificados por separado, como con un aspecto físico superior al promedio. ¿Hay más personas clasificadas con aspecto físico superior al promedio o con aspecto físico inferior al promedio?

ii) Pruebe la hipótesis nula de que en la población las proporciones con un aspecto físico superior al promedio son iguales entre hombres y mujeres. Dé el valor-*p* para la hipótesis de una cola que la proporción de mujeres es mayor. (*Sugerencia:* estimar un modelo de probabilidad lineal simple es lo más fácil.)

iii) Ahora estime el modelo

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

por separado para hombres y para mujeres y dé los resultados en la forma habitual. En ambos casos interprete el coeficiente de *belavg*. Explique en palabras lo que significan las hipótesis H_0: $\beta_1 = 0$ contra H_1: $\beta_1 < 0$ y determine los valores-*p* para hombres y mujeres.

iv) ¿Hay alguna evidencia convincente de que las mujeres con una apariencia superior al promedio ganen más que las mujeres con una apariencia promedio? Explique.

v) Tanto para hombres como para mujeres, agregue las variables explicativas *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity* y *service*. ¿Cambian de manera importante los efectos de las variables de apariencia?

C7.13 Para responder esta pregunta, utilice los datos del archivo APPLE.RAW.

i) Defina una variable binaria $ecobuy = 1$ si $ecolbs > 0$ y $ecobuy = 0$ si $ecolbs = 0$. En otras palabras, *ecobuy* indica si, a los precios dados, una familia comprará manzanas ecológicas. ¿Cuál es la proporción de familias que asegura que comprará manzanas ecoetiquetadas?

ii) Estime el modelo de probabilidad lineal

$$\begin{aligned} ecobuy = {} & \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc \\ & + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u, \end{aligned}$$

y dé los resultados de la manera habitual. Interprete cuidadosamente los coeficientes de las variables de precio (*ecoprc* es el precio de las manzanas ecoetiquetadas y *regprc* el de las manzanas regulares).

iii) ¿Son conjuntamente significativas las variables que no corresponden a precios en el MPL? (Emplee el estadístico *F* usual, aun cuando no sea válido cuando existe heterocedasticidad.) ¿Cuál es la variable explicativa, además de las variables relacionadas con el precio, que parece tener un efecto más importante sobre la decisión de comprar manzanas ecológicas? ¿Le parece que esto tiene sentido?

iv) En el modelo del inciso ii) sustituya *faminc* por log(*faminc*). ¿Qué modelo se ajusta mejor a los datos, el modelo con *faminc* o el modelo con log(*faminc*)? Interprete el coeficiente de log(*faminc*).

v) En la estimación del inciso iv), ¿cuántas probabilidades estimadas son negativas? ¿Cuántas son mayores a uno? ¿Debe preocuparle?

vi) Dada la estimación del inciso iv), calcule el porcentaje predicho correctamente para cada uno de los resultados, $ecobuy = 0$ y $ecobuy = 1$. ¿Cuál de estos resultados se predice mejor con este modelo?

C7.14 Para esta pregunta emplee los datos del archivo CHARITY.RAW. La variable *respond* es una variable binaria igual a uno si la persona responde con una contribución al correo más reciente enviado por una organización de caridad. La variable *resplast* es una variable binaria igual a uno si la persona respondió al correo previo, *avggift* es el promedio de donaciones anteriores (en florines holandeses) y la variable *propresp* es la proporción de veces que la persona ha respondido a correos anteriores.

i) Estime el modelo de probabilidad lineal que relaciona *respond* con *resplast* y *avggift*. Dé los resultados de la manera habitual e interprete el coeficiente de *resplast*.

ii) ¿El valor promedio de las donaciones anteriores parece afectar la probabilidad de respuesta?

iii) Agregue la variable *propresp* al modelo e interprete su coeficiente. (Tenga cuidado aquí: un aumento de uno en *propresp* es la mayor variación posible.)

iv) ¿Qué pasó con el coeficiente de *resplast* al agregar *propresp* al modelo? ¿Tiene sentido esto?

v) Agregue al modelo *mailsyear*, número de correos por año. ¿Qué tan grande es su efecto estimado? ¿Por qué esto puede no ser una buena estimación del efecto causal del número de correos sobre la respuesta?

CAPÍTULO 8

Heterocedasticidad

El supuesto de homocedasticidad, introducido en el capítulo 3 de regresión múltiple, establece que la varianza del error no observable, u, condicional sobre las variables explicativas, es constante. La homocedasticidad no se satisface cuando la varianza de los no observables varía en los diversos segmentos de la población, donde los segmentos están determinados por los diversos valores de las variables explicativas. Por ejemplo, en una ecuación del ahorro, habrá heterocedasticidad si la varianza de los factores no observados que afectan el ahorro aumenta con el ingreso.

En los capítulos 4 y 5, se vio que la homocedasticidad es necesaria para justificar las pruebas t y F usuales y los intervalos de confianza para la estimación por MCO del modelo de regresión lineal, incluso con tamaños de muestra grandes. En este capítulo se verán los remedios con que se cuenta cuando se presenta heterocedasticidad, y se mostrará también cómo probar su presencia. Se comenzará por repasar brevemente las consecuencias de la heterocedasticidad en las estimaciones por mínimos cuadrados ordinarios.

8.1 Consecuencias de la heterocedasticidad para MCO

Considere de nuevo el modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \tag{8.1}$$

En el capítulo 3 se probó el insesgamiento o insesgadez de los estimadores de MCO $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, …, $\hat{\beta}_k$ bajo los primeros cuatro supuestos, RLM.1 a RLM.4, de Gauss-Markov. En el capítulo 5 se mostró que estos mismos cuatro supuestos implican la consistencia de MCO. El supuesto de homocedasticidad RLM.5, dado en términos de la varianza del error como $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$, no desempeñó papel alguno al demostrar si los MCO eran insesgados o consistentes. Es importante recordar que la heterocedasticidad no ocasiona sesgo ni inconsistencia en los estimadores MCO de las β_j, mientras que omitir una variable importante sí tendrá este efecto.

La interpretación de las medidas de bondad de ajuste, R^2 y $\bar{R}^2$, tampoco se ve afectada por la presencia de heterocedasticidad. ¿Por qué? Recuérdese que en la sección 6.3 se vio que la R-cuadrada usual y la R-cuadrada ajustada son dos distintas maneras de estimar la R-cuadrada poblacional, la cual es simplemente $1 - \sigma_u^2/\sigma_y^2$, donde σ_u^2 es la varianza poblacional del error y σ_y^2 es la varianza poblacional de y. El punto clave es que como en la R-cuadrada poblacional ambas varianzas son incondicionales, la R-cuadrada poblacional no se ve afectada por la presencia de heterocedasticidad en $\text{Var}(u|x_1, \dots, x_k)$. Además, SRC/$n$ estima consistentemente σ_u^2, y STC/n estima consistentemente σ_y^2, al margen de si $\text{Var}(u|x_1, \dots, x_k)$ es constante. Lo mismo es cierto cuando se ajustan los grados de libertad. Por tanto, R^2 y $\bar{R}^2$ son estimadores consistentes de la R-cuadrada poblacional, se satisfaga o no el supuesto de homocedasticidad.

Si la heterocedasticidad no causa sesgo ni inconsistencia en los estimadores de MCO, ¿por qué se introduce como uno de los supuestos de Gauss-Markov? Recuerde que en el capítulo 3 se vio que sin el supuesto de homocedasticidad los estimadores de las *varianzas*, $\text{Var}(\hat{\beta}_j)$, son sesgados. Como los errores estándar de MCO se basan directamente en estas varianzas, dejan de ser válidos para la construcción de intervalos de confianza y de estadísticos t. En presencia de heterocedasticidad los estadísticos t usuales de MCO no tienen distribuciones t, y el problema no se resuelve empleando muestras grandes. Esto se verá explícitamente en la siguiente sección para el caso de la regresión simple, donde se obtiene la varianza de los estimadores de pendiente de MCO bajo heterocedasticidad y se propone un estimador válido en presencia de heterocedasticidad. De manera similar, los estadísticos F dejan de seguir una distribución F, y el estadístico *ML* deja de tener una distribución ji-cuadrada asintótica. En resumen, los estadísticos empleados en las pruebas de hipótesis bajo los supuestos de Gauss-Markov ya no son válidos en presencia de heterocedasticidad.

También se sabe que el teorema de Gauss-Markov, que dice que MCO da el mejor estimador lineal de insesgamiento, depende de manera crucial del supuesto de homocedasticidad. Si $\text{Var}(u|\mathbf{x})$ no es constante, MCO ya no es MELI. Además, MCO ya no es asintóticamente eficiente en la clase de los estimadores descritos en el teorema 5.3. Como se verá en la sección 8.4, en presencia de heterocedasticidad es posible hallar estimadores que sean más eficientes que MCO (aunque es necesario conocer la forma de la heterocedasticidad). Con tamaños de muestra relativamente grandes, puede no ser tan importante obtener un estimador eficiente. En la siguiente sección se muestra cómo pueden modificarse los estadísticos de prueba usuales de MCO de manera que sean válidos, por lo menos asintóticamente.

8.2 Inferencia robusta a la heterocedasticidad en la estimación por MCO

Dado que las pruebas de hipótesis son un componente tan importante de cualquier análisis econométrico y que la inferencia usual de MCO general es incorrecta en presencia de heterocedasticidad, debe decidirse si es necesario abandonar por completo MCO. Por fortuna, MCO sigue siendo útil. En las dos décadas pasadas, los econometristas han encontrado cómo ajustar los errores estándar y los estadísticos t, F y *ML* de manera que sean válidos en presencia de **heterocedasticidad de la forma desconocida**. Esto es muy conveniente porque significa que se pueden reportar nuevos estadísticos que funcionen sin importar el tipo de heterocedasticidad presente en la población. Los métodos que se presentan en esta sección se conocen como procedimientos *robustos a la heterocedasticidad*, debido a que son válidos —por lo menos para muestras grandes—, ya sea que los errores tengan o no varianza constante, sin necesidad de saber cuál es el caso.

Se comienza por bosquejar cómo pueden estimarse las varianzas, $\text{Var}(\hat{\beta}_j)$, en presencia de heterocedasticidad. Una deducción detallada de la teoría queda fuera del alcance de este libro, pero el empleo de los métodos robustos a la heterocedasticidad es actualmente muy sencillo debido a que muchos paquetes para estadística y econometría cuentan con una opción para calcular estos estadísticos.

Primero, considere el modelo con una sola variable independiente, en el que para mayor claridad se incluye un subíndice i:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Se supondrá que se satisfacen los cuatro primeros supuestos de Gauss-Markov. Si los errores contienen heterocedasticidad, entonces

$$\mathrm{Var}(u_i|x_i) = \sigma_i^2,$$

donde el subíndice i de σ^2 indica que la varianza del error depende del valor particular de la x_i.

Se pueden expresar los estimadores de MCO como

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Bajo los supuestos RLM.1 a RLM.4 (es decir, sin el supuesto de homocedasticidad) y condicionando sobre los valores x_i de la muestra, pueden utilizarse los mismos argumentos que en el capítulo 2 para mostrar que

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\mathrm{STC}_x^2}, \tag{8.2}$$

donde $\mathrm{STC}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ es la suma total de cuadrados de las x_i. Cuando $\sigma_i^2 = \sigma^2$ para toda i, esta fórmula se reduce a la forma usual, σ^2/STC_x. La ecuación (8.2) muestra explícitamente que, en el caso de la regresión simple, la fórmula para la varianza obtenida bajo homocedasticidad, en presencia de heterocedasticidad ya no es válida.

Puesto que el error estándar de $\hat{\beta}_1$ se basa directamente en la estimación de $\mathrm{Var}(\hat{\beta}_1)$, se necesita una manera de estimar la ecuación (8.2) en presencia de heterocedasticidad. White (1980) mostró cómo puede hacerse esto. Sean $\hat{u}_i$ los residuales de MCO de la regresión inicial de y sobre x. Entonces, un estimador válido de $\mathrm{Var}(\hat{\beta}_1)$, para heterocedasticidad de *cualquier* forma (incluyendo homocedasticidad), es

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{u}_i^2}{\mathrm{STC}_x^2}, \tag{8.3}$$

el cual puede calcularse fácilmente a partir de los datos de la regresión de MCO.

¿En qué sentido es (8.3) un estimador válido de $\mathrm{Var}(\hat{\beta}_1)$? Esto es bastante sutil. Brevemente, se puede demostrar que la ecuación (8.3) multiplicada por el tamaño n de la muestra, converge en probabilidad a $\mathrm{E}[(x_i - \mu_x)^2u_i^2]/(\sigma_x^2)^2$, que es el límite de probabilidad de n multiplicado por (8.2). En principio, esto es lo que se necesita para justificar el uso de los errores estándar para construir intervalos de confianza y estadísticos t. La ley de los números grandes y el teorema del límite central desempeñan papeles importantes en el establecimiento de estas convergencias. Para más detalles se puede consultar el artículo original de White, aunque es muy técnico. Vea también Wooldridge (2002, capítulo 4).

Una fórmula similar funciona con el modelo general de regresión múltiple

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

Puede demostrarse que un estimador válido de $\text{Var}(\hat{\beta}_j)$, bajo los supuestos RLM.1 a RLM.4, es

$$\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{\text{STC}_j^2}, \tag{8.4}$$

donde $\hat{r}_{ij}$ denota el i-ésimo residual de regresar x_j sobre el resto de las variables independientes, y SRC_j es la suma de residuales cuadrados de esta regresión (vea la sección 3.2 para la representación de descuento de efectos parciales de los estimadores de MCO). La raíz cuadrada de la cantidad en (8.4) se conoce como **error estándar** de $\hat{\beta}_j$ **robusto a la heterocedasticidad**. En econometría, estos errores estándar robustos en general se atribuyen a White (1980). Trabajos anteriores en estadística, en especial los de Eicker (1967) y Huber (1967), apuntaban hacia la posibilidad de obtener tales errores estándar robustos. En la práctica, a estos errores se les suele llamar *errores estándar de White, de Huber* o *de Eicker* (o alguna combinación de estos nombres unidos por un guión). Aquí se les llamará *errores estándar robustos a la heterocedasticidad*, o simplemente *errores estándar robustos* cuando el contexto lo permita.

A veces, como una corrección de los grados de libertad, antes de obtener la raíz cuadrada, (8.4) se multiplica por $n/(n - k - 1)$. El razonamiento para hacer este ajuste es que, si los residuales cuadrados de MCO $\hat{u}_i^2$ fueran iguales para todas las observaciones i —la forma más fuerte posible de homocedasticidad en una muestra— se obtendrían los errores estándar usuales de MCO. En MacKinnon y White (1985) se estudian otras modificaciones de (8.4). Puesto que todas las formas sólo tienen justificación asintótica y son asintóticamente equivalentes, ninguna forma es uniformemente preferida a las otras. En general, se utiliza cualquier forma que sea calculada por el paquete de regresión que se emplee.

Una vez que se han obtenido los errores estándar robustos a la heterocedasticidad, es fácil construir un **estadístico *t* robusto a la heterocedasticidad**. Recuerde que la forma general del estadístico t es

$$t = \frac{\textit{estimación} - \textit{valor hipotético}}{\textit{error estándar}}. \tag{8.5}$$

Como aún se siguen utilizando las estimaciones de MCO y como los valores hipotéticos se han elegido de antemano, la única diferencia entre el estadístico t usual de MCO y el estadístico robusto a la heterocedasticidad es la manera en que se calcula el error estándar.

Ejemplo 8.1

[Ecuación para el logaritmo del salario con errores estándar robustos a la heterocedasticidad]

Se estimará el modelo del ejemplo 7.6 y se darán los errores estándar robustos a la heterocedasticidad junto con los errores estándar usuales de MCO. Algunas de las estimaciones se dan con más dígitos de modo que los errores estándar usuales puedan compararse con los errores estándar robustos a la heterocedasticidad:

$$\begin{aligned}
\widehat{\log(wage)} = {} & \underset{\substack{(.100)\\[.109]}}{.321} + \underset{\substack{(.055)\\[.057]}}{.213}\, marrmale - \underset{\substack{(.058)\\[.058]}}{.198}\, marrfem - \underset{\substack{(.056)\\[.057]}}{.110}\, singfem \\
& + \underset{\substack{(.0067)\\[.0074]}}{.0789}\, educ + \underset{\substack{(.0055)\\[.0051]}}{.0268}\, exper - \underset{\substack{(.00011)\\[.00011]}}{.00054}\, exper^2 \\
& + \underset{\substack{(.0068)\\[.0069]}}{.0291}\, tenure - \underset{\substack{(.00023)\\[.00024]}}{.00053}\, tenure^2 \\
n = {} & 526,\ R^2 = .461.
\end{aligned} \tag{8.6}$$

Los errores estándar usuales de MCO aparecen entre paréntesis, ( ), debajo de la estimación correspondiente de MCO, y los errores estándar robustos a la heterocedasticidad aparecen entre corchetes, [ ]. Los números entre corchetes son lo único nuevo, puesto que la ecuación se sigue estimando por MCO.

En la ecuación (8.6) se pueden observar varias cosas. Primero, en esta aplicación particular, las variables que son estadísticamente significativas empleando el estadístico t usual lo siguen siendo usando el estadístico t robusto a la heterocedasticidad. Esto se debe a que los dos conjuntos de errores estándar no son muy diferentes. (Los valores-p correspondientes diferirán ligeramente porque los estadísticos t robustos no son idénticos a los estadísticos t usuales no robustos). La mayor diferencia relativa en los errores estándar está en el coeficiente de *educ:* el error estándar usual es .0067, y el error estándar robusto es .0074. No obstante, el error estándar robusto implica un estadístico t robusto mayor que 10.

La ecuación (8.6) también demuestra que los errores estándar robustos pueden ser mayores o menores que los errores estándar usuales. Por ejemplo, el error estándar robusto de *exper* es .0051, mientras que el error estándar usual es .0055. No se puede saber con antelación cuáles serán mayores. De manera empírica, los errores estándar robustos son a menudo mayores que los errores estándar usuales.

Antes de dejar este ejemplo, debe subrayarse que, en este punto, no se sabe si hay heterocedasticidad en el modelo poblacional subyacente a la ecuación (8.6). Todo lo que se ha hecho es dar, junto con los errores estándar usuales, aquellos que son válidos (asintóticamente) en presencia o no de heterocedasticidad. Se puede ver que, en este ejemplo, al usar los errores estándar robustos ninguna conclusión importante se ha modificado. Esto sucede a menudo en la práctica, pero en otros casos, las diferencias entre los errores estándar usuales y los robustos son mucho más grandes. Como ejemplo en donde hay diferencias sustanciales, vea el ejercicio para computadora C8.2.

Al llegar a este punto, el lector tal vez se pregunte lo siguiente: si los errores estándar robustos a la heterocedasticidad son válidos más a menudo que los errores estándar usuales de MCO, ¿por qué molestarse en calcular los errores estándar usuales? Esta es una pregunta razonable. Una razón por la que se utilizan los errores estándar usuales en el trabajo con cortes transversales es que, si el supuesto de homocedasticidad se satisface y los errores están distribuidos normalmente, los estadísticos t usuales tiene distribuciones t *exactas,* sin importar el tamaño de muestra (vea el capítulo 4). Los errores estándar robustos y los estadísticos t robustos se justifican sólo si el tamaño de muestra se hace grande. Con tamaños de muestra pequeños, el estadístico t robusto puede tener distribuciones que no estén muy próximas a la distribución t y que podrían echar a perder la inferencia.

Cuando se trata de muestras grandes, se justifica que en las aplicaciones con cortes transversales se reporten sólo los errores estándar robustos a la heterocedasticidad, y esto es lo que, en la práctica, se hace cada vez más. Es también usual dar ambos errores estándar, como en la ecuación (8.6), de modo que el lector pueda determinar si algunas conclusiones son sensibles al error estándar empleado.

También pueden obtenerse estadísticos F y ML robustos a una heterocedasticidad de forma desconocida y arbitraria. El **estadístico *F* robusto a la heterocedasticidad** (o una transformación simple del mismo) se conoce también como *estadístico de Wald robusto a la heterocedasticidad*. Un estudio general del estadístico de Wald requiere álgebra de matrices y se bosqueja en el apéndice E; para un estudio más detallado vea Wooldridge (2002, capítulo 4). Sea como sea, el uso de estadísticos robustos a la heterocedasticidad para restricciones de exclusión múltiple es sencillo debido a que, actualmente, muchos paquetes para econometría calculan estos estadísticos de manera rutinaria.

Ejemplo 8.2

[Estadístico *F* robusto a la heterocedasticidad]

Con ayuda de los datos correspondientes al semestre de primavera del archivo GPA3.RAW, se estima la ecuación siguiente:

$$\begin{aligned}\widehat{cumgpa} = {} & \underset{[.22]}{\underset{(.23)}{1.47}} + \underset{[.00019]}{\underset{(.00018)}{.00114}}\, sat - \underset{[.00140]}{\underset{(.00124)}{.00857}}\, hsperc + \underset{[.00073]}{\underset{(.00073)}{.00250}}\, tothrs \\ & + \underset{[.059]}{\underset{(.059)}{.303}}\, female - \underset{[.118]}{\underset{(.147)}{.128}}\, black - \underset{[.110]}{\underset{(.141)}{.059}}\, white \\ & n = 366,\ R^2 = .4006,\ \bar{R}^2 = .3905.\end{aligned} \tag{8.7}$$

Una vez más, las diferencias entre los errores estándar usuales y los errores estándar robustos a la heterocedasticidad no son muy grandes, y el uso de los estadísticos t robustos no modifica la significancia estadística de ninguna de las variables independientes. Las pruebas para significancia conjunta tampoco se ven muy afectadas. Suponga que se desea probar la hipótesis nula de que, una vez controlados los demás factores, no existen diferencias en *cumgpa* de acuerdo con la raza. Esto se indica como H_0: $\beta_{black} = 0$, $\beta_{white} = 0$. El estadístico F usual se obtiene fácilmente, una vez que se tiene la R-cuadrada del modelo restringido; ésta resulta ser .3983. El estadístico F es entonces $[(.4006 - .3983)/(1-.4006)](359/2) \approx .69$. En presencia de heterocedasticidad, esta versión de la prueba no es válida. Para la versión robusta a la heterocedasticidad no existe una forma sencilla, pero se puede calcular usando ciertos paquetes estadísticos. El valor del estadístico F robusto a la heterocedasticidad resulta ser .75, que sólo difiere ligeramente del de la versión no robusta. El valor-p para la prueba robusta es .474, que no está próximo a los niveles de significancia estándar. Con ninguna de las pruebas puede rechazarse la hipótesis nula.

Cálculo de pruebas *ML* robustas a la heterocedasticidad

No todos los paquetes para regresión calculan estadísticos F robustos a la heterocedasticidad. Por tanto, a veces es conveniente contar con una manera de obtener una prueba para restricciones de exclusión múltiple que sea robusta a la heterocedasticidad y que no requiera un

Pregunta 8.1

Evalúe la afirmación siguiente: Los errores estándar robustos a la heterocedasticidad son siempre mayores que los errores estándar usuales.

determinado tipo de software para econometría. Resulta que un **estadístico *ML* robusto a la heterocedasticidad** puede obtenerse con facilidad usando casi cualquier paquete para regresión.

Para ilustrar el cálculo del estadístico *ML* robusto, considere el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u,$$

y suponga que se desea probar H_0: $\beta_4 = 0$, $\beta_5 = 0$. Para obtener el estadístico *ML*, primero se estimará el modelo restringido (es decir, el modelo sin x_4 ni x_5) para obtener los residuales, $\tilde{u}$. Después, se regresa $\tilde{u}$ sobre todas las variables independientes y $ML = n \cdot R^2_{\tilde{u}}$, donde $R^2_{\tilde{u}}$ es la R-cuadrada usual de esta regresión.

Obtener una versión robusta a la heterocedasticidad requiere aún más. Una manera de calcular el estadístico requiere sólo regresiones de MCO. Se necesitan los residuales, por ejemplo, $\tilde{r}_1$, de la regresión de x_4 sobre x_1, x_2, x_3. También, se necesitan los residuales, por ejemplo, $\tilde{r}_2$, de la regresión de x_5 sobre x_1, x_2, x_3. Así, cada una de las variables independientes excluidas bajo la hipótesis nula se regresa sobre todas las variables independientes incluidas. Estos residuales se conservan. El paso final parece algo extraño, pero es, después de todo, sólo un recurso de cálculo. Se corre la regresión de

$$1 \text{ sobre } \tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \tag{8.8}$$

sin intercepto. Sí, efectivamente, se define una variable dependiente igual al valor uno para todas las observaciones. Esta variable se regresa sobre los productos $\tilde{r}_1\tilde{u}$ y $\tilde{r}_2\tilde{u}$. El estadístico *ML* robusto resulta ser $n - \text{SRC}_1$, donde SRC_1 es la suma usual de residuales cuadrados de la regresión (8.8).

La razón de que esto funcione es algo técnica. Básicamente, esto es hacer con la prueba del *ML* lo que los errores estándar robustos hacen con la prueba de *t*. [Vea Wooldridge (1991b) o Davidson y MacKinnon (1993) para un estudio más detallado].

Ahora se resume el cálculo del estadístico *ML* robusto a la heterocedasticidad en el caso general.

UN ESTADÍSTICO *ML* ROBUSTO A LA HETEROCEDASTICIDAD:

1. Obtener los residuales $\tilde{u}$ del modelo restringido.
2. Regresar cada una de las variables independientes excluidas bajo la hipótesis nula sobre todas las variables independientes incluidas; si hay q variables excluidas, esto conduce a q conjuntos de residuales $(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_q)$.
3. Obtener los productos entre cada $\tilde{r}_j$ y $\tilde{u}$ (para todas las observaciones).
4. Correr la regresión de 1 sobre $\tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \ldots, \tilde{r}_q\tilde{u}$, sin intercepto. El estadístico *ML* robusto a la heterocedasticidad es $n - \text{SRC}_1$, donde SRC_1 es precisamente la suma usual de residuales cuadrados de esta regresión final. Bajo H_0, la distribución de *ML* es aproximadamente χ^2_q.

Una vez obtenido el estadístico *ML* robusto, la regla de rechazo y el cálculo de los valores-*p* son los mismos que para el estadístico *ML* usual de la sección 5.2.

Ejemplo 8.3

[Estadístico *ML* robusto a la heterocedasticidad]

Se utilizan los datos del archivo CRIME1.RAW para probar si la longitud media de sentencias cumplidas en el pasado afecta la cantidad de arrestos en el año actual (1986). El modelo estimado es

$$
\begin{aligned}
\widehat{narr86} = {} & .567 - .136\ pcnv + .0178\ avgsen - .00052\ avgsen^2 \\
& (.036)\quad (.040) \qquad (.0097) \qquad\quad (.00030) \\
& [.040]\quad [.034] \qquad [.0101] \qquad\quad [.00021] \\
& - .0394\ ptime86 - .0505\ qemp86 - .00148\ inc86 \\
& \quad (.0087) \qquad\quad (.0144) \qquad\quad (.00034) \\
& \quad [.0062] \qquad\quad [.0142] \qquad\quad [.00023] \\
& + .325\ black + .193\ hispan \\
& \quad (.045) \qquad (.040) \\
& \quad [.058] \qquad [.040] \\
n = {} & 2{,}725,\ R^2 = .0728.
\end{aligned}
\tag{8.9}
$$

En este ejemplo, hay más diferencias sustanciales entre algunos de los errores estándar usuales y los errores estándar robustos. Por ejemplo, el estadístico usual t para $avgsen^2$ es aproximadamente -1.73, mientras que el estadístico t robusto es aproximadamente -2.48. Así, $avgsen^2$ es más significativa empleando el error estándar robusto.

El efecto de *avgsen* sobre *narr86* es un poco más difícil de reconciliar. Dado que es una relación cuadrática, es posible imaginar dónde *avgsen* tiene efecto positivo sobre *narr86* y dónde su efecto llega a ser negativo. El punto de inflexión es $.0178/[2(.00052)] \approx 17.12$; recuérdese que esto se mide en meses. Literalmente, esto significa que *narr86* está relacionada positivamente con *avgsen* cuando *avgsen* es menor a 17 meses; entonces *avgsen* tiene el efecto disuasivo previsto después de 17 meses.

Para ver si la longitud media de la sentencia tiene un efecto estadísticamente significativo sobre *narr86*, se deben probar las hipótesis conjuntas H_0: $\beta_{avgsen} = 0$, $\beta_{avgsen^2} = 0$. Usando el estadístico *ML* usual (vea la sección 5.2), se obtiene $ML = 3.54$; en una distribución ji-cuadrada con dos *gl*, esto da un valor-$p = .170$. Así H_0 no se rechaza ni incluso al nivel de 15%. El estadístico *ML* robusto a la heterocedasticidad es $ML = 4.00$ (redondeado a dos cifras decimales), con un valor-$p = .135$. Ésta aún no es una evidencia muy fuerte contra H_0; *avgsen* no parece tener un efecto fuerte sobre *narr86*. [Incidentalmente, cuando *avgsen* aparece sola en (8.9), es decir, sin el término cuadrático, su estadístico t usual es .658, y su estadístico t robusto es .592].

8.3 Pruebas para heterocedasticidad

Los errores estándar robustos a la heterocedasticidad proporcionan un método simple para calcular estadísticos t que asintóticamente tienen una distribución t, ya sea que exista o no heterocedasticidad. Se vio que también existen estadísticos F y *ML* robustos a la heterocedasticidad. Para realizar estas pruebas no se requiere saber si existe o no heterocedasticidad. A pesar de esto hay algunas razones para emplear pruebas sencillas que puedan detectar su presencia. Primero, como se mencionó en la sección anterior, los estadísticos t usuales tienen, bajo los supuestos del modelo lineal clásico, distribuciones t exactas. Por esta razón, muchos economistas siguen prefiriendo ver reportados los errores estándar de MCO y los estadísticos de prueba usuales, a menos que haya evidencia de heterocedasticidad. En segundo lugar, en presencia de heterocedasticidad, los estimadores de MCO ya no son los mejores estimadores lineales insesgados. Como se verá en la sección 8.4, cuando se conoce la forma de la heterocedasticidad, es posible obtener un estimador mejor que el de MCO.

A lo largo de los años se han sugerido muchas pruebas para heterocedasticidad. Algunas de ellas, aunque pueden detectar heterocedasticidad, no prueban directamente el supuesto de que la varianza del error no depende de las variables independientes. Aquí nos restringiremos a pruebas más modernas, que detectan el tipo de heterocedasticidad que invalida los estadísticos usuales de MCO. Esto también tiene la ventaja de poner todas las pruebas en un mismo marco.

Como de costumbre, se parte del modelo lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u, \tag{8.10}$$

donde los supuestos RLM.l a RLM.4 se mantienen en esta sección. En particular se supone que $E(u|x_1, x_2, \dots, x_k) = 0$, de manera que MCO sea insesgado y consistente.

Como hipótesis nula se toma que el supuesto RLM.5 sea verdadero:

$$H_0\text{: } \mathrm{Var}(u|x_1, x_2, \dots, x_k) = \sigma^2. \tag{8.11}$$

Es decir, se supone que el supuesto ideal de homocedasticidad se satisface, y se requiere de los datos para afirmar otra cosa. Si (8.11) no se puede rechazar a un nivel de significancia suficientemente pequeño, en general se concluye que la heterocedasticidad no es problema. Sin embargo, recuérdese que H_0; nunca se acepta, simplemente no puede rechazarse.

Como se supone que u tiene una esperanza condicional cero, $\mathrm{Var}(u|\mathbf{x}) = E(u^2|\mathbf{x})$, y de esta manera la hipótesis nula de homocedasticidad es equivalente a

$$H_0\text{: } E(u^2|x_1, x_2, \dots, x_k) = E(u^2) = \sigma^2.$$

Esto muestra que, con objeto de probar la violación del supuesto de homocedasticidad, se prueba si u^2 está relacionada (en valor esperado) con una o más de las variables explicativas. Si H_0 es falsa, el valor esperado de u^2, dadas las variables independientes, puede ser prácticamente cualquier función de x_j. Un método sencillo es suponer una función lineal:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + v, \tag{8.12}$$

donde v es un término de error con media cero dadas las x_j. Ponga mucha atención a la variable dependiente de esta ecuación: es el *cuadrado* del error en la ecuación de regresión original, (8.10). La hipótesis nula de homocedasticidad es

$$H_0\text{: } \delta_1 = \delta_2 = \dots = \delta_k = 0. \tag{8.13}$$

Bajo la hipótesis nula, es a menudo razonable suponer que el error en (8.12), v, es independiente de $x_1, x_2, \dots, x_k$. Entonces, se sabe, de acuerdo con la sección 5.2, que el estadístico F o el estadístico ML para significancia general de las variables independientes para explicar u^2 puede utilizarse para probar (8.13). Los dos estadísticos tendrán justificación asintótica, aun cuando u^2 no pueda estar distribuida normalmente. (Por ejemplo, si u está distribuida normalmente, entonces u^2/σ^2 está distribuida como χ^2_1.) Si se pudiera observar la u^2 en la muestra, este estadístico podría calcularse con facilidad ejecutando la regresión por MCO de u^2 sobre $x_1, x_2, \dots, x_k$, usando todas las n observaciones.

Como se ha insistido, en el modelo poblacional nunca se conocen los errores reales, pero se tienen estimaciones de ellos: el residual de MCO, $\hat{u}_i$, es una estimación del error u_i para la observación i. Así, puede estimarse la ecuación

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + error \tag{8.14}$$

y calcular los estadísticos F o ML para la significancia conjunta de $x_1, \dots, x_k$. Resulta que, cuando se trata de muestras grandes, emplear los residuales de MCO en lugar de los errores no afecta la distribución de los estadísticos F o ML aunque demostrar esto es bastante complicado.

Tanto el estadístico F como el estadístico ML dependen de la R-cuadrada de la regresión (8.14); llámesele $R^2_{\hat{u}^2}$ para distinguirla de la R-cuadrada de la ecuación (8.10) estimada. Entonces, el estadístico F es

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n - k - 1)}, \tag{8.15}$$

donde k es el número de regresores en (8.14); éste es igual al número de variables independientes en (8.10). Es raramente necesario calcular (8.15) a mano, ya que la mayoría de los paquetes para regresión calcula automáticamente el estadístico F para la significancia general en una regresión. Este estadístico F tiene (aproximadamente) una distribución $F_{k,n-k-1}$ bajo la hipótesis nula de homocedasticidad.

El estadístico ML para heterocedasticidad es precisamente el tamaño de la muestra multiplicado por la R-cuadrada de (8.14):

$$LM = n \cdot R^2_{\hat{u}^2}. \tag{8.16}$$

Bajo la hipótesis nula, ML se distribuye asintóticamente como χ^2_k. Esto también es muy fácil de obtener después de ejecutar la regresión (8.14).

A la versión ML de la prueba se le llama **prueba Breusch-Pagan para heterocedasticidad (prueba BP)**. Breusch y Pagan (1979) sugirieron una forma diferente de la prueba que supone que los errores están distribuidos normalmente. Koenker (1981) sugirió la forma del estadístico ML en (8.16), que es generalmente preferida debido a su mayor aplicabilidad.

A continuación se resumen los pasos para probar heterocedasticidad usando la prueba de BP:

LA PRUEBA DE BREUSCH-PAGAN PARA HETEROCEDASTICIDAD:

1. Estimar el modelo (8.10) por MCO, como de costumbre. Obtener los residuales cuadrados de MCO, $\hat{u}^2$ (uno para cada observación).
2. Ejecutar la regresión en (8.14). Conservar la R-cuadrada de esta regresión, $R^2_{\hat{u}^2}$.
3. Formar, ya sea el estadístico F o el estadístico ML y calcular el valor-p (usando la distribución de $F_{k,n-k-1}$ en el primer caso y la distribución χ^2_k en el último caso). Si el valor-p es suficientemente pequeño, es decir, menor que el nivel de significancia elegido, se rechaza la hipótesis nula de homocedasticidad.

Si la prueba de BP da como resultado un valor-p suficientemente pequeño, deberá tomarse alguna medida correctiva. Una posibilidad es utilizar los errores estándar y los estadísticos de prueba robustos a la heterocedasticidad vistos en la sección anterior. Otra posibilidad se discute en la sección 8.4.

Ejemplo 8.4

[Heterocedasticidad en las ecuaciones sobre el precio de la vivienda]

Para probar heterocedasticidad en una sencilla ecuación sobre el precio de la vivienda se utilizan los datos del archivo HPRICEl.RAW. La ecuación estimada usando los niveles de todas las variables es

$$\widehat{price} = -21.77 + .00207\ lotsize + .123\ sqrft + 13.85\ bdrms$$
$$(29.48)\quad (.00064) \qquad (.013) \qquad (9.01) \tag{8.17}$$
$$n = 88,\ R^2 = .672.$$

Esta ecuación no nos dice *nada* acerca de si el error en el modelo poblacional es heterocedástico. Es necesario regresar los residuales cuadrados de MCO sobre las variables independientes. La R-cuadrada de la regresión de $\hat{u}^2$ sobre *lotsize*, *sqrft* y *bdrms* es $R^2_{\hat{u}^2} = .1601$. Siendo $n = 88$ y $k = 3$, el estadístico F que se obtiene para la significancia de las variables independientes es $F = [.1601/(1 - .1601)](84/3) \approx 5.34$. El valor-$p$ correspondiente es .002, que es una fuerte evidencia contra la hipótesis nula. El estadístico ML es $88(.1601) \approx 14.09$; esto da un valor-$p \approx .0028$ (empleando la distribución χ^2_3), lo que lleva esencialmente a la misma conclusión que el estadístico F. Esto significa que los errores estándar usuales dados en (8.17) no son confiables.

En el capítulo 6 se dijo que una ventaja de usar la forma funcional logarítmica de la variable dependiente es que suele reducirse la heterocedasticidad. En la aplicación presente, se empleará *price*, *lotsize* y *sqrft* en forma logarítmica, de modo que las elasticidades del *price*, respecto a *lotsize* y *sqrft*, sean constantes. La ecuación estimada es

$$\widehat{\log(price)} = -1.30 + .168\ \log(lotsize) + .700\ \log(sqrft) + .037\ bdrms$$
$$(.65)\quad (.038) \qquad (.093) \qquad (.028) \tag{8.18}$$
$$n = 88,\ R^2 = .643.$$

Regresando los residuales cuadrados de MCO de esta regresión sobre log(*lotsize*), log(*sqrft*) y *bdrms* se obtiene $R^2_{\hat{u}^2} = .0480$. Así, $F = 1.41$ (valor-$p = .245$), y $ML = 4.22$ (valor-$p = .239$). Por tanto, en el modelo con las formas funcionales logarítmicas no puede rechazarse la hipótesis nula de homocedasticidad. La menor ocurrencia de heterocedasticidad cuando la variable dependiente está en forma logarítmica se ha observado en muchas aplicaciones empíricas.

Si se sospecha que la heterocedasticidad sólo depende de ciertas variables independientes, se puede modificar fácilmente la prueba de Breusch-Pagan: simplemente se regresa $\hat{u}^2$ sobre cualquier variable independiente que se elija y se realiza la prueba F o la prueba ML apropiada. Recuérdese que los grados de libertad apropiados dependen del número de variables independientes en la regresión con $\hat{u}^2$ como variable dependiente; el número de las variables independientes que aparecen en la ecuación (8.10) es irrelevante.

Si los residuales cuadrados se regresan sólo sobre una variable independiente, la prueba para la heterocedasticidad es precisamente el estadístico t usual de esa variable. Un estadístico t significativo sugiere que la heterocedasticidad es un problema.

Pregunta 8.2

Considere la ecuación del salario (7.11), donde usted piensa que la varianza condicional de log(*wage*) no depende de *educ*, *exper* o *tenure*. Pero le interesa ver si la varianza de log(*wage*) es diferente entre los cuatro grupos demográficos: hombres casados, mujeres casadas, hombres solteros y mujeres solteras. ¿Qué regresión ejecutaría usted para hacer una prueba de heterocedasticidad? ¿Cuáles son los grados de libertad en la prueba F?

Prueba de White para heterocedasticidad

En el capítulo 5 se mostró que los errores estándar usuales de MCO y los estadísticos de prueba usuales son asintóticamente válidos, siempre que se satisfagan todos los supuestos de Gauss-Markov. Resulta que el supuesto de homocedasticidad, $\mathrm{Var}(u_1|x_1, \ldots, x_k) = \sigma^2$, puede ser sustituido por el supuesto más débil de que el error cuadrado, u^2, *no está correlacionado* con ninguna de las variables independientes (x_j), ni con los cuadrados de las variables independientes (x_j^2), ni con ninguno de los productos cruzados ($x_j x_h$ para $j \neq h$). Esta observación motivó a White (1980) a proponer una prueba para heterocedasticidad que agrega a la ecuación (8.14) los cuadrados y los productos cruzados de todas las variables independientes. La prueba tiene por objeto probar las formas de heterocedasticidad que invalidan los errores estándar usuales de MCO y los estadísticos de prueba usuales.

Si el modelo contiene $k = 3$ variables independientes, la prueba de White se basa en la estimación de

$$\begin{aligned} \hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error. \end{aligned} \tag{8.19}$$

Comparada con la prueba de Breusch-Pagan, esta ecuación tiene seis regresores más. La **prueba de White para heterocedasticidad** es el estadístico *ML* para probar que todas las δ_j en la ecuación (8.19) son cero, a excepción del intercepto. Así, en este caso, se prueban nueve restricciones. Para esta hipótesis también puede utilizarse una prueba *F*; ambas pruebas tienen justificación asintótica.

Aunque sólo hay tres variables independientes en el modelo original, la ecuación (8.19) tiene nueve variables independientes. Si en el modelo original hay seis variables independientes, la regresión de White implicará, en general, 27 regresores (a menos que algunos sean redundantes). Esta abundancia de regresores es una debilidad de la forma pura de la prueba de White: utiliza muchos grados de libertad para modelos que sólo tienen un número moderado de variables independientes.

Es posible obtener una prueba que sea más fácil de realizar que la de White y que preserve mejor los grados de libertad. Para crearla, recuerde que la diferencia entre las pruebas de White y de Breusch-Pagan es que la primera incluye los cuadrados y los productos cruzados de las variables independientes. Tanto el espíritu de la prueba de White como los grados de libertad pueden preservarse empleando los valores ajustados de MCO en una prueba para heterocedasticidad. Recuerde que, para cada observación i, los valores ajustados están definidos por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}.$$

Éstas sólo son funciones lineales de las variables independientes. Si los valores ajustados se elevan al cuadrado, se obtiene una función particular de todos los cuadrados y productos cruzados de las variables independientes. Esto sugiere probar la heterocedasticidad estimando la ecuación

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error, \tag{8.20}$$

donde $\hat{y}$ representa los valores ajustados. En esta ecuación es importante no confundir $\hat{y}$ y y. Se utilizan los valores ajustados porque éstos son funciones de las variables independientes (y de los parámetros estimados); usando y en (8.20) no se obtiene una prueba válida para heterocedasticidad.

Para la hipótesis nula H_0: $\delta_1 = 0$, $\delta_2 = 0$ en la ecuación (8.20) pueden utilizarse los estadísticos *F* o *ML*. Esto da como resultado dos restricciones al probar la hipótesis nula de homocedasticidad, sin importar la cantidad de variables independientes en el modelo original. Conservar los grados de libertad de esta manera suele ser una buena idea y también hace que la prueba sea más fácil de realizar.

Puesto que $\hat{y}$ es una estimación del valor esperado de y, dadas las x_j, usar (8.20) para probar heterocedasticidad es útil en caso de que se crea que la varianza cambia con el nivel del valor esperado, $E(y|\mathbf{x})$. La prueba en (8.20) puede verse como un caso especial de la prueba de White, ya que se puede demostrar que la ecuación (8.20) impone restricciones a los parámetros de la ecuación (8.19).

UN CASO ESPECIAL DE LA PRUEBA DE WHITE PARA HETEROCEDASTICIDAD:

1. Estimar el modelo (8.10) mediante MCO, como de costumbre. Obtener los residuales $\hat{u}$ de MCO y los valores ajustados $\hat{y}$. Calcular los cuadrados de los residuales de MCO, $\hat{u}^2$ y los cuadrados de los valores ajustados $\hat{y}^2$.
2. Ejecutar la regresión de la ecuación (8.20). Conservar la R-cuadrada de esta regresión, $R^2_{\hat{u}^2}$.
3. Formar, ya sea el estadístico F o el estadístico ML, y calcular el valor-p (empleando la distribución $F_{2,n-3}$ en el primer caso y la distribución χ^2_2 en el último).

Ejemplo 8.5

[Forma especial de la prueba de White en la ecuación del logaritmo del precio de la vivienda]

Se aplica el caso especial de la prueba de White a la ecuación (8.18), donde se utiliza la forma *ML* del estadístico. Es importante recordar que la distribución ji-cuadrada siempre tiene dos *gl*. De la regresión de $\hat{u}^2$ sobre $\widehat{lprice}$, $(\widehat{lprice})^2$, donde $\widehat{lprice}$ denota los valores ajustados obtenidos con (8.18), se obtiene $R^2_{\hat{u}^2} = .0392$; así, $ML = 88(.0392) \approx 3.45$, y el valor-$p = .178$. Esta es una evidencia de heterocedasticidad más fuerte que la proporcionada por la prueba de Breusch-Pagan, pero aun así no se puede rechazar la homocedasticidad ni al nivel de 15 por ciento.

Antes de dejar esta sección, hay que analizar una advertencia importante. El rechazo obtenido en una de las pruebas para heterocedasticidad se ha interpretado como evidencia de heterocedasticidad. Esto es correcto siempre que se conserven los supuestos RLM.1 a RLM.4. Pero si se infringe RLM.4 —en particular, si la forma funcional $E(y|\mathbf{x})$ está mal especificada— entonces una prueba para heterocedasticidad puede rechazar H_0, aun cuando $\text{Var}(y|\mathbf{x})$ sea constante. Por ejemplo, si en un modelo de regresión se omiten uno o más términos cuadráticos o se utiliza el modelo lineal cuando debe utilizarse el logarítmico, una prueba para la heterocedasticidad puede ser significativa. Esto ha conducido a algunos economistas a considerar las pruebas para heterocedasticidad como pruebas generales para especificación errónea. Sin embargo, hay pruebas mejores y más directas para formas funcionales mal especificadas y algunas de ellas se verán en la sección 9.1. Es mejor utilizar pruebas explícitas para formas funcionales primero, puesto que la especificación errónea de las formas funcionales es más importante que la heterocedasticidad. Una vez satisfechos con la forma funcional, se puede probar la heterocedasticidad.

8.4 Estimación por mínimos cuadrados ponderados

Si empleando una de las pruebas vistas en la sección 8.3 se detecta heterocedasticidad, se sabe, de acuerdo con lo visto en la sección 8.2, que una posible solución es utilizar estadísticos robustos a la heterocedasticidad después de la estimación por MCO. Antes del desarrollo de los estadísticos robustos a la heterocedasticidad, la solución, cuando se encontraba heterocedasticidad, era especificar su forma y utilizar un método de *mínimos cuadrados ponderados*, el cual se desarrollará en esta sección. Como se verá, si se ha especificado correctamente la forma de la varianza (como función de las variables explicativas), entonces el método de mínimos cuadrados ponderados (MCP) es más eficiente que el de MCO y MCP conduce a nuevos estadísticos t y F que tienen distribuciones t y F. Se verán también las implicaciones de usar una forma incorrecta de la varianza en el procedimiento de MCP.

Heterocedasticidad conocida, salvo una constante multiplicativa

Denótense con **x** todas las variables explicativas de la ecuación (8.10) y suponga que

$$\mathrm{Var}(u|\mathbf{x}) = \sigma^2 h(\mathbf{x}), \tag{8.21}$$

donde $h(\mathbf{x})$ es alguna función de las variables explicativas que determina la heterocedasticidad. Puesto que las varianzas deben ser positivas, $h(\mathbf{x}) > 0$ para todos los posibles valores de las variables independientes. Por ahora, se supondrá que la función $h(\mathbf{x})$ es conocida. El parámetro poblacional σ^2 no se conoce, pero puede ser estimado a partir de una muestra de datos.

Dada una muestra aleatoria de la población, puede escribirse $\sigma_i^2 = \mathrm{Var}(u_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$, donde una vez más se utiliza la notación $\mathbf{x}_i$ para denotar todas las variables independientes de la observación i, y las h_i cambian con cada observación porque las variables independientes cambian de una observación a otra. Por ejemplo, considere la sencilla función del ahorro (*sav*):

$$sav_i = \beta_0 + \beta_1 inc_i + u_i \tag{8.22}$$

$$\mathrm{Var}(u_i|inc_i) = \sigma^2 inc_i. \tag{8.23}$$

Aquí, $h(x) = h(inc) = inc$: la varianza del error es proporcional al nivel de ingreso (*inc*). Esto significa que, a medida que aumenta el ingreso, aumenta la variabilidad en el ahorro. (Si $\beta_1 > 0$, el valor esperado para el ahorro también aumenta con el ingreso). Como *inc* es siempre positivo, se garantiza que en la ecuación (8.23) la varianza sea siempre positiva. La desviación estándar de u_i, condicional sobre inc_i, es $\sigma\sqrt{inc_i}$.

¿Cómo puede utilizarse la información de la ecuación (8.21) para estimar los β_j? En esencia se toma la ecuación original,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i, \tag{8.24}$$

que contiene errores heterocedásticos, y se transforma en una ecuación que tenga errores homocedásticos (y satisfaga los demás supuestos de Gauss-Markov). Como h_i es una función de $\mathbf{x}_i$, $u_i/\sqrt{h_i}$ tiene valor esperado cero condicional sobre $\mathbf{x}_i$. Además, como $\mathrm{Var}(u_i|\mathbf{x}_i) = \mathrm{E}(u_i^2|\mathbf{x}_i) = \sigma^2 h_i$, la varianza de $u_i/\sqrt{h_i}$ (condicional sobre $\mathbf{x}_i$) es σ^2:

$$\mathrm{E}\left((u_i/\sqrt{h_i})^2\right) = \mathrm{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

donde para simplificar se ha suprimido el condicionamiento **sobre** $\mathbf{x}_i$. La ecuación (8.24) puede dividirse entre $\sqrt{h_i}$ para obtener

$$\begin{aligned} y_i/\sqrt{h_i} &= \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \ldots \\ &\quad + \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i}) \end{aligned} \tag{8.25}$$

o

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + u_i^*, \tag{8.26}$$

donde el $x_{i0}^* = 1/\sqrt{h_i}$ y las demás variables con asterisco denotan las variables originales correspondientes divididas entre $\sqrt{h_i}$.

La ecuación (8.26) parece algo peculiar, pero lo que es importante recordar es que se dedujo con objeto de que se pudieran obtener estimadores de β_j que tuvieran mejores propiedades de eficiencia que los de MCO. El intercepto β_0 de la ecuación original (8.24) aparece ahora multiplicando la variable $x_{i0}^* = 1/\sqrt{h_i}$. Cada parámetro de pendiente en β_j multiplica una nueva variable que rara vez tiene una interpretación útil. Esto no deberá causar ningún problema si se recuerda que, para interpretar los parámetros y el modelo, siempre se vuelve a la ecuación original (8.24).

En el ejemplo anterior sobre el ahorro, la ecuación transformada es

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*,$$

donde se utiliza el hecho de que $inc_i/\sqrt{inc_i} = \sqrt{inc_i}$. De cualquier manera, β_1 es la propensión marginal a ahorrar del ingreso, una interpretación que se obtiene de la ecuación (8.22).

La ecuación (8.26) es lineal en sus parámetros (así que satisface RLM.1), y el supuesto de muestreo aleatorio no ha cambiado. Además, u_i^* tiene media cero y varianza constante (σ^2), condicional sobre $\mathbf{x}_i^*$. Esto significa que si la ecuación original satisface los primeros cuatro supuestos de Gauss-Markov, entonces la ecuación transformada (8.26) satisface los cinco supuestos de Gauss-Markov. Asimismo, si u_i tiene una distribución normal, entonces u_i^* tiene una distribución normal con varianza σ^2. Por tanto, la ecuación transformada satisface los supuestos del modelo lineal clásico (RLM.1 a RLM.6) si el modelo original, a excepción del supuesto de homocedasticidad, los satisface.

Puesto que se sabe que MCO tiene atractivas características (es MELI, por ejemplo) bajo los supuestos de Gauss-Markov, el análisis del párrafo anterior sugiere estimar los parámetros de la ecuación (8.26) mediante mínimos cuadrados ordinarios. Estos estimadores, $\beta_0^*, \beta_1^*, \ldots, \beta_k^*$, serán diferentes de los estimadores de MCO de la ecuación original. Los β_j^* son ejemplos de **estimadores de mínimos cuadrados generalizados (MCG)**. En este caso, los estimadores de MCG se utilizan para considerar la heterocedasticidad de los errores. En el capítulo 12 se encontrarán otros estimadores de MCG.

Dado que la ecuación (8.26) satisface todos los supuestos ideales, los errores estándar, los estadísticos t y los estadísticos F pueden obtenerse de regresiones en las que se usen las variables transformadas. La suma de los residuales cuadrados de (8.26) dividida entre los grados de libertad es un estimador insesgado de σ^2. Además, los estimadores de MCG, debido a que son los mejores estimadores lineales insesgados de las β_j, son necesariamente más eficientes que los estimadores $\hat{\beta}_j$ de MCO, obtenidos de la ecuación no transformada. En esencia, después de que se han transformado las variables, simplemente se utiliza el análisis estándar de MCO. Pero debe recordarse que hay que interpretar las estimaciones a la luz de la ecuación original.

La R-cuadrada que se obtiene de estimar (8.26), aunque es útil para calcular estadísticos F, no es especialmente informativa como medida de bondad de ajuste: indica cuánta de la variación en y^* es explicada por las x_j^*, y esto rara vez tiene mucho significado.

Los estimadores de MCG para la corrección de la heterocedasticidad se denominan **estimadores de mínimos cuadrados ponderados (MCP)**. Este nombre proviene del hecho de que los β_j^* minimizan la suma *ponderada* de los residuales cuadrados, donde cada residual cuadrado es ponderado por $1/h_i$. La idea es dar menos peso a las observaciones que tienen una varianza del error mayor; MCO da a cada observación el mismo peso dado que MCO es mejor cuando la varianza del error es idéntica en todas las particiones de la población. Matemáticamente, los estimadores de MCP son los valores de b_j que hacen

$$\sum_{i=1}^{n} (y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \ldots - b_kx_{ik})^2/h_i \qquad \boxed{8.27}$$

tan pequeña como sea posible. Introduciendo la raíz cuadrada de $1/h_i$ dentro del residual cuadrado se muestra que la suma ponderada de los residuales cuadrados es idéntica a la suma de los residuales cuadrados en las variables transformadas:

$$\sum_{i=1}^{n} (y_i^* - b_0x_{i0}^* - b_1x_{i1}^* - b_2x_{i2}^* - \ldots - b_kx_{ik}^*)^2.$$

Como los MCO minimizan la suma de los residuales cuadrados (sin importar las definiciones de la variable dependiente e independiente), se sigue que los estimadores de MCP que minimizan (8.27) son simplemente los estimadores de MCO de (8.26). Observe cuidadosamente que los residuales cuadrados en (8.27) están ponderados por $1/h_i$, mientras que las variables transformadas en (8.26) están ponderadas por $1/\sqrt{h_i}$.

Para todo conjunto de ponderadores positivos puede definirse un estimador de mínimos cuadrados ponderados. MCO es el caso especial que da el mismo peso a todas las observaciones. El procedimiento eficiente, MCG, pondera cada residual cuadrado con el *inverso* de la varianza condicional de u_i dado $\mathbf{x}_i$.

Obtener las variables transformadas de la ecuación (8.25) para realizar manualmente mínimos cuadrados ponderados puede ser tedioso y la probabilidad de cometer errores no es trivial. Por fortuna, la mayoría de los paquetes modernos para regresión cuenta con opciones para calcular mínimos cuadrados ponderados. Por lo general, junto con las variables dependientes e independientes del modelo original, simplemente se especifica la función de peso, $1/h_i$, que aparece en (8.27). Es decir, se especifican ponderadores proporcionales al inverso de la varianza. Además de que se cometen menos errores, esto obliga a interpretar las estimaciones de mínimos cuadrados ponderados en el modelo original. En realidad, la ecuación estimada puede expresarse de la manera usual. Los coeficientes estimados y los errores estándar serán diferentes de los de MCO, pero la manera de *interpretar* esos coeficientes estimados, errores estándar y estadísticos de prueba es la misma.

Ejemplo 8.6

[Ecuación de riqueza financiera]

Ahora se estiman las ecuaciones que explican la riqueza financiera total neta (*nettfa*, medida en miles de dólares) en términos del ingreso (*inc*, también medido en miles de dólares) y de algunas otras variables, como edad, género, y un indicador de si la persona es elegible para un plan de pensiones 401(k). Se utilizan los datos de las personas solteras ($fsize = 1$) del archivo 401KSUBS.RAW. En el ejercicio para computadora C6.12, se encontró que una función cuadrática específica de edad, a saber $(age - 25)^2$, ajusta los datos tan bien como una ecuación cuadrática sin restricción. Pero además, la forma restringida da una interpretación más sencilla debido a que la edad mínima en la muestra es 25 años: *nettfa* es una función creciente de la edad después de $age = 25$ años.

Los resultados se presentan en la tabla 8.1. Como se sospecha que haya heterocedasticidad, se dan los errores estándar de MCO robustos a la heterocedasticidad. Los coeficientes estimados de mínimos cuadrados ponderados, y sus errores estándar, se obtienen bajo el supuesto $\text{Var}(u|inc) = \sigma^2 inc$.

TABLA 8.1

Variable dependiente: *nettfa*

Variables independientes	(1) MCO	(2) MCP	(3) MCO	(4) MCP
inc	.821 (.104)	.787 (.063)	.771 (.100)	.740 (.064)
$(age - 25)^2$	—	—	.0251 (.0043)	.0175 (.0019)
male	—	—	2.48 (2.06)	1.84 (1.56)
e401k	—	—	6.89 (2.29)	5.19 (1.70)
intercepto	−10.57 (2.53)	−9.58 (1.65)	−20.98 (3.50)	−16.70 (1.96)
Observaciones	2,017	2,017	2,017	2,017
R-cuadrada	.0827	.0709	.1279	.1115

Sin controlar otros factores, se estima que un dólar más de ingreso incremente *nettfa* en alrededor de 82¢ cuando se utiliza MCO; el coeficiente estimado de MCP es más pequeño, aproximadamente 79¢. La diferencia no es grande; ciertamente no se esperaba que fueran idénticos. El coeficiente de MCP tiene un error estándar menor que el de MCO, casi 40% menor, siempre que se suponga que el modelo $\text{Var}(nettfa|inc) = \sigma^2 inc$ es correcto.

La adición de los otros controles redujo un poco el coeficiente de *inc*, siendo el coeficiente estimado de MCO todavía mayor que el de MCP. Una vez más, la estimación de MCP de β_{inc} es más precisa. La edad (*age*) tiene un efecto creciente empezando en $age = 25$, mostrando el coeficiente estimado de MCO un efecto mayor. La estimación de MCP de β_{age} es más exacta en este caso. El género masculino (*male*) no tiene un efecto estadísticamente significativo sobre *nettfa*, pero el ser elegible para un plan 401(k) sí: la estimación de MCO es que, manteniendo constantes el ingreso, la edad y el género, aquellos que son elegibles tienen activos financieros totales netos cerca de 6,890 dólares más altos. La estimación de MCP es sustancialmente inferior a la estimación de MCO y sugiere una mala especificación de la forma funcional en la ecuación. (Una posibilidad es interactuar *e401k* e *inc*; vea el ejercicio para computadora C8.11.)

Si se usa la *R*-cuadrada dada en la tabla 8.1, usando MCP, el estadístico *F* para la significancia conjunta de $(age - 25)^2$, *male* y *e401k* es aproximadamente de 30.8. Con 2 y 2,012 grados de libertad, el valor-*p* es cero a más de 15 cifras decimales; por supuesto, esto no es inesperado dado el estadístico *t* tan grande para las variables *age* y *e401k*.

Pregunta 8.3

Utilizando los residuales de MCO obtenidos de la regresión de MCO reportada en la columna (1) de la tabla 8.1, la regresión de $\hat{u}^2$ sobre *inc* da un estadístico *t* de 2.96. ¿Parece que debería preocupar la heterocedasticidad en la ecuación de la riqueza financiera?

Suponer que la varianza del error en la ecuación de la riqueza financiera tenga una varianza proporcional al ingreso es esencialmente arbitrario. En realidad, en la mayoría de los casos, la elección de los ponderadores en MCP tiene cierto grado de arbitrariedad. Sin embargo, existe un caso en el que los ponderadores necesarios para MCP surgen de manera natural del modelo econométrico subyacente. Esto sucede cuando, en lugar de usar datos de individuos, sólo se tienen promedios de datos de un grupo o de una región geográfica. Por ejemplo, suponga que se desea determinar la cantidad con la que un trabajador contribuye a su plan de pensión 401(k) en función de la generosidad del plan. Sea i una determinada empresa y sea e un empleado de esa empresa. Un modelo sencillo es

$$contrib_{i,e} = \beta_0 + \beta_1 earns_{i,e} + \beta_2 age_{i,e} + \beta_3 mrate_i + u_{i,e}, \tag{8.28}$$

donde $contrib_{i,e}$ es la contribución anual del empleado e que trabaja en la empresa i, $earns_{i,e}$ es el ingreso anual de esta persona y $age_{i,e}$ es su edad. La variable $mrate_i$ es la cantidad que la empresa aporta a la cuenta de un empleado por cada dólar con que éste contribuye.

Si (8.28) satisface los supuestos de Gauss-Markov, entonces puede estimarse esta ecuación, dada una muestra de individuos pertenecientes a varios empleadores. Sin embargo, supóngase que sólo se cuenta con valores *promedio* de contribuciones, ingresos y edades por empleador. Es decir, no se cuenta con datos individuales. Así, sea $\overline{contrib}_i$ la contribución promedio de las personas de la empresa i, y de manera similar $\overline{earns}_i$ y $\overline{age}_i$. Sea m_i la cantidad de empleados en la empresa i; se supone que ésta es una cantidad conocida. Entonces, si se promedia la ecuación (8.28) sobre todos los empleados de la empresa i, se obtiene la ecuación a nivel empresa

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \beta_2 \overline{age}_i + \beta_3 mrate_i + \bar{u}_i, \tag{8.29}$$

donde $\bar{u}_i = m_i^{-1} \sum_{e=1}^{m_i} u_{i,e}$ es el error promedio de todos los empleados en el empresa i. Si en la muestra se tienen n empresas, entonces (8.29) es un modelo de regresión lineal múltiple estándar que puede ser estimado por MCO. Los estimadores son insesgados si el modelo original (8.28) satisface los supuestos de Gauss-Markov y si los errores de los individuos $u_{i,e}$ son independientes del tamaño de la empresa, m_i [porque entonces el valor esperado de $\bar{u}_i$, dadas las variables explicativas en (8.29), es cero].

Si la ecuación a nivel individuo (8.28) satisface el supuesto de homocedasticidad, y dentro de la empresa i los errores no están correlacionados entre los empleados, entonces se puede demostrar que la ecuación a nivel de la empresa (8.29) tiene un tipo especial de heterocedasticidad. Específicamente, si $\mathrm{Var}(u_{i,e}) = \sigma^2$ para toda i y e, and $\mathrm{Cov}(u_{i,e},u_{i,g}) = 0$ para todo par de empleados $e \neq g$ de la empresa i, entonces $\mathrm{Var}(\bar{u}_i) = \sigma^2/m_i$; esta es la fórmula usual de la varianza de un promedio de variables aleatorias no correlacionadas con varianza común. En otras palabras, la varianza del término del error $\bar{u}_i$ disminuye con el tamaño de la empresa. En este caso, $h_i = 1/m_i$, y de esta manera el procedimiento más eficiente es mínimos cuadrados ponderados, donde los ponderadores son iguales al número de empleados en la empresa ($1/h_i = m_i$). Esto asegura que a las empresas más grandes se les dé más peso. Esto proporciona una manera eficiente de estimar los parámetros del modelo a nivel individuo cuando sólo se tienen promedios a nivel empresa.

Una ponderación similar surge cuando se utilizan datos *per cápita* a nivel de una ciudad, un condado, un estado o un país. Si la ecuación a nivel individual satisface los supuestos de Gauss-Markov, entonces el error en la ecuación *per cápita* tiene una varianza proporcional a uno entre el tamaño de la población. Por tanto, el método de mínimos cuadrados ponderados con ponderadores iguales a la población es adecuado. Suponga, por ejemplo, que a nivel de una ciudad se tienen datos sobre consumo *per cápita* de cerveza (*beerpc*, en onzas), porcentaje de personas mayores de 21 años en la población (*perc21*), nivel medio de educación de los adultos (*avgeduc*), ingreso promedio (*incpc*) y precio de la cerveza en la ciudad (*price*). Entonces, el modelo a nivel ciudad

$$beerpc = \beta_0 + \beta_1 perc21 + \beta_2 avgeduc + \beta_3 incpc + \beta_4 price + u$$

puede ser estimado mediante mínimos cuadrados ponderados, empleando como ponderadores la población de las ciudades.

La ventaja de ponderar empleando el tamaño de la empresa, la población, la ciudad, etc., radica en que la ecuación individual subyacente sea homocedástica. Si existe heterocedasticidad a nivel individual, entonces la ponderación adecuada depende de la forma de la heterocedasticidad. Además, si hay correlación entre los errores dentro de un grupo (por ejemplo, de una empresa), entonces $\text{Var}(\bar{u}_i) \neq \sigma^2/m_i$; vea el problema 8.7. La incertidumbre acerca de la forma de $\text{Var}(\bar{u}_i)$ en ecuaciones como la (8.29) es la razón por la que cada vez más investigadores utilizan simplemente MCO y calculan errores estándar robustos y estadísticos de prueba robustos al estimar modelos usando datos *per cápita*. Una alternativa es ponderar con el tamaño del grupo, pero reportar los estadísticos robustos a la heterocedasticidad en la estimación de MCP. Esto garantiza que, mientras que la estimación es eficiente si el modelo individual satisface los supuestos de Gauss-Markov, la heterocedasticidad a nivel individual o la correlación dentro del grupo sean explicadas mediante inferencia robusta.

La función de heterocedasticidad debe ser estimada: MCG factibles

En la subsección anterior, se vieron algunos ejemplos en los que se conoce la heterocedasticidad, excepto por una constante multiplicativa. En la mayoría de los casos, la forma exacta de heterocedasticidad no es obvia. Es decir, es difícil encontrar la función $h(\mathbf{x}_i)$ de la sección anterior. Sin embargo, en muchos casos puede modelarse la función h y utilizar los datos para estimar los parámetros desconocidos del modelo. Esto da como resultado una estimación para cada h_i, que se denota como $\hat{h}_i$. Usando $\hat{h}_i$ en lugar de h_i en la transformación de MCG, se obtiene un estimador llamado **estimador de MCG factibles (MCGF)**. A los MCG factibles también se les suele llamar *MCG estimados*, o MCGE.

Hay muchas maneras de modelar la heterocedasticidad, pero aquí se estudiará un método particular, bastante flexible. Suponga que

$$\text{Var}(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k), \tag{8.30}$$

donde $x_1, x_2, \ldots, x_k$ son las variables independientes que aparecen en el modelo de regresión [vea la ecuación (8.1)], y las δ_j son parámetros desconocidos. Pueden encontrarse otras funciones de las x_j pero aquí se enfocará la atención principalmente en (8.30). En la notación de la subsección anterior, $h(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$.

El lector se preguntará por qué se ha utilizado la función exponencial en (8.30). Después de todo, al *probar* heterocedasticidad usando la prueba Breusch-Pagan, se supuso que la heterocedasticidad era una función lineal de las x_j. Las alternativas lineales como en (8.12) son adecuadas cuando se prueba heterocedasticidad, pero pueden ser problemáticas cuando se trata de corregir la heterocedasticidad empleando mínimos cuadrados ponderados. La razón de este problema ya se ha encontrado antes: los modelos lineales no aseguran que los valores predichos sean positivos, y para emplear el método de MCP las varianzas estimadas deben ser positivas.

Si se conocieran los parámetros δ_j simplemente se aplicaría MCP, como en la subsección anterior. Esto no es muy realista. Es mejor utilizar los datos para estimar estos parámetros, y después utilizar estas estimaciones para construir los ponderadores. ¿Cómo pueden estimarse las δ_j? Esencialmente, esta ecuación se transforma en una forma lineal que, con una leve modificación, puede ser estimada por MCO.

Bajo el supuesto (8.30), se puede escribir

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)v,$$

donde v tiene media igual a la unidad, condicional sobre $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. Si se supone que v es realmente independiente de $\mathbf{x}$, puede escribirse

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e, \tag{8.31}$$

donde e tiene media cero y es independiente de $\mathbf{x}$; en esta ecuación el intercepto es diferente a δ_0, pero esto no es importante para emplear MCP. La variable dependiente es el logaritmo del error cuadrado. Puesto que (8.31) satisface los supuestos de Gauss-Markov, pueden obtenerse estimadores insesgados de δ_j usando MCO.

Como de costumbre, debe sustituirse la u no observada por los residuales de MCO. Por tanto, la regresión que se ejecuta es

$$\log(\hat{u}^2) \text{ sobre } x_1, x_2, \ldots, x_k. \tag{8.32}$$

En realidad, lo que se necesita de esta regresión son los valores ajustados; llámeseles $\hat{g}_i$. Entonces, las h_i estimadas son

$$\hat{h}_i = \exp(\hat{g}_i). \tag{8.33}$$

Ahora se utiliza en la ecuación (8.27) MCP con $1/\hat{h}_i$ en lugar de $1/h_i$. A continuación se resumen estos pasos.

PROCEDIMIENTO CON MCG FACTIBLES PARA CORREGIR LA HETEROCEDASTICIDAD:

1. Ejecutar la regresión de y sobre $x_1, x_2, \ldots, x_k$ y obtener los residuales, $\hat{u}$.
2. Obtener $\log(\hat{u}^2)$ elevando primero al cuadrado los residuales de MCO y en seguida tomando el logaritmo natural.
3. Ejecutar la regresión de la ecuación (8.32) y obtener los valores ajustados, $\hat{g}$.
4. Exponenciar los valores ajustados de (8.32): $\hat{h} = \exp(\hat{g})$.
5. Estimar la ecuación

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

por MCP, usando como ponderadores $1/\hat{h}$. En otras palabras, en la ecuación (8.27) se sustituye h_i por $\hat{h}_i$. Recuerde que el residual *cuadrado* de la observación i se pondera con $1/\hat{h}_i$. Si en lugar de esto, se transforman primero todas las variables y se ejecuta MCO, cada variable se multiplica por $1/\sqrt{\hat{h}_i}$, incluido el intercepto.

Si en el procedimiento de MCP pudiera utilizarse h_i en lugar de $\hat{h}_i$ se sabe que nuestros estimadores serían insesgados; de hecho, serían los mejores estimadores lineales insesgados, suponiendo que se haya modelado correctamente la heterocedasticidad. Tener que estimar h_i con los mismos datos significa que el estimador de MCGF deja de ser insesgado (así que tampoco puede ser MELI). Sin embargo, el estimador de MCGF es consistente y *asintóticamente* más eficiente que MCO. Esto es difícil de demostrar debido a la estimación de los parámetros de la varianza. Pero ignorando esto —como se hará— la demostración es similar a demostrar que MCO es eficiente en la clase de estimadores del teorema 5.3. De cualquier manera, para tamaños de muestra grandes, MCGF es una alternativa interesante a MCO cuando hay evidencia de heterocedasticidad que infla los errores estándar de los estimadores de MCO.

Debe recordarse que los estimadores de MCGF son estimadores de los parámetros del modelo poblacional usual

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

Así como las estimaciones de MCO miden el impacto marginal de cada x_j sobre y, así también lo hacen las estimaciones de MCGF. Las estimaciones de MCGF se utilizan en lugar de las de MCO debido a que los estimadores de MCGF son más eficientes y tienen estadísticos de prueba con las distribuciones t y F usuales, por lo menos en muestras grandes. Si se tiene alguna duda acerca de la varianza especificada en la ecuación (8.30), pueden utilizarse errores estándar robustos a la heterocedasticidad y estadísticos de prueba en la ecuación transformada.

Otra alternativa útil para estimar h_i es sustituir las variables independientes de la regresión (8.32) por los valores ajustados de MCO y sus cuadrados. En otras palabras, obtener las $\hat{g}_i$ como valores ajustados de la regresión de

$$\log(\hat{u}^2) \text{ sobre } \hat{y}, \hat{y}^2 \qquad \boxed{8.34}$$

y después obtener $\hat{h}_i$ exactamente como en la ecuación (8.33). Esto sólo modifica el paso (3) del procedimiento anterior.

Si se utiliza la regresión (8.32) para estimar la función de la varianza, el lector puede preguntarse si se puede probar la heterocedasticidad usando simplemente esta misma regresión (se puede utilizar una prueba de F o ML). En realidad, Park (1966) sugirió esto. Por desgracia, en comparación con las pruebas vistas en la sección 8.3, la prueba de Park tiene algunos problemas. Primero, la hipótesis nula debe ser algo más fuerte que la homocedasticidad: efectivamente, u y $\mathbf{x}$ deben ser independientes. Esto no se requiere en las pruebas de Breusch-Pagan o de White. En segundo lugar, usando los residuales de MCO $\hat{u}$ en lugar de u en (8.32) puede hacer que el estadístico F se desvíe de la distribución F incluso en tamaños de muestra grandes. Esto no es problema en otras de las pruebas que se han visto. Por estas razones, la prueba de Park no se recomienda para probar heterocedasticidad. La regresión (8.32) funciona bien para mínimos cuadrados ponderados debido a que sólo se necesitan estimadores consistentes de δ_j, y la regresión (8.32) ciertamente los proporciona.

Ejemplo 8.7

[Demanda de cigarros]

Se emplean los datos del archivo SMOKE.RAW para estimar una función de demanda para el consumo diario de cigarros. Puesto que la mayoría de la gente no fuma, la variable dependiente, *cigs*, es cero en la mayoría de las observaciones. Un modelo lineal no es lo ideal debido a que puede dar lugar a valores de predicción negativos. Sin embargo, se puede conocer algo acerca de los determinantes del tabaquismo usando un modelo lineal.

La ecuación estimada por mínimos cuadrados ordinarios, con los errores estándar usuales de MCO entre paréntesis, es

$$\begin{aligned}\widehat{cigs} &= \underset{(24.08)}{-3.64} + \underset{(.728)}{.880}\log(income) - \underset{(5.773)}{.751}\log(cigpric) \\ &\quad - \underset{(.167)}{.501}\, educ + \underset{(.160)}{.771}\, age - \underset{(.0017)}{.0090}\, age^2 - \underset{(1.11)}{2.83}\, restaurn \\ n &= 807,\ R^2 = .0526,\end{aligned} \tag{8.35}$$

donde

$cigs$ = número de cigarros fumados por día.
$income$ = ingreso anual.
$cigpric$ = precio del paquete de cigarros (en centavos de dólar).
$educ$ = años de escolaridad.
age = edad medida en años.
$restaurn$ = un indicador binario igual a uno si la persona reside en un estado en el que fumar en los restaurantes está prohibido.

Puesto que también se van a emplear mínimos cuadrados ponderados, no se dan los errores estándar de MCO robustos a la heterocedasticidad. (De manera incidental, 13 de los 807 valores ajustados son menores a cero; esto es menos de 2% de la muestra y no es causa de preocupación).

En (8.35) no son estadísticamente significativos el ingreso ni el precio de los cigarros, y sus efectos prácticos no son grandes. Por ejemplo, si aumenta el ingreso 10%, se predice que $cigs$ aumente $(.880/100)(10) = .088$, o menos de un décimo de un cigarro por día. La magnitud del efecto del precio es similar.

Cada año de escolaridad reduce la cantidad promedio de cigarros fumados por día en medio cigarro, y este efecto es estadísticamente significativo. El tabaquismo también se relaciona con la edad, en una forma cuadrática. El tabaquismo aumenta con la edad hasta $age = .771/[2(.009)] \approx 42.83$, y después decrece con la edad. Ambos términos de la función cuadrática son estadísticamente significativos. La presencia de una restricción para fumar en restaurantes disminuye el tabaquismo en casi tres cigarros por día, en promedio.

¿Los errores subyacentes de la ecuación (8.35) contienen heterocedasticidad? La regresión de Breusch-Pagan de los residuales cuadrados de MCO sobre las variables independientes en (8.35) [vea la ecuación (8.14)] produce $R^2_{\hat{u}^2} = .040$. Esta R-cuadrada pequeña parecería indicar que no hay heterocedasticidad, pero no hay que olvidar calcular ya sea el estadístico F o el estadístico ML. Si el tamaño de la muestra es grande, una $R^2_{\hat{u}^2}$ aparentemente pequeña puede dar lugar a un muy fuerte rechazo de la homocedasticidad. El estadístico ML es $ML = 807(.040) = 32.28$, y este es el resultado de una variable aleatoria χ^2_6. El valor-p es menor a .000015, lo que es una fuerte evidencia de heterocedasticidad.

Por tanto, se estima la ecuación usando el procedimiento de MCG factibles basado en la ecuación (8.32). Las estimaciones de mínimos cuadrados ponderados son

$$\begin{aligned}\widehat{cigs} &= \underset{(17.80)}{5.64} + \underset{(.44)}{1.30}\log(income) - \underset{(4.46)}{2.94}\log(cigpric) \\ &\quad - \underset{(.120)}{.463}\, educ + \underset{(.097)}{.482}\, age - \underset{(.0009)}{.0056}\, age^2 - \underset{(.80)}{3.46}\, restaurn \\ n &= 807,\ R^2 = .1134.\end{aligned} \tag{8.36}$$

El efecto del ingreso es ahora estadísticamente significativo y mayor en magnitud. El efecto del precio es también notablemente mayor, pero sigue siendo estadísticamente insignificante. [Una razón de esto es que *cigpric* varía sólo entre los estados de la muestra y, por tanto, hay mucha menos variación en log(*cigpric*) que en log(*income*), *educ* y *age*.]

Las estimaciones de las demás variables tienen, naturalmente, algunos cambios, pero la historia básica sigue siendo la misma. El tabaquismo se relaciona negativamente con la escolaridad, tiene una relación cuadrática con *age* (edad), y es afectado negativamente por restricciones para fumar en los restaurantes.

Después de una estimación por MCP se debe ser un poco cuidadoso al calcular los estadísticos F para probar hipótesis múltiples. (Esto, ya sea que se utilice la fórmula del estadístico F con suma de residuales cuadrados o con la R-cuadrada). Es importante que se utilicen los mismos ponderadores para estimar el modelo no restringido y el restringido. Primero debe estimarse el modelo no restringido mediante MCO. Una vez obtenidos los ponderadores, éstos pueden utilizarse para estimar también el modelo restringido. El estadístico F puede calcularse como de costumbre. Por fortuna, muchos paquetes para regresión tienen un comando sencillo para probar restricciones conjuntas después de la estimación por MCP, así es que no es necesario que uno mismo realice la regresión restringida.

El ejemplo 8.7 hace alusión a un problema que suele presentarse en las aplicaciones de mínimos cuadrados ponderados: las estimaciones por MCO y por MCP pueden ser sustancialmente diferentes. Este no es un gran problema en el caso de la ecuación de la demanda de cigarros debido a que todos los coeficientes conservan el mismo signo, y a que los mayores cambios se presentan en las variables que eran estadísticamente insignificantes cuando la ecuación se estimó por MCO. Las estimaciones de MCO y de MCP diferirán siempre debido al error de muestreo. El problema es si su diferencia es suficiente para modificar conclusiones importantes.

Pregunta 8.4

Sean $\hat{u}_i$ los residuales de MCP de (8.36), los residuales no ponderados, y sean $\widehat{cigs}_i$ los valores ajustados. (Éstos se obtienen usando las mismas fórmulas que en MCO; difieren debido a las diferentes estimaciones de las β_j.) Una manera de determinar si la heterocedasticidad ha sido eliminada es utilizar los $\hat{u}_i^2/\hat{h}_i = (\hat{u}_i/\sqrt{\hat{h}_i})^2$ en una prueba para heterocedasticidad. [Si $h_i = \text{Var}(u_i|\mathbf{x}_i)$, entonces los residuales transformados deberán tener poca evidencia de heterocedasticidad.] Hay muchas posibilidades, pero una —basada en la prueba de White en la ecuación transformada— es regresar $\hat{u}_i^2/\hat{h}_i$ sobre $\widehat{cigs}_i/\sqrt{\hat{h}_i}$ y $\widehat{cigs}_i^2/\hat{h}_i$ (incluyendo un intercepto). El estadístico F conjunto usando SMOKE.RAW es 11.15. ¿Parece que nuestra corrección de heterocedasticidad haya eliminado en realidad la heterocedasticidad?

Si MCO y MCP producen estimaciones estadísticamente significativas que difieren en el signo —por ejemplo, que el precio elasticidad por MCO es positivo y significativo, mientras que el precio elasticidad por MCP sea negativo y significativo— o la diferencia en las magnitudes de las estimaciones es grande en sentido práctico, hay que desconfiar. En general, esto indica que uno de los *otros* supuestos de Gauss-Markov es falso, en particular el de media condicional cero del error (RLM.4). Si $E(y|\mathbf{x}) \neq \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, entonces MCO y MCP tienen valores esperados y límites de probabilidad diferentes. Para que MCP dé estimadores consistentes de las β_j, no basta que u no esté correlacionada con cada una de las x_j; se necesita el supuesto más fuerte RLM.4 en el modelo lineal RLM.1. Por tanto, una diferencia significativa entre MCO y MCP puede indicar una mala especificación de la forma funcional en $E(y|\mathbf{x})$. La *prueba de Hausman* [Hausman (1978)] puede utilizarse para comparar de manera formal las estimaciones de MCO y de MCP para ver si difieren más de lo que el error de muestreo sugiere que deban hacerlo, pero esta prueba está más allá del alcance de este libro. En muchos casos, un vistazo informal de las estimaciones es suficiente para detectar un problema.

¿Qué pasa si la función de heterocedasticidad supuesta es incorrecta?

Como acaba de observarse, si MCO y MCP producen estimaciones muy diferentes, es probable que la media condicional $E(y|\mathbf{x})$ esté mal especificada. ¿Cuáles son las propiedades de MCP si la función de la varianza que se utiliza está mal especificada en el sentido de que $Var(y|\mathbf{x}) \neq \sigma^2 h(\mathbf{x})$ para la función $h(\mathbf{x})$ elegida? Lo más importante es si la mala especificación de $h(\mathbf{x})$ causa sesgo o inconsistencia en el estimador de MCP. Por fortuna, la respuesta es no, por lo menos bajo RLM.4. Recuerde que si $E(u|\mathbf{x}) = 0$, entonces cualquier función de $\mathbf{x}$ no está correlacionada con u y, por tanto, el error ponderado, $u/\sqrt{h(\mathbf{x})}$, no está correlacionado con los regresores ponderados, $x_j/\sqrt{h(\mathbf{x})}$, para *cualquier* función $h(\mathbf{x})$ que sea siempre positiva. Esta es la razón por la cual, como se acaba de ver, grandes diferencias entre los estimadores de MCO y de MCP pueden considerarse como indicativas de una forma funcional mal especificada. Si se estiman parámetros en la función, por ejemplo $h(\mathbf{x}, \hat{\delta})$, entonces no se puede afirmar que MCP sea insesgado, pero por lo general será consistente (ya sea que la función de la varianza esté o no bien especificada).

Si MCP es consistente, por lo menos bajo RLM.1 a RLM.4, ¿cuáles son las consecuencias de usar MCP con una función de la varianza mal especificada?; hay dos. La primera, que es muy importante, es que los errores estándar y los estadísticos de prueba usuales de MCP, calculados bajo el supuesto de que $Var(y|\mathbf{x}) = \sigma^2 h(\mathbf{x})$, ya no son válidos, incluso en muestras grandes. Por ejemplo, en las estimaciones y errores estándar de MCP en la columna (4) de la tabla 8.1 se supone que $Var(nettfa|inc, age, male, e401k) = Var(nettfa|inc) = \sigma^2 inc$; de manera que no se está suponiendo que la varianza sólo depende del ingreso, sino también que es una función lineal del mismo. Si este supuesto es falso, los errores estándar (y cualquier estadístico que se obtenga usándolos) no son válidos. Por fortuna, hay una solución fácil: así como para los coeficientes estimados de MCO es posible obtener errores estándar que sean robustos a una heterocedasticidad arbitraria, también se pueden obtener errores estándar para MCP que permiten que la función de la varianza esté arbitrariamente mal especificada. Es fácil ver por qué funciona esto. Escríbase la ecuación transformada como

$$y_i/\sqrt{h_i} = \beta_0\left(1/\sqrt{h_i}\right) + \beta_1\left(x_{i1}/\sqrt{h_i}\right) + \ldots + \beta_k\left(x_{ik}/\sqrt{h_i}\right) + u_i/\sqrt{h_i}.$$

Ahora, si $Var(u_i|\mathbf{x}_i) \neq \sigma^2 h_i$, entonces el error ponderado $u_i/\sqrt{h_i}$ es heterocedástico. De manera que simplemente pueden aplicarse los errores estándar usuales robustos a la heterocedasticidad después de estimar esta ecuación por MCO —lo cual, recuérdese, es idéntico a MCP.

Para ver cómo funciona esto, en la columna (1) de la tabla 8.2 se reproduce la última columna de la tabla 8.1, y la columna (2) contiene los errores estándar robustos a $Var(u_i|\mathbf{x}_i) \neq \sigma^2 inc_i$.

Los errores estándar en la columna (2) permiten que la función de la varianza esté mal especificada. Se ve que, para las variables del ingreso y de la edad, los errores estándar robustos son algo mayores que los errores estándar usuales de MCP —ciertamente lo suficiente para extender los intervalos de confianza—. Por otro lado, los errores estándar robustos de *male* (hombre) y de *e401k* son en realidad menores que los que suponen una función correcta de la varianza. Se vio que esto también podía ocurrir con errores estándar de MCO robustos a la heterocedasticidad.

Aun si se utilizan formas flexibles de funciones de la varianza, tales como la función en (8.30), no hay ninguna garantía de que se tenga el modelo correcto. Aunque la heterocedasticidad exponencial es interesante y razonablemente flexible, es sólo un modelo, después de todo. Por tanto, siempre es una buena idea calcular errores estándar y estadísticos de prueba completamente robustos de la estimación de MCP.

TABLA 8.2

Estimación por MCP de la ecuación para *nettfa*

Variables independientes	Con errores estándar no robustos	Con errores estándar robustos
inc	.740 (.064)	.740 (.075)
$(age - 25)^2$	.0175 (.0019)	.0175 (.0026)
male	1.84 (1.56)	1.84 (1.31)
e401k	5.19 (1.70)	5.19 (1.57)
intercepto	−16.70 (1.96)	−16.70 (2.24)
Observaciones	2,017	2,017
R-cuadrada	.1115	.1115

Una crítica moderna a MCP es que si la función de la varianza está mal especificada, no hay garantía de que sea más eficiente que MCO. En efecto, este es el caso: si $\text{Var}(y|\mathbf{x})$ no es constante ni igual a $\sigma^2 h(\mathbf{x})$, donde $h(\mathbf{x})$ es el modelo de heterocedasticidad propuesto, entonces no se puede decir si MCO o MCP es más eficiente en términos de las varianzas (o de las varianzas asintóticas cuando los parámetros de la varianza deban ser estimados). Sin embargo, esta crítica, teóricamente correcta, no ve un punto práctico importante. Este punto es que en casos de una fuerte heterocedasticidad, suele ser mejor usar una forma incorrecta de heterocedasticidad y emplear MCP que ignorar por completo la heterocedasticidad y usar MCO. Modelos como el (8.30) pueden aproximar bien una variedad de funciones de heterocedasticidad y puede que produzcan estimadores con varianzas (asintóticas) menores. Incluso en el ejemplo 8.6, en donde se supuso que la forma de heterocedasticidad tenía la sencilla forma $\text{Var}(nettfa|\mathbf{x}) = \sigma^2 inc$, los errores estándar totalmente robustos de MCP son bastante menores que los errores estándar totalmente robustos de MCO. (La comparación de los errores estándar para los dos estimadores, coloca a ambos en igualdad de condiciones: no se supone homocedasticidad ni que la varianza tenga la forma $\sigma^2 inc$.) Por ejemplo, el error estándar robusto de MCP es 25% menor que el error estándar robusto de MCO. Para $(age - 25)^2$, el error estándar robusto de MCP es alrededor de 40% menor que el error estándar robusto de MCO.

Predicción e intervalos de predicción con heterocedasticidad

Si se parte del modelo lineal estándar bajo los supuestos RLM.1 a RLM.4, pero se considera heterocedasticidad de la forma $\text{Var}(y|\mathbf{x}) = \sigma^2 h(\mathbf{x})$ [vea la ecuación (8.21)], la presencia de heterocedasticidad afecta la predicción puntual de y sólo en tanto que afecta la estimación de las β_j. Por supuesto, es natural emplear MCP en una muestra de tamaño n para obtener las $\hat{\beta}_j$. La predicción de un resultado no observado, y^0, dados valores conocidos de las variables explicativas $\mathbf{x}^0$, tiene la misma forma que en la sección 6.4: $\hat{y}^0 = \hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}}$. Esto es razonable: una vez que se conoce $\text{E}(y|\mathbf{x})$, las predicciones se basarán en esto; la estructura de $\text{Var}(y|\mathbf{x})$ no juega ningún papel directo.

Por otro lado, los *intervalos* de predicción sí dependen directamente del carácter de $\text{Var}(y|\mathbf{x})$. Recuerde que en la sección 6.4 se construyó un intervalo de predicción bajo los supuestos del modelo lineal clásico. Suponga ahora que se satisfacen todos los supuestos del MLC, salvo que (8.21) sustituye el supuesto de homocedasticidad, RLM.5. Se sabe que los estimadores de MCP son MELI y, debido a la normalidad, tienen distribuciones (condicionales) normales. Puede obtenerse $\text{ee}(\hat{y}^0)$ usando el mismo método que en la sección 6.4, salvo que ahora se usa MCP. [Un método sencillo es escribir $y_i = \theta_0 + \beta_1(x_{i1} - x_1^0) + \ldots + \beta_k(x_{ik} - x_k^0) + u_i$, donde las x_j^0 son los valores de las variables explicativas para las que se desea predecir el valor de y. Esta ecuación se puede estimar por MCP y obtener $\hat{y}^0 = \hat{\theta}_0$ y $\text{ee}(\hat{y}^0) = \text{ee}(\hat{\theta}_0)$.] También se necesita estimar la desviación estándar de u^0, la parte no observada de y^0. Pero $\text{Var}(u^0|\mathbf{x} = \mathbf{x}^0) = \sigma^2 h(\mathbf{x}^0)$, y de esta manera $\text{ee}(u^0) = \hat{\sigma}\sqrt{h(\mathbf{x}^0)}$, donde $\hat{\sigma}$ es el error estándar de la regresión de la estimación de MCP. Por tanto, un intervalo de predicción de 95% es

$$\hat{y}^0 \pm t_{.025} \cdot \text{ee}(\hat{e}^0) \tag{8.37}$$

donde $\text{ee}(\hat{e}^0) = \{[\text{ee}(\hat{y}^0)]^2 + \hat{\sigma}^2 h(\mathbf{x}^0)\}^{1/2}$.

Este intervalo es exacto sólo si no se tiene que estimar la función de la varianza. Si se estiman parámetros, como en el modelo (8.30), entonces no se puede obtener un intervalo exacto. De hecho, tomar en consideración el error estimado en las $\hat{\beta}_j$ y en las $\hat{\delta}_j$ (los parámetros de la varianza) se vuelve muy difícil. En la sección 6.4 se vieron dos ejemplos en los que el error de estimación de los parámetros fue sobrepasado por la variación en los inobservables, u^0. Por tanto, puede seguirse usando la ecuación (8.37) sustituyendo $h(\mathbf{x}^0)$ por $\hat{h}(\mathbf{x}^0)$. De hecho, si se va a ignorar por completo el error de estimación de los parámetros, se puede eliminar $\text{ee}(\hat{y}^0)$ de $\text{ee}(\hat{e}^0)$. [Recuerde que, $\text{ee}(\hat{y}^0)$ converge a cero a la velocidad $1/\sqrt{n}$, mientras que $\text{ee}(\hat{u}^0)$ es más o menos constante.]

También se puede obtener una predicción para y en el modelo

$$\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \tag{8.38}$$

donde u es heterocedástica. Suponga que u tiene una distribución condicional normal con una forma específica de heterocedasticidad. Considere la forma exponencial en la ecuación (8.30), pero agregue el supuesto de normalidad:

$$u|x_1, x_2, \ldots, x_k \sim \text{Normal}[0, \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)]. \tag{8.39}$$

De manera abreviada, la función de la varianza se escribe como $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$. Entonces como $\log(y)$ dado $\mathbf{x}$ tiene una distribución normal con media $\beta_0 + \mathbf{x}\boldsymbol{\beta}$ y varianza $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$, se sigue que

$$\text{E}(y|\mathbf{x}) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta} + \sigma^2 \exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})/2). \tag{8.40}$$

Ahora se estiman las β_j y las δ_j usando la estimación por MCP de (8.38). Es decir, después de MCO para obtener los residuales, se corre la regresión de (8.32) para obtener los valores ajustados,

$$\hat{g}_i = \hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + \dots + \hat{\delta}_k x_{ik}, \tag{8.41}$$

y después las $\hat{h}_i$ como en (8.33). Usando estas $\hat{h}_i$, se obtienen las estimaciones de MCP, $\hat{\beta}_j$ y también $\hat{\sigma}^2$. Entonces para cada i, se puede obtener un valor ajustado

$$\hat{y}_i = \exp(\widehat{logy_i} + \hat{\sigma}^2 \hat{h}_i/2). \tag{8.42}$$

Estos valores ajustados se pueden usar para obtener una R-cuadrada, como se describió en la sección 6.4: usando el coeficiente de correlación cuadrado entre y_i y $\hat{y}_i$.

Para cualquier valor de la variable explicativa $\mathbf{x}^0$, puede estimarse $\mathrm{E}(y|\mathbf{x} = \mathbf{x}^0)$ como

$$\hat{\mathrm{E}}(y|\mathbf{x} = \mathbf{x}^0) = \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}} + \hat{\sigma}^2 \exp(\hat{\alpha}_0 + \mathbf{x}^0\hat{\boldsymbol{\delta}})/2), \tag{8.43}$$

donde

$\hat{\beta}_j$ = los coeficientes estimados por MCP.
$\hat{\alpha}_0$ = el intercepto en (8.41).
$\hat{\delta}_j$ = las pendientes de la misma regresión.

Obtener un error estándar adecuado para la predicción en (8.42) es analíticamente muy complicado, pero como en la sección 6.4, será bastante fácil obtener un error estándar usando un método de remuestreo como el bootstrap descrito en el apéndice 6A.

Obtener un intervalo de predicción es más complicado cuando se estima la heterocedasticidad de un modelo, y un estudio completo es complejo. Sin embargo, en la sección 6.4 se vieron dos ejemplos en los que la varianza del error domina la estimación del error e ignorando el error de estimación de todos los parámetros sólo se cometerá un pequeño error. Usando argumentos similares a los de la sección 6.4, un intervalo de predicción aproximado de 95% (para muestras grandes) es $\exp[-1.96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ a $\exp[1.96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$, donde $\hat{h}(\mathbf{x}^0)$ es la función de varianza estimada evaluada en $\mathbf{x}^0$, $\hat{h}(\mathbf{x}^0) = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1^0 + \dots + \hat{\delta}_k x_k^0)$. Como en la sección 6.4, este intervalo aproximado se obtiene simplemente exponenciando los extremos.

8.5 Reconsideración del modelo de probabilidad lineal

Como se vio en la sección 7.5, cuando y, la variable dependiente, es una variable binaria, el modelo contendrá heterocedasticidad, a menos que todos los parámetros de pendiente sean cero. Ahora se está en posición de tratar este problema.

La manera más sencilla de tratar la heterocedasticidad en el modelo de probabilidad lineal es continuar usando la estimación de MCO, pero calcular también errores estándar robustos en los estadísticos de prueba. Esto ignora el hecho de que en realidad se conoce la forma de la heterocedasticidad en el MPL. No obstante, la estimación por MCO del MPL es sencilla y con frecuencia produce resultados satisfactorios.

Ejemplo 8.8

[Participación de las mujeres casadas en la fuerza laboral]

En el ejemplo de la participación en la fuerza laboral de la sección 7.5 [vea la ecuación (7.29)], se dieron los errores estándar usuales de MCO. Ahora se calculan también los errores estándar robustos a la heterocedasticidad. Éstos se dan entre corchetes debajo de los errores estándar usuales:

$$\begin{aligned}\widehat{inlf} = &\ \underset{\substack{(.154)\\ [.151]}}{.586} - \underset{\substack{(.0014)\\ [.0015]}}{.0034}\, nwifeinc + \underset{\substack{(.007)\\ [.007]}}{.038}\, educ + \underset{\substack{(.006)\\ [.006]}}{.039}\, exper \\ & - \underset{\substack{(.00018)\\ [.00019]}}{.00060}\, exper^2 - \underset{\substack{(.002)\\ [.002]}}{.016}\, age - \underset{\substack{(.034)\\ [.032]}}{.262}\, kidslt6 + \underset{\substack{(.0132)\\ [.0135]}}{.0130}\, kidsge6 \\ & n = 753,\ R^2 = .264.\end{aligned} \tag{8.44}$$

Con el grado de precisión dado, muchos de los errores estándar de MCO y de los errores estándar robustos son iguales; en todos los casos las diferencias son muy pequeñas. Por tanto, mientras la heterocedasticidad es un problema en la teoría, no lo es en la práctica, por lo menos no en este ejemplo. Con frecuencia resulta que los errores estándar y los estadísticos de prueba usuales de MCO son similares a sus contrapartes robustas a la heterocedasticidad. Además, calcular ambos significa trabajo adicional.

En general, los estimadores de MCO son ineficientes en el MPL. Recuerde que la varianza condicional de y en el MPL es

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \tag{8.45}$$

donde

$$p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{8.46}$$

es la probabilidad de respuesta (la probabilidad de éxito, $y = 1$). Parece que lo indicado es usar mínimos cuadrados ponderados, pero hay un par de problemas. La probabilidad $p(\mathbf{x})$ depende claramente de los parámetros poblacionales desconocidos, β_j. No obstante, se tienen estimadores insesgados de estos parámetros, a saber los estimadores de MCO. Al sustituir los estimadores de MCO en la ecuación (8.46), se obtienen los valores ajustados de MCO. De manera que, para cada observación i, $\text{Var}(y_i|\mathbf{x}_i)$ se estima mediante

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i), \tag{8.47}$$

donde $\hat{y}_i$ es el valor ajustado de MCO para la observación i. Ahora se aplican, como en la sección 8.4, MCG factibles.

Por desgracia, poder estimar h_i para cada i no significa que se pueda proceder de manera directa con la estimación de los MCP. El problema es uno que se vio brevemente en la sección 7.5: los valores ajustados $\hat{y}_i$ no necesitan caer en el intervalo unitario. Si $\hat{y}_i < 0$ o $\hat{y}_i > 1$, la ecuación

(8.47) indica que $\hat{h}_i$ será negativo. Como MCP consiste en multiplicar la observación i por $1/\sqrt{\hat{h}_i}$, este método no se podrá realizar si para alguna observación $\hat{h}_i$ es negativo (o cero). En otras palabras, en MCP todos los ponderadores deben ser positivos.

En algunos casos, $0 < \hat{y}_i < 1$ para toda i, en cuyo caso puede emplearse MCP para estimar el MPL. En los casos en los que hay muchas observaciones y las probabilidades de éxito o fracaso son pequeñas, es muy común que algunos de los valores ajustados se encuentren fuera del intervalo unitario. Si esto ocurre, como en la ecuación (8.44) del ejemplo de la participación en la fuerza laboral, lo más fácil es abandonar los MCP y reportar estadísticos robustos a la heterocedasticidad. Una alternativa es arreglar aquellos valores ajustados que son menores que cero o mayores que la unidad y después aplicar MCP. Una sugerencia es hacer $\hat{y}_i = .01$ si $\hat{y}_i < 0$ y $\hat{y}_i = .99$ si $\hat{y}_i > 1$. Por desgracia, para esto se requiere una decisión arbitraria del investigador —¿por qué no usar .001 y .999 como valores ajustados? Si muchos de los valores ajustados están fuera del intervalo unitario, ajustarlos puede afectar los resultados; en este caso, tal vez lo mejor sea simplemente emplear MCO.

ESTIMACIÓN DEL MODELO DE PROBABILIDAD LINEAL MEDIANTE MÍNIMOS CUADRADOS PONDERADOS:

1. Estimar el modelo mediante MCO y obtener los valores ajustados, $\hat{y}$.
2. Determinar si todos los valores ajustados se encuentran dentro del intervalo unitario. Si es así, seguir con el paso 3). Si no, será necesario realizar algunos ajustes para llevar a todos los valores ajustados al intervalo unitario.
3. Calcular las varianzas estimadas de la ecuación (8.47).
4. Estimar la ecuación

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

por MCP, usando como ponderadores $1/\hat{h}$.

Ejemplo 8.9

[Determinantes de la posesión de una computadora personal]

Se usan los datos del archivo GPA1.RAW para estimar la probabilidad de poseer una computadora. Sea *PC* un indicador binario igual a uno si el estudiante posee una computadora e igual a cero si no es así. La variable *hsGPA* es el GPA (promedio de calificaciones) de la universidad, *ACT* es la puntuación en el examen de admisión y *parcoll* es un indicador binario que es igual a uno si por lo menos uno de los padres asistió a la universidad. (Indicadores de asistencia a la universidad distintos para el padre y la madre no dan resultados individualmente significativos, debido a que éstos están fuertemente correlacionados.)

La ecuación estimada mediante MCO es

$$\begin{aligned}
\widehat{PC} = &-.0004 + .065\ hsGPA + .0006\ ACT + .221\ parcoll \\
&(.4905)\quad (.137)\qquad\quad (.0155)\qquad\quad (.093) \\
&[.4888]\quad [.139]\qquad\quad [.0158]\qquad\quad [.087] \\
n = &\ 141,\ R^2 = .0415.
\end{aligned} \tag{8.48}$$

Como ocurrió en el ejemplo 8.8, no hay diferencias estrictas entre los errores estándar usuales y los errores estándar robustos. Sin embargo, también se estima el modelo mediante MCP. Debido a que todos los valores ajustados se encuentran dentro del intervalo unitario, no es necesario hacer ningún ajuste:

$$\begin{aligned}\widehat{PC} &= \underset{(.477)}{.026} + \underset{(.130)}{.033}\, hsGPA + \underset{(.0155)}{.0043}\, ACT + \underset{(.086)}{.215}\, parcoll \\ n &= 141,\ R^2 = .0464.\end{aligned} \tag{8.49}$$

No hay diferencias importantes entre las estimaciones de MCO y las de MCP. La única variable explicativa significativa es *parcoll*, y en ambos casos se estima que la probabilidad de poseer una computadora es aproximadamente .22 mayor si por lo menos uno de los padres asistió a la universidad.

RESUMEN

Se empezó por repasar las propiedades de mínimos cuadrados ordinarios en presencia de heterocedasticidad. La heterocedasticidad no causa sesgo o inconsistencia en los estimadores de MCO, pero los errores estándar usuales y los estadísticos de prueba dejan de ser válidos. Se mostró cómo calcular errores estándar y estadísticos t robustos a la heterocedasticidad, lo cual es realizado de manera rutinaria por muchos paquetes de regresión. La mayoría de ellos también calculan un estadístico tipo F robusto a la heterocedasticidad.

Se vieron dos maneras de probar heterocedasticidad: la prueba de Breusch-Pagan y un caso especial de la prueba de White. Estos dos estadísticos implican regresar los residuales *cuadrados* de MCO, ya sea sobre las variables independientes (BP) o sobre los valores ajustados y los valores ajustados cuadrados (White). Una prueba F sencilla es asintóticamente válida; hay también versiones del multiplicador de Lagrange de esta prueba.

En presencia de heterocedasticidad, MCO ya no es el mejor estimador lineal insesgado. Cuando se conoce la forma de la heterocedasticidad, puede usarse la estimación por mínimos cuadrados generalizados (MCG). Esto lleva a mínimos cuadrados ponderados como un medio de obtener el estimador MELI. Los estadísticos de prueba de una estimación por MCP son exactamente válidos cuando el término del error está distribuido normalmente o asintóticamente válidos en ausencia de normalidad. Esto, por supuesto, si se tiene el modelo adecuado de heterocedasticidad.

Por lo general, es necesario estimar un modelo para heterocedasticidad antes de aplicar MCP. El estimador de MCG *factibles* resultante ya no es insesgado, pero es consistente y asintóticamente eficiente. Los estadísticos usuales de la regresión de MCP son asintóticamente válidos. Se vio un método que asegura que las varianzas estimadas sean estrictamente positivas para todas las observaciones, lo cual se necesita para aplicar MCP.

Como se vio en el capítulo 7, el modelo de probabilidad lineal para una variable binaria dependiente, necesariamente tiene un término de error heterocedástico. Una manera sencilla de enfrentar este problema es calcular estadísticos robustos a la heterocedasticidad. De manera alterna, si todos los valores ajustados (es decir, las probabilidades estimadas) se encuentran estrictamente entre cero y uno, pueden emplearse mínimos cuadrados ponderados para obtener estimadores asintóticamente eficientes.

TÉRMINOS CLAVE

Error estándar robusto a la heterocedasticidad
Estadístico F robusto a la heterocedasticidad
Estadístico ML robusto a la heterocedasticidad
Estadístico t robusto a la heterocedasticidad
Estimador de MCG factibles (MCGF)
Estimadores de mínimos cuadrados generalizados (MCG)
Estimadores de mínimos cuadrados ponderados (MCP)
Heterocedasticidad de forma desconocida
Prueba de Breusch-Pagan para heterocedasticidad (Prueba BP)
Prueba de White para heterocedasticidad

PROBLEMAS

8.1 ¿Cuál de las siguientes opciones es consecuencia de la heterocedasticidad?

i) Los estimadores de MCO, $\hat{\beta}_j$, son inconsistentes.

ii) El estadístico F usual ya no tiene una distribución F.

iii) Los estimadores de MCO ya no son MELI.

8.2 Considere un modelo lineal para explicar el consumo mensual de cerveza (*beer*):

$$beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u$$

$$\mathrm{E}(u|inc,price,educ,female) = 0$$

$$\mathrm{Var}(u|inc,price,educ,female) = \sigma^2 inc^2.$$

Escriba la ecuación transformada que tiene un término de error homocedástico.

8.3 Verdadero o falso: MCP se prefiere a MCO cuando se ha omitido una variable importante del modelo.

8.4 Usando los datos del archivo GPA3.RAW se estimó la siguiente ecuación para los estudiantes que en otoño están en el segundo semestre:

$$\begin{aligned}
\widehat{trmgpa} = {} & -2.12 + .900\ crsgpa + .193\ cumgpa + .0014\ tothrs \\
& \quad (.55) \quad (.175) \qquad\quad (.064) \qquad\quad (.0012) \\
& \quad [.55] \quad [.166] \qquad\quad [.074] \qquad\quad [.0012] \\
& + .0018\ sat - .0039\ hsperc + .351\ female - .157\ season \\
& \quad (.0002) \qquad (.0018) \qquad\quad (.085) \qquad\quad (.098) \\
& \quad [.0002] \qquad [.0019] \qquad\quad [.079] \qquad\quad [.080] \\
n = {} & 269,\ R^2 = .465.
\end{aligned}$$

Aquí *trmgpa* es el promedio general de calificaciones (GPA) del semestre actual, *crsgpa* es un promedio ponderado de calificaciones de los cursos que se están tomando, *cumgpa* es el promedio general de calificaciones antes del semestre presente, *tothrs* es el total de horas crédito antes de este semestre, *sat* es la puntuación en la prueba SAT de admisión a la universidad, *hsperc* es el percentil que ocupó el alumno entre los graduados del bachillerato, *female* es una variable binaria para el género femenino, y *season* es una variable binaria que es igual a uno si el deporte del estudiante se practica en otoño. Entre paréntesis y entre corchetes se dan, respectivamente, los errores estándar usuales y los errores estándar robustos a la heterocedasticidad.

i) ¿Tienen las variables *crsgpa*, *cumgpa* y *tothrs* los efectos estimados esperados? ¿Cuáles de estas variables son estadísticamente significativas al nivel de 5%? ¿Importa qué error estándar se use?

ii) ¿Por qué es razonable la hipótesis H_0: $\beta_{crsgpa} = 1$? Pruebe esta hipótesis contra la alternativa de dos colas al nivel de 5%, usando los dos errores estándar. Describa sus conclusiones.

iii) Pruebe si el hecho de que el deporte del estudiante se practique en otoño tiene efecto sobre el GPA del semestre, usando ambos errores estándar. ¿El nivel de significancia al que se puede rechazar la prueba depende de cuál error estándar se emplee?

8.5 La variable *smokes* es una variable binaria igual a uno si la persona fuma e igual a cero si no es así. Utilizando los datos del archivo SMOKE.RAW, se estima un modelo de probabilidad lineal para *smokes*:

$$
\begin{aligned}
\widehat{smokes} = {} & \underset{\substack{(.855)\\[.856]}}{.656} - \underset{\substack{(.204)\\[.207]}}{.069}\log(cigpric) + \underset{\substack{(.026)\\[.026]}}{.012}\log(income) - \underset{\substack{(.006)\\[.006]}}{.029}\,educ \\
& + \underset{\substack{(.006)\\[.005]}}{.020}\,age - \underset{\substack{(.00006)\\[.00006]}}{.00026}\,age^2 - \underset{\substack{(.039)\\[.038]}}{.101}\,restaurn - \underset{\substack{(.052)\\[.050]}}{.026}\,white \\
& n = 807,\ R^2 = .062.
\end{aligned}
$$

La variable *white* es igual a uno si el entrevistado es blanco e igual a cero si no es así; las demás variables independientes fueron definidas en el ejemplo 8.7. Se dan tanto los errores estándar usuales como los errores estándar robustos a la heterocedasticidad.

i) ¿Hay alguna diferencia importante entre los dos conjuntos de errores estándar?

ii) Permaneciendo todos los demás factores constantes, si la educación aumenta cuatro años, ¿qué ocurre con la probabilidad estimada de fumar?

iii) ¿En qué punto un año más de edad reduce la probabilidad de fumar?

iv) Interprete el coeficiente de la variable binaria *restaurn* (una variable que es igual a uno si la persona vive en un lugar en el que hay restricciones para fumar en los restaurantes).

v) La persona número 206 de la base de datos tiene las características siguientes: $cigpric = 67.44$, $income = 6{,}500$, $educ = 16$, $age = 77$, $restaurn = 0$, $white = 0$ y $smokes = 0$. Calcule la probabilidad predicha de que esta persona fume. Comente sobre el resultado.

8.6 Existen diversas maneras de combinar las características de las pruebas de Breusch-Pagan y de White para heterocedasticidad. Una posibilidad no vista en el libro es correr la regresión

$$\hat{u}_i^2 \text{ sobre } x_{i1}, x_{i2}, \ldots, x_{ik}, \hat{y}_i^2,\ i = 1, \ldots, n,$$

donde las $\hat{u}_i$ son los residuales de MCO y las $\hat{y}_i$ son los valores ajustados de MCO. Después se prueba la significancia conjunta de $x_{i1}, x_{i2}, \ldots, x_{ik}$ y $\hat{y}_i^2$. (Por supuesto, en esta regresión siempre se incluye un intercepto.)

i) ¿Cuáles son los *gl* correspondientes a la prueba *F* propuesta para heterocedasticidad?

ii) Explique por qué la *R*-cuadrada de la regresión indicada arriba siempre será por lo menos tan grande como la *R*-cuadrada de la regresión BP y del caso especial de la prueba de White.

iii) ¿Implica el inciso ii) que con la nueva prueba siempre se obtiene un valor-*p* menor que el del estadístico BP o que el del caso especial del estadístico de White? Explique.

iv) Suponga que alguien sugiere agregar también $\hat{y}_i$ a la prueba recién propuesta. ¿Qué le parece esta idea?

8.7 Considere un modelo para los empleados,

$$y_{i,e} = \beta_0 + \beta_1 x_{i,e,1} + \beta_2 x_{i,e,2} + \ldots + \beta_k x_{i,e,k} + f_i + v_{i,e},$$

donde la variable inobservada f_i es un "efecto de la empresa" para cada empleado en una empresa dada i. El término del error $v_{i,e}$ es específico para cada empleado e en la empresa i. El *error compuesto* es $u_{i,e} = f_i + v_{i,e}$, como en la ecuación (8.28).

i) Suponga que $\mathrm{Var}(f_i) = \sigma_f^2$, $\mathrm{Var}(v_{i,e}) = \sigma_v^2$, y que f_i y $v_{i,e}$ no están correlacionadas. Muestre que $\mathrm{Var}(u_{i,e}) = \sigma_f^2 + \sigma_v^2$; llame a esto σ^2.

ii) Ahora suponga que para $e \neq g$, $v_{i,e}$ y $v_{i,g}$ no están correlacionadas. Muestre que $\mathrm{Cov}(u_{i,e}, u_{i,g}) = \sigma_f^2$.

iii) Sea $\bar{u}_i = m_i^{-1}\sum_{\hat{e}=1}^{m_i} u_{i,e}$ el promedio de los errores compuestos dentro de una empresa. Muestre que $\mathrm{Var}(\bar{u}_i) = \sigma_f^2 + \sigma_v^2/m_i$.

iv) Analice la relevancia del inciso iii) para la estimación por MCP empleando datos promediados a nivel de las empresas, donde el ponderador empleado para la observación i es el tamaño de la firma, como de costumbre.

EJERCICIOS EN COMPUTADORA

C8.1 Considere el modelo siguiente para explicar el comportamiento del sueño:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u.$$

i) Dé un modelo que permita que la varianza de u difiera entre hombres (*male*) y mujeres. La varianza no debe depender de otros factores.

ii) Emplee los datos del archivo SLEEP75.RAW para estimar los parámetros de la ecuación de heterocedasticidad. (Tiene que estimar la ecuación para *sleep* primero mediante MCO para obtener los residuales de MCO.) ¿Es la varianza estimada de u mayor para los hombres o para las mujeres?

iii) ¿Es la varianza de u diferente estadísticamente para hombres y para mujeres?

C8.2 i) Emplee los datos del archivo HPRICE1.RAW para obtener errores estándar robustos a la heterocedasticidad para la ecuación (8.17). Analice cualquier diferencia importante con los errores estándar usuales.

ii) Repita el inciso i) para la ecuación (8.18).

iii) ¿Qué indican estos ejemplos acerca de la heterocedasticidad y de la transformación usada para la variable dependiente?

C8.3 Aplique la prueba completa de White para heterocedasticidad [ver ecuación (8.19)] a la ecuación (8.18). Usando la forma ji-cuadrada del estadístico, obtenga el valor-p. ¿Qué concluye usted?

C8.4 Para este ejercicio emplee el archivo VOTE1.RAW.

i) Estime un modelo en el que *voteA* sea la variable dependiente y *prtystrA*, *democA*, log(*expendA*) y log(*expendB*) sean las variables independientes. Obtenga los residuales de MCO, $\hat{u}_i$, y regrese éstos sobre todas las variables independientes. Explique por qué obtiene $R^2 = 0$.

ii) Ahora calcule la prueba de Breusch-Pagan para heterocedasticidad. Emplee la versión del estadístico F y dé el valor-p.

iii) Calcule el caso especial de la prueba de White para heterocedasticidad, usando de nuevo la forma del estadístico F. ¿Qué tan fuerte es ahora la evidencia de heterocedasticidad?

C8.5 Para este ejercicio emplee los datos del archivo PNTSPRD.RAW.

i) La variable *sprdcvr* es una variable binaria que es igual a uno si en un partido de baloncesto universitario se cubrió la diferencia de puntos predicha en Las Vegas. El valor esperado para *sprdcvr*, por ejemplo μ, es la probabilidad de que la diferencia sea cubierta en un partido seleccionado al azar. Pruebe H_0: $\mu = .5$ contra H_1: $\mu \neq .5$ al nivel de significancia de 10% y analice sus hallazgos. (*Sugerencia:* esto puede hacerse fácilmente usando una prueba t en la regresión de *sprdcvr* sobre sólo un intercepto.)

ii) ¿Cuántos de los 553 partidos se jugaron en un campo neutral?

iii) Estime el modelo de probabilidad lineal

$$sprdcvr = \beta_0 + \beta_1 favhome + \beta_2 neutral + \beta_3 fav25 + \beta_4 und25 + u$$

y reporte los resultados en la forma habitual. (Dé los errores estándar usuales de MCO y los errores estándar robustos a la heterocedasticidad.) ¿Qué variable es más significativa tanto estadística como prácticamente?

iv) Explique por qué bajo la hipótesis nula H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, no hay heterocedasticidad en el modelo.

v) Emplee el estadístico F usual para probar la hipótesis del inciso iv). ¿Qué concluye usted?

vi) Dado el análisis previo, ¿diría que es posible predecir de manera sistemática si la diferencia de puntos predicha en Las Vegas se logrará usando la información disponible antes del partido?

C8.6 En el ejemplo 7.12 se estimó un modelo de probabilidad lineal para determinar si un joven había sido arrestado durante 1986:

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u.$$

i) Estime este modelo mediante MCO y verifique que todos los valores ajustados estén estrictamente entre cero y uno. ¿Cuáles son los valores ajustados menor y mayor?

ii) Estime la ecuación mediante mínimos cuadrados ponderados, como se analizó en la sección 8.5.

iii) Use las estimaciones de MCP para determinar si *avgsen* y *tottime* son conjuntamente significativas al nivel de 5%.

C8.7 Para este ejercicio emplee los datos del archivo LOANAPP.RAW.

i) Estime la ecuación del inciso iii) del ejercicio para computadora C7.8, calculando los errores estándar robustos a la heterocedasticidad. Compare el intervalo de confianza de 95% para β_{white} con el intervalo de confianza no robusto.

ii) Obtenga los valores ajustados con la regresión del inciso i). ¿Son algunos de ellos menores a cero? ¿Qué significa esto acerca de la aplicación de mínimos cuadrados ponderados?

C8.8 Para este ejercicio emplee los datos del archivo GPA1.RAW.

i) Emplee MCO para estimar un modelo que relacione *colGPA* con *hsGPA*, *ACT*, *skipped* y *PC*. Obtenga los residuales de MCO.

ii) Calcule el caso especial de la prueba de White para heterocedasticidad. En la regresión de $\hat{u}_i^2$ sobre $\widehat{colGPA}_i$, $\widehat{colGPA}_i^2$, obtenga los valores ajustados, es decir $\hat{h}_i$.

iii) Verifique que los valores ajustados del inciso ii) sean todos estrictamente positivos. Después obtenga las estimaciones de mínimos cuadrados ponderados usando como ponderadores $1/\hat{h}_i$. Compare las estimaciones de mínimos cuadrados ponderados con las correspondientes estimaciones de MCO para el efecto del número de clases perdidas por semana (*skipped*) y el efecto de poseer una computadora (PC). ¿Qué puede decir de su significancia estadística?

iv) En la estimación por MCP del inciso iii), obtenga los errores estándar robustos a la heterocedasticidad. En otras palabras, considere la posibilidad de que la función de varianza estimada en el inciso ii) pueda estar mal especificada. (Vea la pregunta 8.4.) ¿Varían mucho los errores estándar de los del inciso iii)?

C8.9 En el ejemplo 8.7 se calcularon las estimaciones de MCO y un conjunto de estimaciones de MCP en una ecuación sobre la demanda de cigarros.

i) Obtenga las estimaciones de MCO en la ecuación (8.35).

ii) Obtenga los $\hat{h}_i$ usados en la estimación de MCP de la ecuación (8.36) y reproduzca la ecuación (8.36). A partir de esta ecuación, obtenga los residuales *no ponderados* y los valores ajustados; llámeles $\hat{u}_i$ y $\hat{y}_i$, respectivamente. (Por ejemplo, en Stata, los residuales no ponderados y los valores ajustados se dan automáticamente.)

iii) Sean $\breve{u}_i = \hat{u}_i/\sqrt{\hat{h}_i}$ y $\breve{y}_i = \hat{y}_i/\sqrt{\hat{h}_i}$ las cantidades ponderadas. Realice el caso especial de la prueba de White para heterocedasticidad regresando $\breve{u}_i^2$ sobre $\breve{y}_i$, $\breve{y}_i^2$, teniendo cuidado de incluir un intercepto, como siempre. ¿Encuentra heterocedasticidad en los residuales ponderados?

iv) ¿Qué implican los hallazgos del inciso iii) respecto a la forma de heterocedasticidad usada para obtener (8.36)?

v) Obtenga errores estándar válidos para las estimaciones de MCP que permitan que la función de la varianza esté mal especificada.

C8.10 Para este ejercicio emplee la base de datos del archivo 401KSUBS.RAW.

i) Empleando MCO estime un modelo de probabilidad lineal para *e401k*, usando como variables explicativas *inc*, *inc*^{2}, *age*, *age*^{2} y *male*. Obtenga tanto los errores estándar usuales de MCO como sus versiones robustas a la heterocedasticidad. ¿Existen diferencias importantes?

ii) En el caso especial de la prueba de White para heterocedasticidad en la que se regresan los residuales cuadrados de MCO sobre un cuadrado de los valores ajustados por MCO, $\hat{u}_i^2$ sobre $\hat{y}_i, \hat{y}_i^2$, $i = 1, \ldots, n$, argumente que el límite de probabilidad del coeficiente de $\hat{y}_i$ debe de ser uno, el límite de probabilidad del coeficiente de $\hat{y}_i^2$ debe de ser -1, y que el límite de probabilidad del intercepto debe de ser cero. {*Sugerencia:* recuerde que $\text{Var}(y|x_1, \ldots, x_k) = p(\mathbf{x})[1 - p(\mathbf{x})]$, donde $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.}

iii) Para el modelo estimado en el inciso i), obtenga la prueba de White y vea si las estimaciones de los coeficientes corresponden aproximadamente a los valores teóricos descritos en el inciso ii).

iv) Después de verificar que todos los valores ajustados del inciso i) se encuentran entre cero y uno, obtenga las estimaciones de mínimos cuadrados ponderados del modelo de probabilidad lineal. ¿Difieren éstas, de manera importante, de las estimaciones de MCO?

C8.11 Para esta pregunta utilice los datos del archivo 401KSUBS.RAW, restringiendo la muestra a *fsize* = 1.

i) Al modelo estimado en la tabla 8.1, agregue el término de interacción, *e401k* · *inc*. Estime esta ecuación por MCO y obtenga los errores estándar usuales y robustos. ¿Qué concluye acerca de la significancia estadística del término de interacción?

ii) Ahora estime el modelo más general por MCP empleando los mismos ponderadores, $1/inc_i$, que en la tabla 8.1. Calcule los errores estándar usuales y robustos para los estimadores de MCP. Usando los errores estándar robustos, ¿es el término de interacción estadísticamente significativo?

iii) En el modelo más general, analice el coeficiente de MCP de *e401k*. ¿Es este coeficiente de mucho interés en sí mismo? Explique.

iv) Vuelva a estimar el modelo por MCP, pero use el término de interacción $e401k \cdot (inc - 30)$; el ingreso promedio en la muestra es aproximadamente 29.44. Interprete ahora el coeficiente de *e401k*.

C8.12 Para responder esta pregunta emplee los datos del archivo MEAP00_01.RAW.

i) Estime el modelo

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 \log(enroll) + \beta_3 \log(exppp) + u$$

mediante MCO y obtenga los errores estándar usuales y los errores estándar completamente robustos. ¿Qué observa al compararlos?

ii) Aplique el caso especial de la prueba de White para heterocedasticidad. ¿Cuál es el valor de la prueba F? ¿Qué concluye usted?

iii) Obtenga los $\hat{g}_i$ como los valores ajustados de la regresión de $\log(\hat{u}_i^2)$ sobre $\widehat{math4}_i$, $\widehat{math4}_i^2$, donde $\widehat{math4}_i$ son los valores ajustados de MCO y las $\hat{u}_i$ son los residuales de MCO. Sea $\hat{h}_i = \exp(\hat{g}_i)$. Use las $\hat{h}_i$ para obtener las estimaciones de MCP. ¿Hay grandes diferencias con los coeficientes de MCO?

iv) Obtenga los errores estándar de MCP que permiten una mala especificación de la función de la varianza. ¿Difieren mucho estos errores estándar de los errores estándar usuales de MCP

v) Para estimar el efecto del gasto (*exppp*) sobre *math*4, ¿parece ser más preciso MCO o MCP?

CAPÍTULO 9

Más sobre especificación y temas de datos

En el capítulo 8 se trató la falla de uno de los supuestos de Gauss-Markov. Aunque la heterocedasticidad de los errores puede verse como un problema de un modelo, es uno relativamente menor. La presencia de heterocedasticidad no causa sesgo o inconsistencia en los estimadores de MCO. Además, es bastante fácil ajustar los intervalos de confianza y los estadísticos t y F para obtener inferencias válidas después de la estimación de MCO, o incluso obtener estimadores más eficientes usando mínimos cuadrados ponderados.

En este capítulo se vuelve al problema mucho más serio de la correlación entre el error, u, y una o más de las variables explicativas. Recuerde que en el capítulo 3 se vio que si u está, por la razón que sea, correlacionada con la variable explicativa x_j, entonces se dice que x_j es una **variable explicativa endógena**. Se proporciona también un análisis más detallado sobre tres razones por las que una variable explicativa puede ser endógena; en algunos casos se analizan posibles soluciones.

En los capítulos 3 y 5 se vio que omitir una variable clave puede causar la correlación entre el error y algunas de las variables explicativas, lo que en general conduce al sesgo e inconsistencia en *todas* las estimaciones de MCO. En el caso especial en que la variable omitida sea una función de una variable explicativa ya incluida en el modelo, éste sufre de una **especificación incorrecta de la forma funcional**.

En la primera sección se empieza por analizar las consecuencias de la especificación incorrecta de la forma funcional y cómo detectarla. En la sección 9.2 se muestra cómo el uso de variables proxy puede solucionar, o al menos atenuar, el sesgo de las variables omitidas. En la sección 9.3 se deduce y explica el sesgo en MCO que puede presentarse bajo ciertas formas de **error de medición**. Los problemas de datos adicionales se estudian en la sección 9.4.

Todos los procedimientos de este capítulo se basan en la estimación por MCO. Como se verá, ciertos problemas que causan correlación entre el error y algunas variables explicativas no se pueden resolver usando MCO en un corte transversal sencillo. El estudio de los métodos de estimación alternativos se pospone hasta la parte 3.

9.1 Especificación incorrecta de la forma funcional

Un modelo de regresión múltiple sufre de especificación incorrecta de la forma funcional cuando no explica de manera correcta la relación entre la variable dependiente y las variables explicativas observadas. Por ejemplo, si el salario por hora se determina mediante $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$, pero se omite el término cuadrado de experiencia, $exper^2$, entonces se comete un error de especificación de la forma funcional. Se sabe ya, de acuerdo con el capítulo 3, que esto en general lleva a estimadores sesgados de β_0, β_1 y β_2. (No se estima β_3 porque $exper^2$ se ha excluido del modelo.) Así, especificar de manera incorrecta cómo *exper* afecta log(*wage*) en general da como resultado un estimador sesgado para el rendimiento de la educación, β_1. La magnitud de este sesgo depende del tamaño de β_3 y de la correlación entre *educ*, *exper* y $exper^2$.

Es peor cuando se estima el rendimiento de la experiencia: aunque podría obtenerse un estimador insesgado de β_2, el rendimiento de la experiencia no podría estimarse, debido a que ésta es igual a $\beta_2 + 2\beta_3 exper$ (en forma decimal). Usar el estimador sesgado de β_2 puede ser engañoso, en especial en los valores extremos de *exper*.

Para dar otro ejemplo, suponga que la ecuación de log(*wage*) es

$$\begin{aligned}\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 \\ + \beta_4 female + \beta_5 female{\cdot}educ + u,\end{aligned} \tag{9.1}$$

donde *female* es una variable binaria. Si se omite el término de interacción, *female·educ*, entonces se estará especificando de manera incorrecta la forma funcional. En general, no se obtendrán estimadores insesgados de ninguno de los demás parámetros, y puesto que el rendimiento de la educación depende del género, no es claro qué se estará estimando al omitir el término de interacción.

La omisión de funciones de variables independientes no es la única manera en la que un modelo puede sufrir de especificación incorrecta de la forma funcional. Por ejemplo, si (9.1) es el verdadero modelo que satisface los primeros cuatro supuestos de Gauss-Markov, pero como variable dependiente se utiliza *wage* en lugar de log(*wage*), entonces no se obtendrán estimadores insesgados o consistentes de los efectos parciales. La prueba que sigue tiene cierta capacidad para detectar esta clase de problema de la forma funcional, pero existen mejores pruebas, las cuales se mencionarán en la subsección sobre alternativas no anidadas.

La especificación incorrecta de la forma funcional de un modelo puede ciertamente tener serias consecuencias. Sin embargo, el problema es menor en un aspecto importante: por definición, se tienen datos de todas las variables necesarias para obtener una relación funcional que se ajuste bien a los datos. Esto puede compararse con el problema que se trata en la sección siguiente, donde se omite una variable clave sobre la cual no es posible recolectar datos.

Ya se tiene una herramienta muy poderosa para detectar formas funcionales incorrectamente especificadas: la prueba F para restricciones de exclusión conjunta. Suele ser recomendable agregar a un modelo términos cuadráticos de cualquiera de las variables significativas y realizar una prueba de significancia conjunta. Si los términos cuadráticos adicionados son significativos, pueden ser agregados al modelo (al costo de complicar la interpretación del modelo). Sin embargo, los términos cuadráticos significativos pueden ser sintomáticos de otros problemas de la forma funcional, tales como usar el nivel de una variable cuando el logaritmo es más apropiado, o viceversa. Establecer con claridad la razón exacta por la que una forma funcional está incorrectamente especificada puede ser difícil. Por fortuna, en muchos casos, usar los logaritmos de ciertas variables y agregar términos cuadráticos es suficiente para detectar muchas relaciones no lineales importantes en la economía.

Ejemplo 9.1

[Modelo económico de la delincuencia]

En la tabla 9.1 se presentan estimaciones de MCO correspondientes al modelo económico de la delincuencia (vea el ejemplo 8.3). Primero se estima el modelo sin ningún término cuadrático; estos resultados se presentan en la columna (1).

TABLA 9.1

Variable dependiente: *narr86*

Variables independientes	(1)	(2)
pcnv	–.133 (.040)	.533 (.154)
$pcnv^2$	—	–.730 (.156)
avgsen	–.011 (.012)	–.017 (.012)
tottime	.012 (.009)	.012 (.009)
ptime86	–.041 (.009)	.287 (.004)
$ptime86^2$	—	–.0296 (.0039)
qemp86	–.051 (.014)	–.014 (.017)
inc86	–.0015 (.0003)	–.0034 (.0008)
$inc86^2$	—	–.000007 (.000003)
black	.327 (.045)	.292 (.045)
hispan	.194 (.040)	.164 (.039)
intercepto	.596 (.036)	.505 (.037)
Observaciones *R*-cuadrada	2,725 .0723	2,725 .1035

En la columna (2) se han agregado los cuadrados de *pcnv*, *ptime86* e *inc86*. Los cuadrados de estas variables se incluyen porque en la columna (1) cada uno de los términos lineales es significativo. La variable *qemp86* es una variable discreta que sólo toma cinco valores, de manera que en la columna (2) no se incluye su cuadrado.

Cada uno de estos términos cuadrados es significativo y juntos son conjuntamente muy significativos ($F = 31.37$, con $gl = 3$ y 2,713; el valor-p es esencialmente cero). Por tanto, parece que el modelo inicial pasó por alto algunas no linealidades potencialmente importantes.

Pregunta 9.1

¿Por qué no se incluyen los cuadrados de *black* y de *hispan* en la columna (2) de la tabla 9.1?

La presencia de los términos cuadráticos dificulta un poco la interpretación del modelo. Por ejemplo, *pcnv* ya no tiene un efecto estrictamente disuasivo: la relación entre *narr86* y *pcnv* es positiva hasta $pcnv = .365$, y después la relación es negativa. Puede concluirse que a valores más bajos de *pcnv*; el efecto disuasivo es poco o nulo; el efecto sólo entra en funcionamiento a tasas superiores de condenas previas. Para verificar esta conclusión tendrían que utilizarse formas funcionales más sofisticadas que las cuadráticas. Puede ser que *pcnv* no sea completamente exógena. Por ejemplo, los hombres que no hayan sido condenados en el pasado (por lo que $pcnv = 0$) quizá son delincuentes ocasionales, y por tanto es menos probable que sean arrestados en 1986. Esto podría sesgar las estimaciones.

De manera similar, la relación entre *narr86* y *ptime86* es positiva hasta que $ptime86 = 4.85$ (casi cinco meses en prisión), y después esta relación es negativa. La mayoría de hombres de la muestra no estuvieron en la prisión en 1986, de manera que, de nuevo, se debe tener cuidado al interpretar los resultados.

El ingreso legal tiene un efecto negativo sobre *narr86 hasta inc86* = 242.85; puesto que el ingreso se mide en cientos de dólares, esto significa un ingreso anual de 24,285 dólares. Sólo 46 de los hombres de la muestra tienen ingresos superiores a este nivel. Por tanto, se puede concluir que *narr86* e *inc86* están relacionadas negativamente con un efecto decreciente.

El ejemplo 9.1 es un difícil problema de la forma funcional debido a la naturaleza de la variable dependiente. Otros modelos son teóricamente más adecuados para manejar variables dependientes que toman una cantidad reducida de valores enteros. Estos modelos se verán brevemente en el capítulo 17.

RESET como una prueba general para especificación incorrecta de formas funcionales

Se han propuesto algunas pruebas para detectar en general la especificación incorrecta de formas funcionales. La **prueba de error de especificación de la regresión (RESET)** de Ramsey (1969) ha demostrado ser útil a este respecto.

La idea detrás de RESET es bastante sencilla. Si el modelo original

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u \tag{9.2}$$

satisface RLM.4, ninguna función no lineal de las variables independientes será significativa al agregarla a la ecuación (9.2). En el ejemplo 9.1 se agregaron cuadrados de variables explicativas significativas. Aunque esto suele detectar problemas de la forma funcional, tiene la desventaja de requerir muchos grados de libertad si en el modelo original hay muchas variables explicativas (consume tantos grados de libertad como la forma directa de la prueba de White para heterocedasticidad). Además, ciertos tipos de no linealidades omitidas no se detectarán agregando términos cuadráticos. RESET, para detectar tipos generales de especificación incorrecta de la forma funcional, agrega a la ecuación (9.2) polinomios en los valores ajustados de MCO.

Para realizar RESET, hay que decidir cuántas funciones de los valores ajustados incluir en la regresión ampliada. No existe respuesta correcta para esta pregunta, pero los términos cuadrados y cúbicos han demostrado ser útiles en la mayoría de las aplicaciones.

Sea $\hat{y}$ el valor ajustado obtenido de la estimación por MCO de (9.2). Considere la ecuación ampliada

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error. \qquad \textbf{9.3}$$

Esta ecuación resulta un poco extraña, porque ahora como variables explicativas aparecen funciones de los valores ajustados de la estimación inicial. En realidad lo que interesa no son los parámetros estimados con (9.3); esta ecuación sólo se usa para probar si (9.2) ha ignorado no linealidades importantes. Lo que no hay que olvidar es que $\hat{y}^2$ y $\hat{y}^3$ son funciones lineales de las x_j.

La hipótesis nula es que (9.2) está correctamente especificada. De manera que RESET es el estadístico F para probar H_0: $\delta_1 = 0$, $\delta_2 = 0$ en el modelo ampliado (9.3). Un estadístico F significativo sugiere algún tipo de problema de la forma funcional. La distribución del estadístico F es aproximadamente $F_{2,n-k-3}$ en muestras grandes bajo la hipótesis nula (y los supuestos de Gauss-Markov). El número de *gl* en la ecuación ampliada (9.3) es $n - k - 1 - 2 = n - k - 3$. También existe una versión de *ML* (y la distribución ji-cuadrada tendrá dos *gl*). Además, la prueba puede hacerse robusta a la heterocedasticidad empleando los métodos estudiados en la sección 8.2.

Ejemplo 9.2

[Ecuación para el precio de la vivienda]

Se estiman dos modelos para los precios de las viviendas. En el primero, todas las variables están en forma lineal:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u. \qquad \textbf{9.4}$$

En el segundo se emplean los logaritmos de todas las variables, excepto de *bdrms*:

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u. \qquad \textbf{9.5}$$

Empleando $n = 88$ viviendas del archivo HPRICE1.RAW, el estadístico RESET de la ecuación (9.4) resulta ser 4.67; este es el valor de una variable aleatoria $F_{2,82}$ ($n = 88$, $k = 3$), y el correspondiente valor-p es .012. Esto es una evidencia de especificación incorrecta de la forma funcional en (9.4).

El estadístico RESET en (9.5) es 2.56 y su valor-p = .084. Por tanto, no se rechaza (9.5) al nivel de significancia de 5% (aunque sí se rechazaría al nivel de 10%). De acuerdo con RESET, se prefiere el modelo log-log de (9.5).

En el ejemplo anterior, se probaron dos modelos para explicar los precios de la vivienda. Uno fue rechazado por RESET, mientras que el otro no (por lo menos en el nivel de 5%). A menudo, las cosas no son tan sencillas. Una desventaja con RESET es que no proporciona ninguna indicación sobre cómo proceder si se rechaza el modelo. Rechazar (9.4) usando RESET no sugiere inmediatamente que (9.5) sea el paso siguiente. La ecuación (9.5) se estimó debido a que los modelos de elasticidad constantes son fáciles de interpretar y pueden tener características estadísticas agradables. En este ejemplo, casualmente ocurre que también pasa la prueba de la forma funcional.

Algunos opinan que RESET es una prueba muy general para la especificación incorrecta de modelos, incluyendo variables omitidas no observadas y heterocedasticidad. Por desgracia, tal uso de RESET está equivocado en gran parte. Se puede demostrar que RESET no tiene ninguna potencia para detectar variables omitidas cuando éstas tienen esperanzas lineales en las variables independientes incluidas en el modelo [vea Wooldridge (1995) para una explicación exacta]. Además, si la forma funcional está especificada correctamente, RESET no tiene potencia para detectar heterocedasticidad. El fondo del asunto es que RESET es una prueba para formas funcionales, y nada más.

Pruebas contra alternativas no anidadas

Obtener pruebas para otros tipos de especificaciones incorrectas de la forma funcional —por ejemplo, tratar de decidir si una variable independiente debe aparecer en forma lineal o logarítmica— llevaría fuera del campo de las pruebas de hipótesis clásicas. Es posible probar el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{9.6}$$

contra el modelo

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u, \tag{9.7}$$

y viceversa. Sin embargo, éstos son **modelos no anidados** (vea el capítulo 6), y, por tanto, no es posible simplemente utilizar una prueba F estándar. Se han sugerido dos métodos diferentes. El primero es construir un modelo que contenga cada modelo como un caso especial y después probar las restricciones que conduzcan a cada uno de los modelos. En el caso del ejemplo actual, el modelo es

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u. \tag{9.8}$$

Se puede probar primero H_0: $\gamma_3 = 0$, $\gamma_4 = 0$ como prueba de (9.6). También se puede probar H_0: $\gamma_1 = 0$, $\gamma_2 = 0$ como prueba de (9.7). Este método fue sugerido por Mizon y Richard (1986).

Otro método es el que ha sido sugerido por Davidson y MacKinnon (1981). Estos investigadores señalan que, si (9.6) es verdadero, entonces los valores ajustados del otro modelo, (9.7), no serán significativos en (9.6). Por tanto, para probar (9.6), primero se estima el modelo (9.7) por MCO para obtener los valores ajustados. Llámense estos $\hat{\hat{y}}$. La **prueba de Davidson-MacKinnon** se basa en el estadístico t para $\hat{\hat{y}}$ en la ecuación

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + error.$$

Un estadístico t significativo (contra una alternativa de dos colas) es un rechazo de (9.6).

De manera similar, si $\hat{y}$ denota los valores ajustados obtenidos de la estimación (9.6), la prueba de (9.7) es el estadístico t para $\hat{y}$ en el modelo

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \hat{y} + error;$$

Un estadístico t significativo es una evidencia contra (9.7). Estas mismas dos pruebas pueden utilizarse para probar cualesquiera dos modelos no anidados con la misma variable dependiente.

Las pruebas no anidadas presentan algunos problemas. Primero, no necesariamente surgirá un claro ganador. Ambos modelos pueden ser rechazados o ninguno de los dos puede ser rechazado. En el último caso, puede utilizarse la R-cuadrada ajustada para elegir entre ellos. Si ambos modelos se rechazan, será necesario más trabajo. Sin embargo, es importante conocer las consecuencias prácticas de usar una forma o la otra: si los efectos de las variables independientes clave sobre y no son muy diferentes, entonces no importa qué modelo se utilice.

Un segundo problema es que rechazar (9.6) usando, por ejemplo, la prueba de Davidson-MacKinnon, no significa que (9.7) sea el modelo correcto. El modelo (9.6) se puede rechazar para una variedad de especificaciones incorrectas de la forma funcional.

Un problema incluso más difícil es obtener pruebas no anidadas cuando los modelos que se comparan tienen variables dependientes diferentes. El caso más importante es y contra $\log(y)$. En el capítulo 6 se vio que sólo obtener medidas de la bondad de ajuste que puedan compararse requiere un cierto cuidado. Diversas pruebas han sido propuestas para solucionar este problema, pero están fuera del alcance de este libro. [Vea en Wooldridge (l994a) una prueba que tiene una interpretación sencilla y es fácil de realizar.]

9.2 Uso de las variables proxy para las variables explicativas no observadas

Un problema más difícil se presenta cuando en un modelo se excluye una variable clave, en general por no disponer de datos. Considere una ecuación para el salario en la que explícitamente se reconozca que la capacidad (*abil*) afecta el log(*wage*):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u. \quad \boxed{9.9}$$

Este modelo muestra de manera explícita que se desea mantener la capacidad constante para medir el rendimiento de *educ* y de *exper*. Si, por ejemplo, *educ* está correlacionada con *abil*, entonces colocar *abil* en el término de error hace que el estimador de MCO de β_1 (y de β_2) sea sesgado, un tema que ha aparecido en varias ocasiones.

Lo que más interesa en la ecuación (9.9) son los parámetros de pendiente β_1 y β_2. En realidad no interesa si se obtiene un estimador insesgado o consistente del intercepto β_0; pues como se verá en breve, esto en general no es posible. Tampoco se puede esperar estimar β_3 porque *abil* no es una variable observada; en realidad, de cualquier manera no se sabría cómo interpretar β_3 puesto que la capacidad es, en el mejor de los casos, un concepto vago.

¿Cómo puede resolverse, o al menos atenuarse, el sesgo de las variables omitidas en una ecuación como la (9.9)? Una posibilidad es obtener una **variable proxy** para la variable omitida. En términos vagos, una variable proxy es algo que está relacionado con la variable no observada que se desea controlar en el análisis que se realiza. En la ecuación del salario, una posibilidad es utilizar el cociente intelectual, o *IQ* (por sus siglas en inglés), como variable proxy para capacidad. Para esto *no* se requiere que *IQ* sea lo mismo que capacidad; lo que se necesita es que el *IQ* esté correlacionado con la capacidad, algo que se aclarará en el siguiente análisis.

Todas las ideas clave pueden ilustrarse mediante un modelo con tres variables independientes, dos de las cuales son observadas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u. \quad \boxed{9.10}$$

Se supone que se dispone de datos sobre y, x_1 y x_2 —en el ejemplo del salario, estas variables son log(*wage*), *educ* y *exper*, respectivamente—. La variable explicativa x_3^* no es observada, pero se tiene una variable proxy para x_3^*. Llámese x_3 a esta variable proxy.

¿Qué se requerirá de x_3? Como mínimo, deberá tener alguna relación con x_3^*. Esto se capta mediante la sencilla ecuación de regresión

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3, \quad \boxed{9.11}$$

donde v_3 es un error debido al hecho de que la relación entre x_3^* y x_3 no es exacta. El parámetro δ_3 mide la relación entre x_3^* y x_3; en general se considera que x_3^* y x_3 están relacionadas positivamente, de modo que $\delta_3 > 0$. Si $\delta_3 = 0$, entonces x_3 no es una proxy adecuada para x_3^*. El intercepto δ_0 en (9.11), que puede ser positivo o negativo, simplemente permite que x_3^* y x_3 se midan en escalas diferentes. (Por ejemplo, ciertamente no se requiere que en la población de Estados Unidos la capacidad no observada tenga el mismo valor promedio que el *IQ*.)

¿Cómo puede utilizarse x_3 para obtener estimadores insesgados (o al menos consistentes) de β_1 y β_2? La propuesta es pretender que x_3 y x_3^* son iguales, de manera que se corra la regresión de

$$y \text{ sobre } x_1, x_2, x_3. \tag{9.12}$$

A esto se le llama **solución suplente al problema de variables omitidas** debido a que x_3 simplemente suple a x_3^* antes de correr los MCO. Si x_3 en realidad está relacionada con x_3^*, esto parece ser algo razonable. Sin embargo como x_3 y x_3^* no son iguales, habrá que determinar cuándo este procedimiento realmente da estimadores consistentes de β_1 y β_2.

Los supuestos necesarios para que la solución suplente proporcione estimadores consistentes de β_1 y β_2 pueden dividirse en supuestos acerca de u y de v_3:

1. El error u no está correlacionado con x_1, x_2 y x_3^*, que es el supuesto estándar en el modelo (9.10). Además u no está correlacionada con x_3. Este último supuesto significa que una vez incluidas x_1, x_2 y x_3^*, en el modelo poblacional x_3 es irrelevante. Esto es así por definición, ya que x_3 es una variable proxy para x_3^*: la que afecta de forma directa a y es x_3^*, no x_3. Por tanto, el supuesto de que u no está correlacionada con x_1, x_2, x_3^* y x_3 no es muy controversial. (Otra manera de enunciar este supuesto es que el valor esperado de u dadas todas esas variables es cero.)
2. El error v_3 no está correlacionado con x_1, x_2 y x_3. Suponer que v_3 no esté correlacionado con x_1 y x_2 requiere que x_3 sea una "buena" proxy para x_3^*. La manera más fácil de ver esto es expresar los análogos de estos supuestos en términos de esperanzas condicionales:

$$\mathrm{E}(x_3^*|x_1, x_2, x_3) = \mathrm{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3. \tag{9.13}$$

La primera igualdad, que es la más importante, indica que, una vez controlado x_3 el valor esperado de x_3^* no depende de x_1 o de x_2. De manera alterna, x_3^* tiene correlación cero con x_1 y x_2 una vez que se les han descontado los efectos parciales de x_3.

En la ecuación (9.9) sobre el salario, donde *IQ* es la proxy para capacidad, la condición (9.13) es

$$\mathrm{E}(abil|educ, exper, IQ) = \mathrm{E}(abil|IQ) = \delta_0 + \delta_3 IQ.$$

Por tanto, el nivel promedio de capacidad sólo varía con el *IQ*, no con *educ* y *exper*. ¿Es esto razonable? Quizá no sea totalmente así, pero está cercano a serlo. Seguro vale la pena incluir el *IQ* en la ecuación para el salario con objeto de ver lo que ocurre con el rendimiento estimado de la educación.

Es fácil ver por qué los supuestos anteriores son suficientes para que funcione la solución suplente. Si en la ecuación (9.10) se sustituye con la (9.11) y se hacen las sencillas manipulaciones algebraicas necesarias, se obtiene

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3.$$

Al error compuesto en esta ecuación llámesele $e = u + \beta_3 v_3$; este error depende del error en el modelo que interesa, (9.10), y del error en la ecuación de la variable proxy, v_3. Como u y

v_3 tienen ambas media cero y ninguna está correlacionada con x_1, x_2 y x_3, e también tiene media cero y no está correlacionada con x_1, x_2 y x_3. Escriba esta ecuación como

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

donde $\alpha_0 = (\beta_0 + \beta_3\delta_0)$ es el nuevo intercepto y $\alpha_3 = \beta_3\delta_3$ es el parámetro de pendiente de la variable proxy x_3. Como se dijo antes, cuando se corre la regresión en (9.12) no se obtienen estimadores insesgados de β_0 y β_3; lo que se obtiene son estimadores insesgados (o al menos consistentes) de α_0, β_1, β_2 y α_3. Lo importante es que se obtengan buenas estimaciones de los parámetros β_1 y β_2.

En la mayoría de los casos, la α_3 estimada es en realidad más interesante que la β_3 estimada. Por ejemplo, en la ecuación sobre el salario, α_3 mide el rendimiento en el salario por un punto más en la puntuación del *IQ*.

Ejemplo 9.3

[*IQ* como variable proxy para capacidad]

El archivo WAGE2.RAW, de Blackburn y Neumark (1992), contiene información sobre remuneraciones mensuales, educación, varias variables demográficas y puntuaciones de *IQ* de 935 hombres en 1980. Como método para explicar el sesgo por omitir capacidad, se agrega el *IQ* a una ecuación estándar para logaritmo de salario. Los resultados se muestran en la tabla 9.2.

El principal interés se refiere a lo que ocurre con el rendimiento estimado de la educación. En la columna (1) se presentan las estimaciones sin usar *IQ* como variable proxy. El rendimiento estimado de la educación es 6.5%. Si se considera que la capacidad que se ha omitido está relacionada positivamente con *educ*, entonces se supone que esta estimación es demasiado alta. (De manera más precisa, el promedio de estimaciones de todas las muestras aleatorias será demasiado alto.) Cuando a la ecuación se le agrega el *IQ*, el rendimiento de la educación disminuye a 5.4%, lo que corresponde a la creencia anterior acerca del sesgo por la omisión de la capacidad.

El efecto del *IQ* en los resultados socioeconómicos ha sido documentado en el controvertido libro *La curva de campana* de Herrnstein y Murray (1994). En la columna (2) se muestra que el *IQ* tiene un efecto positivo estadísticamente significativo sobre las remuneraciones una vez controlados otros factores. Permaneciendo todo lo demás constante, se predice que un aumento de 10 puntos en el *IQ* incremente las remuneraciones mensuales 3.6%. En la población de Estados Unidos, la desviación del *IQ* es 15, por lo que un incremento de una desviación estándar en el *IQ* corresponde a ganancias 5.4% mayores. Este incremento es idéntico al que se predice en el salario por un año más de educación. De acuerdo con la columna (2) es claro que la educación sigue teniendo un papel importante para el incremento de las remuneraciones, aun cuando este efecto no sea tan grande como originalmente se estimó.

En las columnas (1) y (2) pueden hacerse otras observaciones interesantes. Agregar *IQ* a la ecuación sólo hace que la *R*-cuadrada se incremente de .253 a .263. La mayor parte de la variación en log(*wage*) no es explicada por los factores que aparecen en la columna (2). La adición de *IQ* a la ecuación tampoco elimina la diferencia entre blancos y negros en las remuneraciones estimadas: se predice que un negro que tenga el mismo *IQ*, la misma educación, la misma experiencia, etc., que un blanco, gane cerca de 14.3% menos, y esta diferencia es estadísticamente muy significativa.

En la columna (3) de la tabla 9.2 se incluye el término de interacción *educ*·*IQ*. Esto permite que *educ* y *abil* interactúen en la determinación de log(*wage*). Se podría pensar que el rendimiento de la educación fuera mayor en el caso de personas con

Pregunta 9.2

¿Qué piensa del pequeño y estadísticamente no significativo coeficiente de *educ* en la columna (3) de la tabla 9.2? (*Sugerencia:* cuando en la ecuación aparece *educ*·*IQ*, ¿cuál es la interpretación del coeficiente de *educ*?)

TABLA 9.2

Variable dependiente: log(*wage*)

Variables independientes	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	–.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	–.188 (.038)	–.143 (.039)	−.147 (.040)
IQ	—	.0036 (.0010)	−.0009 (.0052)
educ·IQ	—	—	.00034 (.00038)
intercepto	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observaciones *R*-cuadrada	935 .253	935 .263	935 .263

mayor capacidad, pero esto resulta no ser el caso: el término de interacción no es significativo y hace a *educ* y a *IQ* individualmente no significativas complicando al mismo tiempo el método. Por tanto, se prefieren las estimaciones de la columna (2).

No hay razón para quedarse con una sola variable proxy en este ejemplo. La base de datos WAGE2. RAW contiene también una puntuación sobre la prueba *Knowledge of the World of Work* (KWW) (Conocimiento del mundo del trabajo) de cada individuo. Esto proporciona otra medida de la capacidad, que puede usarse en lugar del *IQ* o junto con el *IQ* para estimar el rendimiento de la educación (vea el ejercicio para computadora C9.2).

Es fácil ver cómo el uso de una variable proxy puede conducir a sesgo si no satisface los supuestos anteriores. Suponga que, en lugar de (9.11), la variable proxy no observada, x_3^*, está relacionada con todas las variables observadas mediante

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3, \tag{9.14}$$

donde v_3 tiene media cero y no está correlacionada con x_1, x_2 y x_3. En la ecuación (9.11) se supone que tanto δ_1 como δ_2 son cero. Sustituyendo la ecuación (9.14) en la (9.10), se obtiene

$$\begin{aligned} y = {} & (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 \\ & + \beta_3\delta_3 x_3 + u + \beta_3 v_3, \end{aligned} \tag{9.15}$$

de acuerdo con la cual se concluye que $\text{plim}(\hat{\beta}_1) = \beta_1 + \beta_3\delta_1$ y que $\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_3\delta_2$. [Esto se sigue debido a que el error en (9.15), $u + \beta_3 v_3$, tiene media cero y no está correlacionado con x_1, x_2 y x_3.] En el ejemplo anterior en el que $x_1 = educ$ y $x_3^* = abil$, $\beta_3 > 0$, de manera que si *abil* tiene una correlación parcial positiva con *educ* ($\delta_1 > 0$) existirá un sesgo (inconsistencia) positivo. Por tanto, si *IQ* no es una buena proxy, podría obtenerse un sesgo hacia arriba usando *IQ* como proxy para *abil*. Pero es razonable esperar que este sesgo sea menor que si se ignorara por completo el problema de omitir la capacidad.

Las variables proxy también pueden presentarse en forma de información binaria. En el ejemplo 7.9 [vea la ecuación (7.15)] se vieron las estimaciones de Krueger (1993) sobre el rendimiento de emplear una computadora en el trabajo. Krueger también incluyó una variable binaria para indicar si el trabajador emplea computadora en casa (así como un término de interacción entre el uso de computadora en el trabajo y el uso en casa). Su principal razón para incluir en la ecuación el uso de computadora en casa fue emplear una proxy para la variable no observada "destreza técnica" que podía afectar directamente al salario y estar relacionada con el uso de computadora en el trabajo.

Utilización de variables dependientes rezagadas como variables proxy

En algunas aplicaciones, como en el ejemplo anterior sobre el salario, se tiene al menos una idea vaga de cuáles son los factores no observados que se desearía controlar. Esto facilita la elección de las variables proxy. En otras aplicaciones se sospecha que una o más variables independientes estén correlacionadas con una variable omitida, pero no se tiene idea de cómo obtener una proxy para esa variable omitida. En tales casos, puede incluirse como control el valor de la variable dependiente en un periodo anterior. Esto es especialmente útil en análisis de políticas.

El uso de una **variable dependiente rezagada** en una ecuación con datos de corte transversal aumenta el requerimiento de datos, pero proporciona también una manera sencilla de dar cuenta de factores históricos que ocasionan diferencias *actuales* en la variable dependiente que son difíciles de explicar de otra manera. Por ejemplo, algunas ciudades han tenido altos índices de delincuencia en el pasado. Muchos de los mismos factores no observados contribuyen tanto a los índices de criminalidad en el presente como en el pasado. De la misma manera, algunas universidades son mejores académicamente que otras. Factores de inercia se captan empleando rezagos en y.

Considere una ecuación sencilla para explicar los índices de criminalidad urbanos:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{-1} + u, \tag{9.16}$$

donde *crime* es una medida de la delincuencia *per cápita*, *unem* es el índice de desempleo urbano, *expend* es el gasto *per cápita* en seguridad pública y $crime_{-1}$ es el índice de delincuencia en alguno de los años anteriores (puede ser el año anterior o varios años anteriores). Lo que interesa es el efecto de *unem* sobre *crime*, así como el gasto en seguridad pública sobre *crime*.

¿Cuál es el objeto de incluir en la ecuación $crime_{-1}$? Se espera que $\beta_3 > 0$ porque la delincuencia tiene inercia. Pero la principal razón para incluir esta variable en la ecuación es que quizá las ciudades con altos índices históricos de delincuencia gasten más en prevención de la misma. Por tanto, es posible que factores no observados por nosotros (los econometristas) que afectan a *crime* estén correlacionados con *expend* (y *unem*). Si se emplea un análisis de corte transversal puro, es poco probable que se obtenga un estimador insesgado del efecto del gasto en seguridad pública sobre la delincuencia. Pero incluyendo $crime_{-1}$ en la ecuación, al menos se puede hacer el siguiente experimento: si dos ciudades tienen el mismo índice de delincuencia previo y el mismo índice de desempleo, entonces β_2 mide el efecto de un dólar más en seguridad pública sobre la delincuencia.

Ejemplo 9.4

[Índice de delincuencia urbana]

Se estima una versión de elasticidad constante del modelo de delincuencia dado en la ecuación (9.16) (*unem*, se deja de forma lineal debido a que es un porcentaje). Los datos en el archivo CRIME2.RAW son del año 1987 y corresponden a 46 ciudades. Se cuenta también con el índice de delincuencia de 1982 y se usa como variable independiente adicional para tratar de controlar los no observables urbanos que afectan la delincuencia y que pueden estar correlacionados con el gasto presente en seguridad pública. Los resultados se muestran en la tabla 9.3.

Sin el índice rezagado de delincuencia, los efectos del índice de desempleo y del gasto en seguridad pública son contraintuitivos; ninguno es estadísticamente significativo, aunque el estadístico *t* para $\log(lawexpc_{87})$ es 1.17. Una posibilidad es que el aumento del gasto en seguridad pública mejore los

TABLA 9.3

Variable dependiente: $\log(crmrte_{87})$

Variables independientes	(1)	(2)
$unem_{87}$	−.029 (.032)	.009 (.020)
$\log(lawexpc_{87})$	.203 (.173)	−.140 (.109)
$\log(crmrte_{82})$	—	1.194 (.132)
intercepto	3.34 (1.25)	.076 (.821)
Observaciones *R*-cuadrada	46 .057	46 .680

métodos de denuncia y haga que más delitos sean *reportados*. Pero también puede ser que las ciudades con altas tasas recientes de criminalidad gasten más en seguridad pública.

Agregar el logaritmo del índice de delincuencia de cinco años atrás tiene un gran efecto sobre el coeficiente del gasto. La elasticidad del índice de delincuencia respecto al gasto se vuelve $-.14$, con $t = -1.28$. Esto no es fuertemente significativo, pero sugiere que un modelo más sofisticado que tenga más ciudades en la muestra podría producir resultados significativos.

Como era de esperarse, el índice de delincuencia actual está fuertemente relacionado con el índice de delincuencia del pasado. El coeficiente estimado indica que si el índice de delincuencia en 1982 fue 1% más alto, se predice que el índice de delincuencia en 1987 sea aproximadamente 1.19% más alto. No se puede rechazar la hipótesis de que la elasticidad de la delincuencia actual respecto a la delincuencia en el pasado sea igual a uno [$t = (1.194 - 1)/.132 \approx 1.47$]. Agregar el índice de delincuencia del pasado aumenta marcadamente la potencia explicativa de la regresión, pero esto era de esperarse. La principal razón para incluir el índice de delincuencia rezagado es obtener una mejor estimación del efecto *ceteris paribus* de $\log(lawexpc_{87})$ sobre $\log(crmrte_{87})$.

La práctica de emplear una y rezagada como medio general para controlar variables no observadas dista de ser perfecta. Pero puede ayudar a obtener una mejor estimación de los efectos de variables de políticas sobre diversos resultados.

Agregar un valor rezagado de y no es la única manera de usar datos de dos años para controlar factores omitidos. Cuando se estudien los métodos para datos de panel, en los capítulos 13 y 14, se verán varias maneras de usar datos repetidos en las mismas unidades de corte transversal en distintos puntos en el tiempo.

Un enfoque diferente de la regresión múltiple

El estudio de las variables proxy en esta sección sugiere una manera alterna de interpretar un análisis de regresión múltiple cuando no se observan todas las variables explicativas relevantes. Hasta ahora, el modelo poblacional de interés se ha especificado con un error aditivo, como en la ecuación (9.9). El análisis en ese ejemplo dependía de si se tenía una variable proxy adecuada (la puntuación correspondiente al *IQ* en este caso, u otras puntuaciones de manera más general) para la variable explicativa no observada, a la que se le llamó "capacidad".

Una manera menos estructurada y más general de la regresión múltiple es renunciar a especificar modelos con no observables. Antes bien, se parte de la premisa de que se tiene acceso a un conjunto de variables explicativas observables —que comprenden la variable de mayor interés, como años de escolaridad, y los controles, como puntuaciones de examen observables. Con esto se modela la media condicionada de y sobre las variables explicativas observadas. Por ejemplo, en el caso del salario, denotando *lwage* como log(*wage*), puede estimarse $E(lwage|educ,exper,tenure,south,urban,black,IQ)$— exactamente lo que se reporta en la tabla 9.2. La diferencia es que ahora los objetivos se fijan de una manera más modesta. Es decir, en lugar de introducir en la ecuación (9.9) el nebuloso concepto de "la capacidad", desde un principio se establece que se va a estimar el efecto *ceteris paribus* de la educación, manteniendo constantes el *IQ* (y los demás factores observables). No es necesario analizar si el *IQ* es una variable proxy adecuada para capacidad. Por tanto, aunque puede que no se esté dando respuesta a la pregunta que subyace a la ecuación (9.9), se está dando respuesta a una cuestión de interés: si dos personas tienen el mismo *IQ* (y los mismos valores para experiencia, antigüedad, etc.), pero su nivel de educación difiere en un año, ¿cuál es la diferencia que se espera entre los logaritmos de sus salarios?

Como otro ejemplo, si en una regresión a nivel escolar se incluye como variable explicativa el índice de pobreza para evaluar los efectos de los gastos sobre las puntuaciones en exámenes estandarizados, habrá que reconocer que el índice de pobreza únicamente capta de una manera cruda las diferencias relevantes entre los niños y padres de las distintas escuelas. Pero, a menudo, eso es todo lo que se tiene y es mejor controlar el índice de pobreza que no hacer nada porque no se puedan hallar proxys para la "capacidad" de los estudiantes, la "participación" de los padres, etc. Es casi seguro que controlar el índice de pobreza permita acercarse más a los efectos *ceteris paribus* del gasto que si se deja dicho índice fuera del análisis.

En algunas aplicaciones del análisis de regresión, lo único que interesa es predecir un resultado y, dado un conjunto de variables explicativas $(x_1, ..., x_k)$. En esos casos no tiene mucho sentido pensar en términos del "sesgo" en los coeficientes estimados a causa de las variables omitidas. Más bien habrá que preocuparse por obtener un modelo que prediga tan bien como sea posible y tener cuidado de no incluir como regresores variables que no puedan ser observados en el momento de la predicción. Por ejemplo, el responsable del departamento de admisión de una universidad está interesado en predecir el éxito, medido en términos del promedio general de calificaciones, que tendrán los estudiantes en los estudios universitarios en términos de variables que puedan medirse al momento de la inscripción del alumno. Estas variables pueden ser desempeño en el bachillerato (tal vez sólo el promedio general o las calificaciones en determinadas materias), calificaciones en exámenes estandarizados, participación en determinadas actividades (como en declamación o en concursos de matemáticas) e incluso variables sobre antecedentes familiares. No se incluirán variables como asistencia a clases en la universidad, porque esta es una variable que aún no se ha observado al momento de la inscripción. Ni será motivo de preocupación el potencial "sesgo" debido a que no se mida una variable de asistencia a clases: no interesa medir el efecto del promedio general de calificaciones manteniendo constante la asistencia a clases en la universidad. De igual manera, no será motivo de preocupación el sesgo en los coeficientes debido a no poder observar factores tales como la motivación. Naturalmente, para fines de predicción, tal vez ayudaría bastante contar con una medida de la motivación, pero en ausencia de esta variable se ajustará el mejor modelo posible que se obtenga dadas las variables explicativas observadas.

9.3 Modelos con pendientes aleatorias

En el estudio hasta ahora realizado de la regresión, se ha supuesto que los coeficientes de pendiente son los mismos para todos los individuos de la población, o que, si las pendientes difieren, lo harán en características medibles, lo cual lleva a los modelos de regresión con términos de interacción. Por ejemplo, como se vio en la sección 7.4, pueden considerarse las diferencias en el rendimiento a la educación entre hombres y mujeres haciendo que la educación interactúe con una variable binaria de género en una ecuación para el logaritmo del salario.

Aquí la pregunta de interés es parecida pero diferente: ¿Qué pasa si el efecto parcial de una variable depende de factores no observados que varían entre las unidades de la población? Teniendo sólo una variable explicativa, x, el modelo general (para un elemento, i, tomado en forma aleatoria de la población, para hacer hincapié) es

$$y_i = a_i + b_i x_i, \tag{9.17}$$

donde a_i es el intercepto para la unidad i y b_i es la pendiente. En el modelo de regresión simple, visto en el capítulo 2, se supuso que $b_i = \beta$ y a_i se consideró como el error, u_i. Al modelo (9.17) suele llamársele **modelo de coeficiente aleatorio** o **modelo de pendiente aleatoria** porque el coeficiente de pendiente no observado, b_i, se considera como una observación muestreada aleatoriamente de la población, al igual que los datos observados (x_i, y_i), y el intercepto no observado, a_i. Por ejemplo, si $y_i = \log(wage_i)$ y $x_i = educ_i$, entonces (9.17) considera que el rendimiento de la educación, b_i varía de acuerdo con la persona. Si, por ejemplo, en b_i está contenida la capacidad no medida (así como en a_i), el efecto parcial de un año más de escolaridad puede depender de la capacidad.

Cuando se tiene una muestra aleatoria de tamaño n, se muestrean (implícitamente) n valores b_i y n valores para a_i (así como los datos observados para x y y). Naturalmente, no se puede estimar una pendiente —o, en realidad, un intercepto— para cada i. Pero se puede esperar estimar la pendiente promedio (y el intercepto promedio), donde el promedio es sobre toda la población. Por tanto, se define $\alpha = \mathrm{E}(a_i)$ y $\beta = \mathrm{E}(b_i)$. Entonces β es el promedio del efecto parcial de x sobre y, y por esto a β se le llama **efecto parcial promedio (EPP)**, o *efecto marginal promedio (EMP)*. En el contexto de una ecuación de logaritmos del salario, β es el rendimiento promedio, en la población, de un año de escolaridad.

Si se escribe $a_i = \alpha + c_i$ y $b_i = \beta + d_i$, entonces d_i es la desviación individual específica del EPP. Por construcción, $\mathrm{E}(c_i) = 0$ y $\mathrm{E}(d_i) = 0$. Sustituyendo en (9.17) se obtiene

$$y_i = \alpha + \beta x_i + c_i + d_i x_i \equiv \alpha + \beta x_i + u_i, \qquad \textbf{9.18}$$

donde $u_i = c_i + d_i x_i$. (Para hacer esta notación más fácil de entender, ahora se usará α, el valor medio de las a_i, como el intercepto, y β, la media de las b_i, como la pendiente.) En otras palabras, el modelo de coeficientes aleatorios puede escribirse como un modelo de coeficientes constantes, pero donde el término del error contiene una interacción entre un no observable, d_i, y la variable explicativa observada, x_i.

¿Cuándo producirá una regresión simple de y_i sobre x_i una estimación insesgada de β (y de α)? Se pueden emplear los resultados para el insesgamiento del capítulo 2. Si $\mathrm{E}(u_i|x_i) = 0$, entonces MCO es generalmente insesgado. Cuando $u_i = c_i + d_i x_i$, es suficiente que $\mathrm{E}(c_i|x_i) = \mathrm{E}(c_i) = 0$ y que $\mathrm{E}(d_i|x_i) = \mathrm{E}(d_i) = 0$. Esto se puede expresar en términos del intercepto y de la pendiente específicos a cada observación como

$$\mathrm{E}(a_i|x_i) = \mathrm{E}(a_i) \quad \text{y} \quad \mathrm{E}(b_i|x_i) = \mathrm{E}(b_i); \qquad \textbf{9.19}$$

es decir, tanto a_i como b_i son media independientes de x_i. Este es un hallazgo útil: si se consideran pendientes específicos a cada observación, MCO estima consistentemente el promedio poblacional de esas pendientes cuando sean media independientes de la variable explicativa. (Vea en el problema 9.6 condiciones más débiles que implican la consistencia de MCO.)

Es casi seguro que el término del error en (9.18) contenga heterocedasticidad. En realidad, si $\mathrm{Var}(c_i|x_i) = \sigma_c^2$, $\mathrm{Var}(d_i|x_i) = \sigma_d^2$, y $\mathrm{Cov}(c_i, d_i|x_i) = 0$, entonces

$$\mathrm{Var}(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2, \qquad \textbf{9.20}$$

y por tanto deberá haber heterocedasticidad en u_i a menos que $\sigma_d^2 = 0$, lo que significa que $b_i = \beta$ para toda i. Se sabe qué hacer con la heterocedasticidad de este tipo. Se puede usar MCO y calcular errores estándar y estadísticos de prueba robustos a la heterocedasticidad, o se puede estimar la función de varianza en (9.20) y emplear mínimos cuadrados ponderados. Por supuesto que la última estrategia impone homocedasticidad al intercepto y a la pendiente aleatoria, y de esta manera se deseará hacer un análisis de MCP completamente robusto a las violaciones de (9.20).

Debido a la ecuación (9.20), a algunos autores les gusta considerar la heterocedasticidad de los modelos de regresión como resultado de los coeficientes de la pendiente aleatoria. Pero hay que recordar que la forma de (9.20) es especial y no considera heterocedasticidad en a_i ni en b_i. No se puede distinguir de manera convincente entre un modelo de pendiente aleatoria, donde el intercepto y la pendiente son independientes de x_i, de un modelo de pendiente constante con heterocedasticidad en a_i.

El tratamiento en el caso de la regresión múltiple es similar. En general, se escribe

$$y_i = a_i + b_{i1}x_{i1} + b_{i2}x_{i2} \ldots + b_{ik}x_{ik}. \qquad \textbf{9.21}$$

Entonces, escribiendo $a_i = \alpha + c_i$ y $b_{ij} = \beta_j + d_{ij}$, se tiene

$$y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i, \tag{9.22}$$

donde $u_i = c_i + d_{i1}x_{i1} + \dots + d_{ik}x_{ik}$. Si se mantienen los supuestos de media independiente $E(a_i|\mathbf{x}_i) = E(a_i)$ y $E(b_{ij}|\mathbf{x}_i) = E(b_{ij})$, $j = 1, \dots, k$, entonces $E(y_i|\mathbf{x}_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, y de esta manera usando una muestra aleatoria MCO produce estimadores insesgados de α y de las β_j. Como en el caso de la regresión simple, $\mathrm{Var}(u_i|\mathbf{x}_i)$ es casi seguro que sea heterocedástica.

Se puede considerar que las b_{ij} dependan de variables explicativas observables así como de no observables. Por ejemplo, suponga que para $k = 2$ el efecto de x_{i2} dependa de x_{i1}, y que se escriba $b_{i2} = \beta_2 + \delta_1(x_{i1} - \mu_1) + d_{i2}$, donde $\mu_1 = E(x_{i1})$. Si se supone que $E(d_{i2}|\mathbf{x}_i) = 0$ (y de manera similar para c_i y d_{i1}), entonces $E(y_i|x_{i1}, x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1(x_{i1} - \mu_1)x_{i2}$, lo que significa que se tiene una interacción entre x_{i1} y x_{i2}. Como de μ_1 se ha restado la media x_{i1}, β_2 es el efecto parcial promedio de x_{i2}.

Lo esencial de esta sección es que considerar pendientes aleatorias es bastante sencillo si las pendientes son independientes, o al menos media independientes, de las variables explicativas. Además, es fácil modelar las pendientes como funciones de las variables exógenas, lo que conduce a modelos con cuadrados e interacciones. Por supuesto, en el capítulo 6, sin haber introducido la noción de una pendiente aleatoria, se analizó cómo pueden ser útiles estos modelos. La especificación de pendientes aleatorias proporciona una justificación distinta para tales modelos. La estimación se vuelve considerablemente más difícil si el intercepto aleatorio, así como algunas pendientes están relacionadas con algunos de los regresores. El problema de variables explicativas endógenas se verá en el capítulo 15.

9.4 Propiedades de MCO bajo error de medición

Algunas veces, en las aplicaciones económicas, no es posible recolectar datos sobre la variable que realmente afecta el comportamiento económico. Un buen ejemplo es la tasa marginal de impuesto al ingreso que enfrenta una familia al tratar de elegir con cuánto contribuir a obras de caridad. La tasa marginal suele ser difícil de obtener o de resumir en un solo número para todos los niveles de ingreso. En cambio, sí se puede calcular la tasa de impuesto promedio basándose en el ingreso total y en los pagos de impuestos.

Cuando en un modelo de regresión se emplea una medición no precisa de una variable económica, el modelo contiene un error de medición. En esta sección se hallarán las consecuencias de un error de medición en la estimación de mínimos cuadrados ordinarios. MCO será consistente bajo ciertos supuestos, pero existen otros bajo los cuales no lo es. En algunos de estos casos, se puede obtener la magnitud del sesgo asintótico.

Como se verá, el problema del error de medición tiene una estructura estadística similar a la del problema de la variable omitida-variable proxy visto en la sección anterior, sin embargo, son conceptualmente diferentes. En el caso de la variable proxy, se busca una variable que esté de alguna manera relacionada con la variable no observada. En el caso del error de medición, la variable que no se observa tiene un significado cuantitativo bien definido (por ejemplo, tasa de ingreso marginal o ingreso anual), pero la medida que se ha registrado de esa variable puede contener error. Por ejemplo, el ingreso anual reportado es una medida del ingreso anual real, mientras que la puntuación en el *IQ* es una proxy para capacidad.

Otra diferencia importante entre una variable proxy y los problemas de error de medición es que en el último caso con frecuencia la variable independiente mal medida es la de mayor interés. En el caso de la variable proxy, el efecto parcial de la variable omitida raramente es de interés central: en general nos preocupan los efectos de las otras variables independientes.

Antes de entrar en detalles, hay que recordar que el error de medición constituye un problema sólo cuando las variables para las cuales los dedicados a la econometría pueden recolectar datos difieren de las variables que influyen en las decisiones de los individuos, las familias, las empresas, etcétera.

Error de medición en la variable dependiente

Se empieza con el caso en el que sólo la variable dependiente se mide con error. Sea y^* la variable (en la población, como siempre) que se desea explicar. Por ejemplo, y^* puede ser el ahorro familiar anual. El modelo de regresión tiene la forma usual

$$y^* = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u, \tag{9.23}$$

y se supone que satisface los supuestos de Gauss-Markov. Sea y la medición observable de y^*. En el caso del ahorro, y es ahorro anual reportado. Por desgracia, las familias no reportan perfectamente su ahorro familiar anual; es fácil dejar algunas categorías fuera o sobrestimar el monto con el que se contribuyó al ahorro. Por lo general, puede esperarse que y y y^* difieran, al menos en algún subconjunto de familias de la población.

El error de medición (en la población) está definido como la diferencia entre el valor observado y el valor real:

$$e_0 = y - y^*. \tag{9.24}$$

En el caso de una observación i muestreada aleatoriamente de la población, se puede escribir $e_{i0} = y_i - y_i^*$, pero lo importante es cómo está relacionado el error de medición en la población con otros factores. Para obtener un modelo estimable, se escribe $y^* = y - e_0$, esto se sustituye en la ecuación (9.23) y se reordena:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u + e_0. \tag{9.25}$$

En la ecuación (9.25), el término del error es $u + e_0$. Como $y, x_1, x_2, \dots, x_k$ son observados, este modelo puede estimarse por MCO. En efecto, simplemente se ignora el hecho de que y sea una medición imperfecta de y^* y se procede como de costumbre.

¿Cuándo produce MCO con y en lugar de y^* estimadores consistentes de las β_j? Como el modelo original (9.23) satisface los supuestos de Gauss-Markov, u tiene media cero y no está correlacionada con ninguna de las x_j. Resulta natural suponer que el error de medición tenga media cero; si no es así, simplemente se obtiene un estimador sesgado del intercepto, β_0, lo que rara vez es causa de preocupación. De mucha más importancia es el supuesto acerca de la relación entre el error de medición, e_0, y las variables explicativas, x_j. El supuesto usual es que el error de medición en y es estadísticamente independiente de cada una de las variables explicativas. Si esto es así, entonces los estimadores de MCO de (9.25) son insesgados y consistentes. Además, los procedimientos usuales de inferencia de MCO (estadísticos t, F y ML) son válidos.

Si e_0 y u no están correlacionados, como se supone usualmente, entonces $\text{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$. Esto significa que el error de medición en la variable dependiente da como resultado una mayor varianza del error que cuando no ocurre ningún error de medición; esto, por supuesto, da como resultado varianzas mayores de los estimadores de MCO. Esto es de esperarse, y no hay nada que se pueda hacer (salvo recabar datos mejores). Lo importante es que, si el error de medición no está correlacionado con las variables independientes, entonces la estimación por MCO tiene propiedades buenas.

Ejemplo 9.5

[Función de ahorro con error de medición]

Considere una función de ahorro

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age + u,$$

pero en la que en el ahorro real (sav^*) puede ser distinto del ahorro reportado (sav). La cuestión es si la magnitud del error de medición en sav está sistemáticamente relacionada con las otras variables. Parece razonable suponer que el error de medición no está correlacionado con *inc* (ingreso), *size* (tamaño del hogar), *educ* y *age* (edad). Por otro lado, se puede pensar que las familias con mayores ingresos o más educación reporten con mayor exactitud sus ahorros. Nunca se podrá saber si el error de medición está correlacionado con *inc* o con *educ*, a menos que se puedan recolectar datos sobre sav^*; entonces puede calcularse el error de medición en cada observación como $e_{i0} = sav_i - sav_i^*$.

Cuando la variable dependiente está en forma logarítmica, de manera que la variable dependiente es $\log(y^*)$, es natural que la ecuación del error de medición tenga la forma

$$\log(y) = \log(y^*) + e_0. \tag{9.26}$$

Esto sigue de un **error de medición multiplicativo** para y: $y = y^*a_0$, donde $a_0 > 0$ y $e_0 = \log(a_0)$.

Ejemplo 9.6

[Error de medición en las tasas de desperdicio]

En la sección 7.6 se analizó un ejemplo en el que se deseaba determinar si una subvención para la capacitación en el trabajo reducía la tasa de desperdicio en las fábricas. Se puede pensar que es muy probable que las tasas de desperdicio que reportan las empresas estén medidas con error. (En realidad, la mayoría de las empresas de la muestra no reportan tasa de desperdicio.) En un marco de una regresión simple esto se expresa mediante

$$\log(scrap^*) = \beta_0 + \beta_1 grant + u,$$

donde $scrap^*$ es la verdadera tasa de desperdicio y *grant* es una variable binaria que indica si una empresa recibió subvención. La ecuación del error de medición es

$$\log(scrap) = \log(scrap^*) + e_0.$$

¿Es el error de medición, e_0, independiente de si la empresa recibió subvención? Se puede pensar que es más probable que la empresa que recibe una subvención reporte una tasa de desperdicio menor a la real con objeto de que parezca que la subvención tuvo efecto. Si pasa esto, entonces, en la ecuación estimable,

$$\log(scrap) = \beta_0 + \beta_1 grant + u + e_0,$$

el error $u + e_0$ está correlacionado negativamente con *grant.* Esto producirá un sesgo hacia abajo en β_1, lo cual tenderá a hacer que el programa de capacitación parezca más eficiente de lo que en realidad es. (Recuerde que una β_1 más negativa significa que el programa fue más eficiente, dado que una mayor productividad de los trabajadores está relacionada con una menor tasa de desperdicio.)

La conclusión de esta subsección es que un error de medición en la variable dependiente *puede* causar sesgo en MCO en el caso de que esté relacionado sistemáticamente con una o más de las variables explicativas. Si el error de medición es sólo un error aleatorio asociado al reporte de los datos que es independiente de las variables explicativas, como suele suponerse, entonces MCO es perfectamente apropiado.

Error de medición en las variables explicativas

Tradicionalmente se ha considerado que un error de medición en una variable explicativa es un problema mucho más importante que un error de medición en una variable dependiente. En esta subsección se verá a qué se debe esto.

Se parte del modelo de regresión simple

$$y = \beta_0 + \beta_1 x_1^* + u, \tag{9.27}$$

y se supone que este modelo satisface al menos los cuatro primeros supuestos de Gauss-Markov. Esto significa que la estimación por MCO de (9.27) dará estimadores insesgados y consistentes de β_0 y β_1. El problema es que x_1^* no es observada. En cambio, se tiene una medición de x_1^*; a la que se le llamará x_1. Así, por ejemplo, x_1^* puede ser el ingreso real y x_1 el ingreso reportado.

El error de medición en la población es simplemente

$$e_1 = x_1 - x_1^*, \tag{9.28}$$

y puede ser positivo, negativo o cero. Se supone que en la población el error de medición *promedio* es cero: $\mathrm{E}(e_1) = 0$. Esto es natural y, en todo caso, no afecta la importante conclusión que se obtendrá. Un supuesto que se conserva en lo que sigue es que u no está correlacionada con x_1^* *ni* con x_1. En términos de la esperanza condicional, esto se puede expresar como $\mathrm{E}(y|x_1^*, x_1) = \mathrm{E}(y|x_1^*)$, lo que sólo dice que x_1 no afecta a y una vez que se ha controlado x_1^*. El mismo supuesto se empleó en el caso de la variable proxy, y este supuesto no es controversial; se sigue casi por definición.

Se desea conocer las propiedades de MCO cuando simplemente se sustituye x_1^* por x_1 y se corre la regresión de y sobre x_1. Estas propiedades dependen de manera crucial de los supuestos hechos acerca del error de medición. Dos supuestos han sido el centro en la literatura econométrica, y representan extremos polarmente opuestos. El primer supuesto es que e_1 no está correlacionado con la medición *observada*, x_1:

$$\mathrm{Cov}(x_1,e_1) = 0. \tag{9.29}$$

De acuerdo con la relación en (9.28), si el supuesto (9.29) es verdadero, entonces e_1 debe estar correlacionado con la variable no observada x_1^*. Para determinar las propiedades de MCO en este caso, se escribe $x_1^* = x_1 - e_1$ y esto se sustituye en la ecuación (9.27):

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1). \tag{9.30}$$

Como se ha supuesto que tanto u como e_1 tienen media cero y no están correlacionados con x_1, $u - \beta_1 e_1$ tiene media cero y no está correlacionado con x_1. Se sigue que la estimación de MCO con x_1 en lugar de x_1^* produce un estimador consistente de β_1 (y también de β_0). Como u no está correlacionada con e_1, la varianza del error en (9.30) es $\mathrm{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2\sigma_{e_1}^2$. Por tanto, salvo cuando $\beta_1 = 0$, el error de medición hace que aumente la varianza del error. Pero esto no afecta a ninguna de las propiedades de MCO (salvo que las varianzas de las $\hat{\beta}_j$ serán mayores que si se observara directamente x_1^*).

El supuesto de que e_1 no está correlacionado con x_1 es análogo al supuesto que se hizo para la variable proxy en la sección 9.2. Como este supuesto implica que MCO tiene todas sus propiedades agradables, esto no suele ser lo que los econometristas tienen en mente al referirse al error de medición en una variable explicativa. El supuesto de **errores clásicos en las variables (ECV)** es que el error de medición no está correlacionado con la variable explicativa *no observada*:

$$\mathrm{Cov}(x_1^*, e_1) = 0. \tag{9.31}$$

Este supuesto proviene de expresar la medición observada como la suma de la verdadera variable explicativa y el error de medición,

$$x_1 = x_1^* + e_1,$$

y suponer después que los dos componentes de x_1 no están correlacionados. (Esto no tiene nada que ver con los supuestos acerca de u; siempre se sostiene que u no está correlacionada con x_1^* ni con x_1 y, por tanto, tampoco con e_1.)

Si se satisface el supuesto (9.31), entonces x_1 y e_1 *deben* estar correlacionadas:

$$\mathrm{Cov}(x_1, e_1) = \mathrm{E}(x_1 e_1) = \mathrm{E}(x_1^* e_1) + \mathrm{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2. \tag{9.32}$$

Por tanto, bajo el supuesto ECV la covarianza entre x_1 y e_1 es igual a la varianza del error de medición.

Volviendo a la ecuación (9.30), se puede ver que la correlación entre x_1 y e_1 va a ocasionar problemas. Dado que u y x_1 no están correlacionados, la covarianza entre x_1 y el error compuesto $u - \beta_1 e_1$ es

$$\mathrm{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \mathrm{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2.$$

Por tanto, en el caso de ECV, la regresión por MCO de y sobre x_1 da un estimador sesgado e inconsistente.

Empleando el resultado asintótico del capítulo 5, puede determinarse la magnitud de la inconsistencia de MCO. El límite de probabilidad de $\hat{\beta}_1$ es β_1 más el cociente de la covarianza entre x_1 y $u - \beta_1 e_1$ y la varianza de x_1:

$$\begin{aligned}
\mathrm{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\mathrm{Cov}(x_1, u - \beta_1 e_1)}{\mathrm{Var}(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\
&= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) \\
&= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right),
\end{aligned} \tag{9.33}$$

donde se ha usado el hecho de que $\mathrm{Var}(x_1) = \mathrm{Var}(x_1^*) + \mathrm{Var}(e_1)$.

La ecuación (9.33) es muy interesante. El término que multiplica a β_1, que es el cociente Var (x_1^*)/Var(x_1), es siempre menor a uno [consecuencia del supuesto (9.31) de ECV]. Por tanto, el plim$(\hat{\beta}_1)$ se encuentra más cercano a cero que β_1. A esto se le llama el **sesgo de atenuación** en MCO debido a los errores clásicos en las variables: en promedio (o en muestras grandes), el efecto en los estimados de MCO será *atenuado*. En particular, si β_1 es positivo, $\hat{\beta}_1$ tenderá a subestimarlo. Esta es una conclusión importante, pero que proviene del supuesto ECV.

Si la varianza de x_1^* es grande en relación con la varianza del error de medición, entonces la inconsistencia en MCO será pequeña. Esto se debe a que Var(x_1^*)/Var(x_1) tendrá un valor cercano a la unidad cuando $\sigma^2_{x_1^*}/\sigma^2_{e_1}$ es grande. Por tanto, dependiendo de qué tanta sea la variación que exista en x_1^*, en relación con e_1, el error de medición no necesariamente causará sesgos grandes.

Las cosas son más complicadas cuando se agregan más variables explicativas. Como ilustración, considere el modelo

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + u, \tag{9.34}$$

donde la primera de las tres variables explicativas se ha medido con error. Se supone, como es natural, que u no está correlacionada con x_1^*, x_2, x_3, ni con x_1. Una vez más el supuesto crucial es el que se refiere al error de medición e_1. En casi todos los casos, se supone que e_1 no está correlacionado con x_2 y x_3 —las variables explicativas no medidas con error—. El punto clave es si e_1 no está correlacionado con x_1. Si esto es así, entonces la regresión por MCO de y sobre x_1, x_2 y x_3 da estimadores consistentes. Esto se ve con facilidad escribiendo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1, \tag{9.35}$$

donde tanto u como e_1 no están correlacionados con ninguna de las variables explicativas.

Bajo el supuesto ECV en (9.31), MCO será sesgado e inconsistente, debido a que en la ecuación (9.35) e_1 está correlacionado con x_1. Recuerde que esto significa que, en general, *todos* los estimadores de MCO serán sesgados, no sólo los $\hat{\beta}_1$. ¿Qué pasa con el sesgo de atenuación obtenido en la ecuación (9.33)? Resulta que aún existe un sesgo de atenuación de β_1: se puede demostrar que

$$\text{plim}(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma^2_{r_1^*}}{\sigma^2_{r_1^*} + \sigma^2_{e_1}} \right), \tag{9.36}$$

donde r_1^* es el error poblacional en la ecuación $x_1^* = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + r_1^*$. La fórmula (9.36) también funciona en el caso general de k variables cuando x_1 es la única variable mal medida.

Las cosas son menos claras en la estimación de las β_j para las variables no medidas con error. En el caso especial en que x_1^* no esté correlacionada con x_2 y x_3, $\hat{\beta}_2$ y $\hat{\beta}_3$ son consistentes. Pero esto es raro en la práctica. En general, el error de medición en una sola variable ocasiona inconsistencia en todos los estimadores. Por desgracia, el tamaño, e incluso la dirección de los sesgos, no son fácilmente obtenibles.

Ejemplo 9.7

[Ecuación para el GPA con error de medición]

Considérese el problema de estimar el efecto del ingreso familiar sobre el promedio general de calificaciones (GPA, por sus siglas en inglés), una vez controlados *hsGPA* (promedio general de calificaciones en el bachillerato) y *SAT* (puntuación en la prueba de admisión a la universidad). Puede ser que aun cuando el ingreso familiar sea importante para el desempeño antes de la universidad, no tenga efecto directo sobre el desempeño en la universidad. Para probar esto, se puede postular el modelo

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

donde *faminc** será el ingreso familiar anual real. (Éste puede aparecer en forma logarítmica, pero para los fines de este ejemplo se ha dejado en forma lineal.) Es relativamente fácil obtener datos precisos sobre *colGPA*, *hsGPA* y *SAT*. Pero el ingreso familiar, en especial como lo reportan los estudiantes, es fácil que esté mal medido. Si $faminc = faminc^* + e_1$ y se satisface el supuesto de ECV, entonces utilizar el ingreso familiar reportado en lugar del ingreso familiar real sesgará el estimador de MCO de β_1 hacia cero. Una consecuencia de este sesgo hacia abajo es que en la prueba de H_0: $\beta_1 = 0$ se tendrá menos posibilidad de detectar $\beta_1 > 0$.

Por supuesto, el error de medición puede estar presente en más de una de las variables explicativas, o en algunas variables explicativas y en la variable dependiente. Como se vio antes, en general se supone que cualquier error de medición en la variable dependiente no está correlacionado con ninguna de las variables explicativas, ya sean observadas o no. Obtener el sesgo de los estimadores de MCO bajo extensiones del supuesto de ECV es complicado y no conduce a resultados claros.

En algunos casos, es claro que el supuesto de ECV en (9.31) no puede ser verdadero. Considere una variable del ejemplo 9.7:

$$colGPA = \beta_0 + \beta_1 smoked^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

donde *smoked** es la cantidad real de veces que un estudiante ha fumado marihuana en los últimos 30 días. La variable *smoked* es la respuesta a esta pregunta: ¿En cuántas ocasiones distintas ha fumado marihuana en los últimos 30 días? Suponga que se postula el modelo estándar del error de medición

$$smoked = smoked^* + e_1.$$

Aun cuando se supone que los estudiantes tratan de decir la verdad, es poco probable que se satisfaga el supuesto de ECV. Es posible que la gente que no fuma marihuana —de manera que $smoked^* = 0$— reporte $smoked = 0$, por lo que es probable que en el caso de estudiantes que nunca fuman marihuana el error de medición sea cero. Cuando $smoked^* > 0$, es mucho más probable que los estudiantes cuenten mal el número de veces que fumaron marihuana en los últimos 30 días. Esto significa que el error de medición e_1 y la cantidad *real* de veces que fumaron marihuana, *smoked**, están correlacionados, lo cual viola el supuesto (9.31) de ECV. Por desgracia, obtener las consecuencias derivadas de un error de medición que no satisface (9.29) o (9.31) es difícil y queda fuera del alcance de este libro.

Pregunta 9.3

Sea *educ** la escolaridad real, medida en años (pudiendo ser un número no entero), y sea *educ* el grado más alto alcanzado que se reporta. ¿Considera usted que *educ* y *educ** estén relacionadas en el modelo de errores clásicos en las variables?

Antes de dejar esta sección, se señala que el supuesto (9.31) ECV, aunque es más creíble que el supuesto (9.29) es, sin embargo, un supuesto muy fuerte. La verdad se encuentra quizá en algún punto intermedio, y si e_1 está correlacionado tanto con x_1^* como con x_1, MCO es inconsistente. Esto plantea una pregunta importante: ¿Es necesario vivir con estimadores inconsistentes bajo errores clásicos en las variables, u otro tipo de errores de medición que estén correlacionados con x_1? Por fortuna la respuesta es no. En el capítulo 15 se muestra cómo, bajo ciertos supuestos, los parámetros pueden ser estimados consistentemente en presencia de un error general de medición. Este estudio se pospone para después porque requiere salir del campo de la estimación de MCO. (Vea en el problema 9.7 cómo emplear mediciones múltiples para reducir el sesgo de atenuación.)

9.5 Datos faltantes, muestras no aleatorias y observaciones aberrantes

El problema del error de medición tratado en la sección anterior puede verse como un problema de datos: no es posible obtener datos sobre las variables de interés. Además, bajo el modelo de errores clásicos en las variables, el término de error compuesto está correlacionado con la variable independiente mal medida, violando los supuestos de Gauss-Markov.

Otro problema de datos encontrado con frecuencia en capítulos anteriores es la multicolinealidad entre las variables explicativas. Recuerde que la correlación entre las variables explicativas no viola ningún supuesto. Cuando dos variables independientes están fuertemente correlacionadas, puede ser difícil estimar el efecto parcial de cada una. Pero esto se refleja de manera adecuada en los estadísticos usuales de MCO.

En esta sección se presenta una introducción a los problemas de datos que pueden violar el supuesto del muestreo aleatorio, RLM.2. Se pueden aislar casos en los que el muestreo no aleatorio no tiene ningún efecto parcial sobre MCO. En otros casos, el muestreo no aleatorio ocasiona que los estimadores de MCO sean sesgados e inconsistentes. En el capítulo 17 se presenta un estudio más completo que fundamenta varias de las afirmaciones que se hacen aquí.

Datos faltantes

El problema de **datos faltantes** puede presentarse de diferentes maneras. Con frecuencia se toma una muestra aleatoria de personas, escuelas, ciudades, etc., y más tarde se descubre que a algunas de las unidades de la muestra les faltan datos sobre algunas de las variables clave. Por ejemplo, en la base de datos del archivo BWGHT.RAW, 197 de las 1,388 observaciones no tienen información sobre la educación de la madre, o del padre o de ambos. En la base de datos LAWSCH85.RAW, seis de las 156 escuelas no cuentan con información sobre las puntuaciones LSAT de los estudiantes de primer ingreso; a otras escuelas de leyes les falta la información sobre otras variables.

Si a una observación le faltan datos, ya sea sobre la variable dependiente o sobre una de las variables independientes, la observación no puede emplearse en un análisis estándar de regresión múltiple. De hecho, siempre que se indiquen adecuadamente los datos faltantes, todos los paquetes modernos para regresión llevan cuenta de los datos faltantes y, al hacer los cálculos para la regresión, simplemente ignoran esas observaciones. Esto ya se vio de manera explícita en el ejemplo 4.9 sobre el peso al nacer, donde se eliminaron 197 observaciones por no contar con la información sobre la educación de los padres.

Además de que se reduce el tamaño de la muestra para la regresión, ¿tiene la falta de datos alguna otra consecuencia *estadística*? Eso depende de la razón por la que faltan los datos. Si los datos que faltan son al azar, simplemente se reduce el tamaño de la muestra aleatoria tomada de la población. Aunque esto hace que la estimación sea menos precisa, no introduce ningún sesgo: el supuesto de muestreo aleatorio, RLM.2, se sigue satisfaciendo. Hay maneras de usar la información de observaciones en las que sólo falta la información de algunas de las variables, pero en la práctica no se suele hacer esto. Los estimadores sólo mejoran ligeramente y los métodos son un poco complicados. En la mayoría de los casos se ignoran las observaciones en las que falte información.

Muestras no aleatorias

La falta de datos es más problemática cuando es resultado de una **muestra no aleatoria** de la población. Por ejemplo, ¿qué pasa en la base de datos sobre el peso al nacer si la probabilidad de que falte el dato sobre la educación es mayor en aquellas personas con un nivel de educación inferior al promedio? O en la sección 9.2 se usó un conjunto de datos sobre el salario en el que se incluían las puntuaciones de *IQ*. Este conjunto de datos se construyó eliminando de la muestra a varias personas de las que no se tenía el dato sobre la puntuación en el *IQ*. Si obtener la puntuación sobre el *IQ* es más fácil para aquellos que tienen un *IQ* alto, entonces la muestra no es representativa de la población. El supuesto de muestreo aleatorio RLM.2 no se satisface, y habrá que considerar las consecuencias de esto en la estimación por MCO.

Por fortuna, ciertos tipos de muestreo no aleatorio *no* causan sesgo ni inconsistencia en MCO. Bajo los supuestos de Gauss-Markov (pero sin RLM.2), resulta que la muestra puede tomarse con base en las variables *independientes* sin causar ningún problema estadístico. A esto se le llama *elección de la muestra con base en las variables independientes*, y es un ejemplo de **selección muestral exógena**. Para ilustrar esto, suponga que se estima una función de ahorro, en la que el ahorro anual depende del ingreso (*income*), la edad (*age*), el tamaño de la familia (*size*) y de algunos otros factores. Un modelo sencillo es

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u. \tag{9.37}$$

Suponga que el conjunto de datos que se tiene se obtuvo por medio de una encuesta realizada entre personas mayores de 35 años, con lo que se obtiene una muestra no aleatoria de todos los adultos. Aunque esto no es lo ideal, usando esta muestra no aleatoria es posible obtener estimadores insesgados y consistentes de los parámetros del modelo poblacional (9.37). Esto no se demostrará aquí formalmente, pero la razón por la que MCO es insesgado en esta muestra no aleatoria es que la función de regresión $E(saving|income,age,size)$ es la misma para cualquier subconjunto de la población descrito por *income*, *age* o *size*. Siempre que en la población haya suficiente variación en las variables independientes, la selección de la muestra con base en las variables independientes no es un problema serio, salvo que se obtiene una muestra de menor tamaño.

En el ejemplo del *IQ* que se acaba de mencionar, las cosas no son tan claras, porque para incluir a una persona en la muestra no se ha usado ninguna regla fija que se base en el *IQ*. Más bien, la *probabilidad* de estar en la muestra aumenta con el *IQ*. Si en la ecuación del salario los demás factores que determinan ser seleccionado para la muestra son independientes del término de error, entonces se tiene otro caso de selección muestral exógena y, empleando la muestra seleccionada, MCO tendrá todas las propiedades deseables bajo los supuestos de Gauss-Markov.

La situación es muy diferente cuando la selección se basa en la variable dependiente, *y*, a lo que se le llama *selección muestral basada en la variable dependiente* y es un ejemplo de **selección muestral endógena**. Si la muestra está basada en si la variable dependiente es mayor o menor que un valor dado, al estimar el modelo poblacional siempre se presentarán sesgos en MCO. Por ejemplo, suponga que en la población de adultos se desea estimar la relación entre la riqueza individual y otros factores:

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u. \tag{9.38}$$

Suponga que en la muestra sólo se incluyen personas cuya riqueza sea menor a $250,000. Esta es una muestra no aleatoria de la población de interés, que está basada en el valor de la variable dependiente. Emplear una muestra de personas cuya riqueza sea menor a $250,000 dará como resultado estimadores sesgados e inconsistentes de los parámetros de la ecuación (9.32). En concreto, esto se debe a que la regresión poblacional $E(wealth|educ,exper,age)$ no es lo mismo que el valor esperado condicional sobre *wealth* (riqueza) cuando ésta es menor a $250,000.

Hay otros esquemas de muestreo que conducen a muestras no aleatorias de la población, por lo general intencionalmente. Un método usual de recolección de datos es el **muestreo estratificado**, en el que la población se divide en grupos exhaustivos que no se superponen, o estratos. Después, algunos de estos grupos se muestrean con mayor frecuencia que lo indicado por su representación en la población y otros se muestrean con menor frecuencia. Por ejemplo, en algunos estudios a propósito se sobremuestrean grupos minoritarios o grupos de bajos ingresos. Si se requieren o no métodos especiales depende una vez más de si la estratificación es exógena (basada en variables explicativas exógenas) o endógena (basada en variables dependientes). Suponga que se hace un estudio sobre el personal militar sobremuestreando las mujeres debido a que el interés inicial es estudiar los factores que determinan el salario de las mujeres en el ejército. (Sobremuestrear un grupo que es relativamente pequeño en la población es usual en muestras estratificadas.) Siempre que también se muestreen los hombres, se puede usar MCO en la muestra estratificada para estimar cualquier diferencial de salarios entre géneros, al mismo tiempo que los rendimientos de la educación y de la experiencia para todo el personal del ejército. (Se puede estar dispuesto a suponer que los rendimientos de la educación y de la experiencia no sean específicos del género.) MCO es insesgado y consistente debido a que la estratificación se hace respecto a una variable explicativa, a saber, el género.

En cambio, si en el estudio se muestrea el personal del ejército peor pagado, entonces, usando una muestra estratificada, MCO no estima de manera consistente los parámetros de la ecuación del salario en el ejército debido a que la estratificación es endógena. En tales casos, es necesario emplear métodos econométricos especiales [vea Wooldridge (2002, capítulo 17)].

El muestreo estratificado es una forma bastante obvia de muestreo no aleatorio. Otros problemas relacionados con la selección de la muestra son más sutiles. Por ejemplo, en varios de los ejemplos vistos, se han estimado los efectos de diversas variables, particularmente educación y experiencia, sobre el salario por hora. La base de datos WAGE1.RAW, que se ha usado ya en repetidas ocasiones, es en esencia una muestra aleatoria de individuos *trabajadores*. A los economistas laborales con frecuencia les interesa estimar el efecto, por ejemplo, de la educación sobre la *oferta* de salario. La idea es ésta: toda persona en edad laboral se encuentra con una oferta de salario por hora y puede trabajar o no a ese salario. Para una persona que trabaja, la oferta de salario es precisamente su salario. En el caso de las personas que no trabajan, por lo general no se puede observar la oferta de salario. Ahora, dado que la ecuación para la oferta de salario

$$\log(wage^o) = \beta_0 + \beta_1 educ + \beta_2 exper + u \qquad \mathbf{9.39}$$

representa la población de todas las personas en edad laboral, no puede ser estimada usando una muestra aleatoria de esta población; más bien, se tienen datos sobre la oferta de salarios sólo de la población activa (aun cuando es posible obtener datos sobre *educ* y *exper* de la población no activa). Si se usa una muestra aleatoria de la población activa para estimar (9.39), ¿se obtendrán estimadores insesgados? Este caso no es claro. Ya que la muestra se selecciona con base en la decisión de las personas de trabajar (frente a la magnitud de la oferta de salario), este caso no es como el anterior. Sin embargo, como la decisión de trabajar puede estar relacionada con factores no observados que afecten la oferta salarial, la selección puede ser endógena, y esto puede dar como resultado un sesgo de selección de la muestra en los estimadores de MCO. En el capítulo 17 se verán métodos que pueden usarse para probar y corregir sesgos en la selección de la muestra.

Pregunta 9.4

Suponga que interesa conocer los efectos de los gastos de campaña de los mandatarios titulares que buscan la reelección sobre la respuesta de los votantes. Algunos mandatarios titulares deciden no competir para la reelección. Si sólo se pueden recolectar resultados de la votación y de los gastos de los que en realidad compiten, ¿es posible que haya una selección muestral endógena?

Observaciones influyentes y observaciones aberrantes

En algunas aplicaciones, en especial, pero no únicamente, en las que el conjunto de datos es pequeño, las estimaciones de MCO son sensibles a la inclusión de una o varias observaciones. Un estudio completo de las **observaciones aberrantes** y de las **observaciones influyentes** queda fuera del alcance de este libro, debido a que para su desarrollo formal se necesita álgebra de matrices. En términos vagos, una observación es influyente si al eliminarla del análisis las estimaciones clave de MCO se modifican en una cantidad "grande" prácticamente hablando. La noción de observación aberrante es también un poco vaga, porque requiere que se comparen los valores de las variables en una observación con los del resto de la muestra. Sin embargo, se debe estar atento a la presencia de observaciones "inusuales" porque éstas pueden afectar enormemente las estimaciones de MCO.

MCO es sensible a las observaciones aberrantes porque minimiza la suma de residuales cuadrados: a los residuales grandes (positivos o negativos) se les da un peso muy grande en el problema de minimización de mínimos cuadrados. Si al modificar ligeramente la muestra la estimación varía en una cantidad grande en sentido práctico, habrá que poner atención.

Cuando quienes se dedican a la estadística o los econometristas estudian teóricamente el problema de las observaciones aberrantes, algunas veces consideran que los datos provienen de una muestra aleatoria de una población dada —aunque con una distribución inusual que puede dar como resultado valores extremos— y algunas veces se supone que las observaciones aberrantes provienen de una población diferente. Desde una perspectiva práctica, las observaciones aberrantes pueden presentarse por dos razones. El caso más sencillo es cuando se comete un error al registrar el dato. Agregar demasiados ceros a un número o colocar un punto decimal en la posición incorrecta puede echar a perder las estimaciones de MCO, en especial cuando se trata de muestras pequeñas. Siempre es aconsejable calcular estadísticos de resumen, en especial mínimos y máximos, para evitar errores en el registro de los datos. Por desgracia, datos incorrectamente anotados no son siempre obvios.

Las observaciones aberrantes también pueden presentarse cuando la muestra se toma de una población pequeña y uno o varios de los miembros de la población difieren del resto en algún aspecto relevante. La decisión de conservar o de eliminar tales observaciones de un análisis de regresión puede ser difícil y las propiedades estadísticas de los estimadores que se obtienen son complicadas. Las observaciones aberrantes pueden proporcionar información importante incrementando la variación de las variables explicativas (lo que reduce los errores estándar). Los resultados de MCO quizá deban ser dados con y sin las observaciones aberrantes en caso de que uno o varios de los datos modifiquen de manera sustancial los resultados.

Ejemplo 9.8

[Intensidad de la IyD y tamaño de la empresa]

Suponga que los gastos en IyD como porcentaje de las ventas (*rdintens*) estén relacionados con las ventas (*sales*) (en millones) y con las ganancias como porcentaje de las ventas (*profmarg*):

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u. \tag{9.40}$$

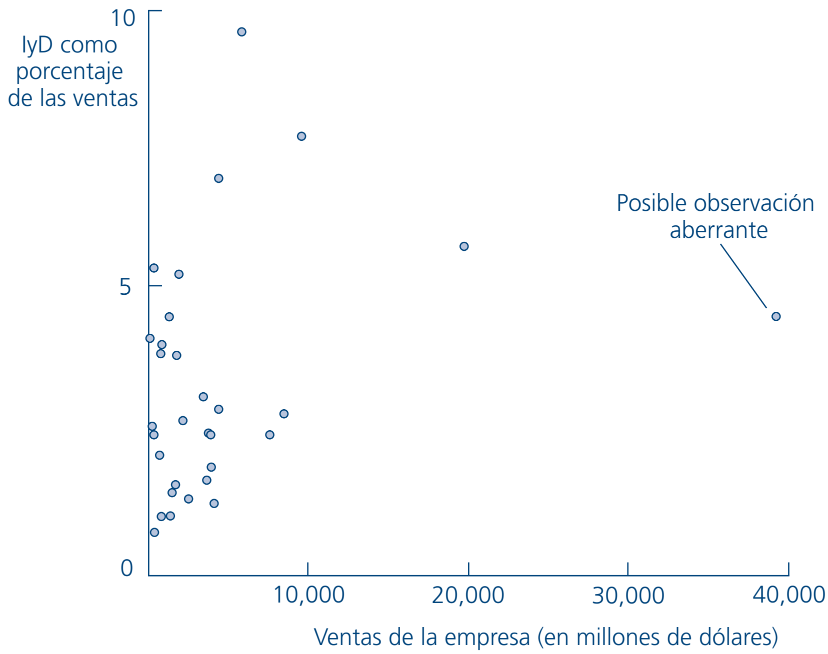
La ecuación de MCO empleando los datos de 32 empresas químicas del archivo RDCHEM.RAW es

$$\begin{aligned} \widehat{rdintens} &= \underset{(0.586)}{2.625} + \underset{(.000044)}{.000053}\, sales + \underset{(.0462)}{.0446}\, profmarg \\ n &= 32,\ R^2 = .0761,\ \bar{R}^2 = .0124. \end{aligned}$$

En esta regresión ni *sales* ni *profmarg* son estadísticamente significativos ni aun al nivel de 10%.

FIGURA 9.1

Diagrama de dispersión de intensidad de IyD contra ventas de la empresa.



De las 32 empresas, 31 tienen ventas anuales menores a \$20 mil millones. Una de las empresas tiene ventas anuales de casi \$40 mil millones. En la figura 9.1 se muestra cuán alejada está esta empresa del resto de la muestra. En términos de ventas, es más de dos veces mayor que cualquiera de las otras empresas, por lo que es recomendable estimar el modelo eliminando este dato. Al hacer esto, se obtiene

$$\widehat{rdintens} = \underset{(0.592)}{2.297} + \underset{(.000084)}{.000186}\, sales + \underset{(.0445)}{.0478}\, profmarg$$
$$n = 31,\ R^2 = .1728,\ \bar{R}^2 = .1137.$$

Eliminando la empresa más grande de la regresión, el coeficiente de *sales* se vuelve más del triple y ahora tiene un estadístico *t* sobre dos. Empleando la muestra de las firmas pequeñas, se concluye que hay un efecto positivo estadísticamente significativo entre la intensidad de IyD y el tamaño de la empresa. El margen de ganancia sigue sin ser significativo, y su coeficiente no cambió mucho.

Algunas veces, en una regresión de MCO, en la que se usan todas las observaciones, las aberrantes se definen mediante el tamaño del residual. En general, esto *no* es aconsejable debido a que las estimaciones de MCO se ajustan para hacer que la suma de residuales cuadrados sea tan pequeña como sea posible. En el ejemplo anterior, incluir las empresas más grandes hace que la línea de regresión sea considerablemente menos inclinada, lo que hace que el residual de esa estimación no sea en especial grande. En realidad, cuando se usan las 32 observaciones el residual de la firma más grande es -1.62. Este valor del residual no está ni a una desviación estándar estimada, $\hat{\sigma} = 1.82$, de la media de los residuales, que es cero por construcción.

Los **residuales estudentizados** se obtienen de los residuales originales de MCO dividiéndolos entre una estimación de su desviación estándar (condicional sobre las variables explicativas de la muestra). La fórmula para los residuales estudentizados se basa en álgebra de matrices, pero resulta haber un truco sencillo para calcular un residual estudentizado para cualquier observación. A saber, se define una variable binaria que es igual a uno para esa observación —por ejemplo, para la observación *h*— y se incluye en la regresión (usando todas las observaciones) junto con las otras variables explicativas. El coeficiente de la variable binaria tiene una interpretación útil: es el residual de la observación *h* calculado a partir de la línea de regresión usando sólo las *otras* observaciones. Por tanto, el coeficiente de la variable binaria puede emplearse para ver qué tanto se aleja esa observación de la línea de regresión obtenida sin ella. Más aún, el estadístico *t* para la variable binaria es igual al residual estudentizado de la observación *h*. Bajo los supuestos del modelo lineal clásico, este estadístico *t* tiene una distribución t_{n-k-1}. Por tanto, un valor grande (en valor absoluto) del estadístico *t* implica un residual grande en relación con su desviación estándar estimada.

En el ejemplo 9.8, si se define una variable binaria para la firma más grande (décima observación en el archivo de los datos) y esta variable se incluye como un regresor más, su coeficiente es -6.57, comprobándose que la observación correspondiente a la empresa más grande está muy alejada de la línea de regresión obtenida usando las demás observaciones. Sin embargo, si se estudentiza, el residual es sólo -1.82. Aunque este es un estadístico *t* marginalmente significativo (valor-*p* de dos colas $=$ .08), no es el mayor residual estudentizado de la muestra. Si se emplea el mismo método para la observación que tiene el mayor valor para *rdintens* —la primera observación, para la cual $rdintens \approx 9.42$— el coeficiente de la variable binaria será 6.72 con un estadístico *t* de 4.56. Por tanto, de acuerdo con esta medición, la primera observación tiene más de aberrante que la décima. Sin embargo, eliminando la primera observación, el coeficiente de *sales* varía sólo en una cantidad pequeña (de .000051 a .000053), aun cuando el coeficiente de *profmarg* se vuelve mayor y estadísticamente significativo. Entonces, ¿es también la primera observación una observación aberrante? Estos cálculos muestran la confusión a la que, aun cuando el conjunto de datos sea pequeño, puede uno llegar al tratar de determinar observaciones que deben ser excluidas del análisis de regresión. Por desgracia, la magnitud del residual estudentizado no necesita corresponder a lo influyente que pueda ser una observación para las estimaciones de pendiente de MCO, y seguramente no para todas a la vez.

Un problema general al usar residuales estudentizados es que, en efecto, todas las demás observaciones se utilizan para estimar la línea de regresión para calcular el residual de una determinada observación. En otras palabras, cuando se obtiene el residual estudentizado de la primera observación, se emplea la décima observación para estimar el intercepto y la pendiente. Dado lo plana que es la línea de regresión con la empresa más grande incluida (observación décima), no es de sorprender que la primera observación, que tiene un valor alto para *rdintens*, se encuentre alejada de la línea de regresión.

Por supuesto, se pueden agregar dos variables binarias a la vez —una para la primera observación y otra para la décima— con lo que para estimar la línea de regresión se usan sólo las 30 observaciones restantes. Si la ecuación se estima sin la primera y la décima observación el resultado es

$$\widehat{rdintens} = 1.939 + .000160\, sales + .0701\, profmarg$$
$$(0.459) \quad (.00065) \qquad\qquad (.0343)$$
$$n = 30,\, R^2 = .2711,\, \bar{R}^2 = .2171$$

El coeficiente para la variable binaria de la primera observación es 6.47 ($t = 4.58$), y para la décima observación es -5.41 ($t = -1.95$). Observe que los coeficientes de *sales* y de *profmarg* son estadísticamente significativos, el último a un nivel sólo de aproximadamente 5% contra una alternativa de dos colas (valor-p = .051). Incluso en esta regresión hay todavía dos observaciones con residuales estandarizados mayores a dos (que corresponden a las dos observaciones restantes con intensidades de IyD mayores a seis).

Ciertas formas funcionales son menos sensibles a las observaciones aberrantes. En la sección 6.2 se dijo que en el caso de la mayoría de las variables económicas, la transformación logarítmica reduce significativamente el rango de los datos y proporciona también formas funcionales —como los modelos de elasticidad constante— que pueden explicar un rango más amplio de datos.

Ejemplo 9.9

[Intensidad de IyD]

Se desea probar si la intensidad de IyD aumenta con el tamaño de la empresa partiendo del modelo

$$rd = sales^{\beta_1}\exp(\beta_0 + \beta_2 profmarg + u). \tag{9.41}$$

Manteniendo los demás factores constantes, la intensidad de IyD aumenta con *sales* si y sólo si $\beta_1 > 1$. Tomando el logaritmo de (9.41) se obtiene

$$\log(rd) = \beta_0 + \beta_1\log(sales) + \beta_2 profmarg + u. \tag{9.42}$$

Usando las 32 empresas, la ecuación de regresión es

$$\widehat{\log(rd)} = -4.378 + 1.084\,\log(sales) + .0217\, profmarg,$$
$$(.468) \quad (.062) \qquad\qquad (.0128)$$
$$n = 32,\, R^2 = .9180,\, \bar{R}^2 = .9123,$$

mientras que eliminando la empresa más grande se obtiene

$$\widehat{\log(rd)} = -4.404 + 1.088\,\log(sales) + .0218\, profmarg,$$
$$(.511) \quad (.067) \qquad\qquad (.0130)$$
$$n = 31,\, R^2 = .9037,\, \bar{R}^2 = .8968.$$

Estos resultados son prácticamente los mismos. En ninguno de los casos se rechaza la hipótesis nula H_0: $\beta_1 = 1$ contra H_1: $\beta_1 > 1$. (¿Por qué?)

En algunos casos, desde un principio se sospecha que ciertas observaciones sean fundamentalmente diferentes del resto de la muestra. Esto suele ocurrir cuando se emplean datos a niveles de agregación grandes, tales como la ciudad, el campo o el estado. El siguiente es un ejemplo de esto.

Ejemplo 9.10

[Tasas estatales de mortalidad infantil]

Datos sobre mortalidad infantil, ingreso *per cápita* y mediciones de atención a la salud, a nivel estatal, pueden obtenerse de *Statistical Abstract of the United States*. Aquí se proporciona un análisis muy sencillo, sólo para ilustrar el efecto de las observaciones aberrantes. Los datos pertenecen al año 1990 y comprenden los 50 estados de Estados Unidos, más el Distrito de Columbia (D.C.). La variable *infmort* es la cantidad de decesos durante el primer año por cada 1,000 nacimientos vivos, *pcinc* es el ingreso *per cápita*, *physic* es médicos por cada 100,000 miembros de la población civil y *popul* es la población (en miles). Los datos se encuentran en el archivo INFMRT.RAW. Todas las variables independientes se presentan en forma logarítmica:

$$\begin{aligned}\widehat{infmort} = {} & \underset{(20.43)}{33.86} - \underset{(2.60)}{4.68}\log(pcinc) + \underset{(1.51)}{4.15}\log(physic) \\ & -\underset{(.287)}{.088}\log(popul) \\ & n = 51,\ R^2 = .139,\ \bar{R}^2 = .084.\end{aligned} \tag{9.43}$$

Se estimó que un mayor ingreso *per cápita* reduce la mortalidad infantil, lo cual es un resultado esperado. Pero más médicos *per cápita* está relacionado con índices mayores de mortalidad infantil, cosa que es contraintuitiva. El índice de mortalidad infantil no parece estar relacionado con el tamaño de la población.

El Distrito de Columbia es inusual en el sentido de que presenta sectores de extrema pobreza y de enorme riqueza en un área pequeña. El índice de mortalidad infantil en 1990 en D.C. fue de 20.7 en comparación con 12.4 en el siguiente estado con mortalidad más alta. En el Distrito de Columbia hay 615 médicos por 100,000 miembros de la sociedad civil, en comparación con 337 para el siguiente estado con mayor cantidad. La elevada cantidad de médicos en D.C., unida a su alto índice de mortalidad infantil, seguramente puede influir en los resultados. Si de la regresión se elimina D.C. se obtiene

$$\begin{aligned}\widehat{infmort} = {} & \underset{(12.42)}{23.95} - \underset{(1.64)}{.57}\log(pcinc) - \underset{(1.19)}{2.74}\log(physic) \\ & + \underset{(.191)}{.629}\log(popul) \\ & n = 50,\ R^2 = .273,\ \bar{R}^2 = .226.\end{aligned} \tag{9.44}$$

Ahora, se encuentra que un mayor número de médicos *per cápita* hace que disminuya la mortalidad infantil y el estimador es estadísticamente distinto de cero al nivel de 5%. El efecto del ingreso *per cápita* disminuyó notablemente y ya no es estadísticamente significativo. En la ecuación (9.44), los índices de mortalidad infantil son mayores en estados más poblados y esta relación es estadísticamente significativa. Además, cuando se elimina D.C. de la regresión, se explica mucha más de la variación en *infmort*. Es claro que D.C. tuvo una influencia sustancial sobre las estimaciones iniciales y es probable que se deje fuera de cualquier análisis posterior.

Como demuestra el ejemplo 9.8, analizar las observaciones para tratar de determinar cuáles son aberrantes y cuáles tienen una influencia sustancial sobre las estimaciones de MCO es una tarea difícil. Análisis más avanzados permiten emplear métodos formales para determinar qué observaciones tienen probabilidad de ser influyentes. Empleando álgebra de matrices, Belsley, Kuh y Welsh (1980) definieron el *influjo* de una observación, lo que formaliza la noción de que una observación tenga una influencia grande o pequeña sobre las estimaciones de MCO. Estos autores proporcionaron también un análisis más profundo de los residuales estandarizados y estudentizados.

9.6 Estimación por mínimas desviaciones absolutas

En lugar de tratar de determinar qué observaciones, si las hubiera, tienen una influencia indebida sobre las estimaciones de MCO, una manera para protegerse de las observaciones aberrantes es usar un método que sea menos sensible que MCO a las observaciones aberrantes. Uno de estos métodos, muy usual entre los econometristas aplicados, es el llamado **mínimas desviaciones absolutas (MDA)**. En un modelo lineal, los estimadores de MDA de las β_j minimizan la suma de los valores absolutos de los residuales,

$$\min_{b_0, b_1, ..., b_k} \sum_{i=1}^{n} |y_i - b_0 - b_1 x_{i1} - ... - b_k x_{ik}|. \quad \boxed{9.45}$$

A diferencia de MCO, que minimiza la suma de los residuales cuadrados, las estimaciones de MDA no pueden obtenerse de forma cerrada —es decir, no es posible dar una fórmula para ellos—. Históricamente, resolver el problema planteado en la ecuación (9.45) era difícil desde el punto de vista de los cálculos, en especial si se tienen muestras grandes y muchas variables explicativas. Pero con el enorme adelanto en las últimas dos décadas en la velocidad computacional, es bastante fácil obtener las estimaciones de MDA aun con bases de datos grandes.

Como MDA no da un mayor peso a los residuales más grandes, es mucho menos sensible que MCO a las variaciones en los valores extremos de los datos. De hecho, se sabe que MDA está diseñado para estimar los parámetros de la **mediana condicional**, y no de la media condicional, de y dados $x_1, x_2, ..., x_k$. Como a la mediana no le afectan variaciones grandes en las observaciones extremas, se sigue que las estimaciones MDA de los parámetros son más invariantes a las observaciones aberrantes. (Vea en la sección A.1 un breve análisis de la mediana muestral.) Para elegir las estimaciones, MCO eleva al cuadrado cada residual, y por tanto las estimaciones de MCO pueden ser muy sensibles a las observaciones aberrantes, como se vio en los ejemplos 9.8 y 9.10.

MDA, además de necesitar muchos más cálculos que MCO, tiene otra desventaja: que toda inferencia estadística en la que intervengan los estimadores MDA sólo se justifica en la medida en que el tamaño de la muestra aumente. [Las fórmulas son un poco complicadas y requieren álgebra de matrices, pero no se necesitan aquí. Koenker (2005) proporciona un estudio exhaustivo.] Recuerde que de acuerdo con los supuestos del modelo lineal clásico, los estadísticos t de MCO tienen distribuciones t exactas, y los estadísticos F tienen distribuciones F exactas. Aunque para MDA existen versiones asintóticas de estos estadísticos —que son dadas de manera rutinaria por el software que calcula estimaciones MDA—, éstas sólo se justifican en el caso de muestras grandes. Al igual que el problema adicional que significa calcular las estimaciones MDA, la falta de una inferencia exacta para MDA es un problema de menor importancia, ya que en la mayoría de las aplicaciones de MDA se tienen cientos, si no es que miles, de observaciones. Por supuesto, sería inapropiado si se emplearan aproximaciones para muestras grandes en ejemplos como el 9.8, en el que $n = 32$. En cierto sentido, esto no difiere mucho de MCO porque, la mayoría de las veces, es necesario apelar a aproximaciones para muestras grandes con objeto de justificar las inferencias de MCO cuando alguno de los supuestos del MLC falla.

Otra desventaja más sutil pero importante de MDA es que no siempre estima de manera consistente los parámetros que aparecen en la función de media condicional, $E(y|x_1, ..., x_k)$. Como ya se dijo, MDA está diseñado para estimar los efectos sobre la mediana condicional. En general, la media y la mediana sólo son iguales cuando la distribución de y dadas las covariadas $x_1, ..., x_k$ es simétrica respecto a $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$. (De manera equivalente, el término del error poblacional, u, es simétrico respecto a cero.) Recuerde que MCO produce estimadores insesgados y consistentes de los parámetros en la media condicional, ya sea que la distribución del error sea o no simétrica; la simetría no aparece entre los supuestos de Gauss-Markov. Cuando se aplican MDA y MCO a casos con distribuciones asimétricas, el efecto parcial estimado de, por ejemplo, x_1, obtenido con MDA puede ser muy diferente del efecto parcial obtenido con MCO. Pero esta diferencia puede reflejar únicamente la diferencia entre la mediana y la media y no tener nada que ver con observaciones aberrantes. Vea un ejemplo en el ejercicio para computadora C9.9.

Si se supone que en el modelo (9.2) el error poblacional u es *independiente* de $(x_1, ..., x_k)$, entonces las estimaciones de pendiente de MCO y de MDA diferirán sólo en un error de muestreo, ya sea que la distribución de u sea simétrica o no. En general, las estimaciones del intercepto serán diferentes reflejando el hecho de que si la media de u es cero, entonces bajo asimetría su mediana es diferente de cero. Por desgracia, la independencia entre el error y las variables explicativas suele ser muy poco realista cuando se emplea MDA. En particular, la independencia elimina la heterocedasticidad, problema que suele presentarse en aplicaciones con distribuciones asimétricas.

Mínimas desviaciones absolutas es un caso especial de lo que se suele conocer como *regresión robusta*. Por desgracia, la manera en que se usa aquí "robusta" puede llevar a confusión. En la literatura estadística, un estimador de regresión robusto es relativamente insensible a observaciones extremas. En efecto, a las observaciones con residuales grandes se les da menos peso que en mínimos cuadrados. [Berk (1990) contiene un estudio introductorio a los estimadores que son robustos a observaciones aberrantes.] Con base en la discusión anterior, en la jerga econométrica, MDA no es un estimador robusto de la media condicional porque requiere de más supuestos con objeto de estimar de manera consistente los parámetros de media condicional. En la ecuación (9.2), o bien la distribución de u dadas $(x_1, ..., x_k)$ debe de ser simétrica respecto a cero, o bien u debe ser independiente de $(x_1, ..., x_k)$. Ninguna de estas dos cosas se requiere para MCO.

RESUMEN

Se continuó investigando temas importantes sobre especificación y datos que suelen surgir en el análisis empírico de corte transversal. La especificación incorrecta de la forma funcional hace que la ecuación estimada sea difícil de interpretar. Sin embargo, una forma funcional incorrecta puede ser detectada agregando términos cuadráticos, calculando el estadístico RESET o probando contra un modelo alterno no anidado usando la prueba de Davidson-MacKinnon. No es necesario recolectar datos adicionales.

Resolver el problema de variables omitidas es más difícil. En la sección 9.2 se vio una posible solución basada en el uso de una variable proxy para la variable omitida. Bajo supuestos razonables, incluir una variable proxy en una regresión de MCO elimina, o al menos reduce, los sesgos. La dificultad de aplicar este método es que las variables proxy son difíciles de encontrar. Una posibilidad general es usar datos de la variable dependiente de algún año anterior.

Los economistas aplicados con frecuencia se encuentran con errores de medición. Bajo los supuestos de errores clásicos en las variables (ECV) un error de medición en la variable dependiente no tiene ningún efecto sobre las propiedades estadísticas de MCO. En cambio, bajo los supuestos de ECV para una variable independiente, el estimador MCO del coeficiente de la variable mal medida está sesgado hacia cero. El sesgo en los coeficientes de las otras variables puede ir en cualquier dirección y es difícil de determinar.

Muestras no aleatorias de una población pueden conducir a sesgos en MCO. Cuando la selección de la muestra está correlacionada con el término de error u, MCO por lo general es sesgado e inconsistente. Por otro lado, una selección muestral exógena —que está basada en las variables explicativas o que es independiente de u— no causa problemas a MCO. Las observaciones aberrantes en un conjunto de datos pueden tener un impacto grande sobre las estimaciones de MCO, en especial en muestras pequeñas. Es importante identificar las observaciones aberrantes y volver a estimar los modelos sin las observaciones sospechosas de ser aberrantes.

La estimación de las mínimas desviaciones absolutas es una alternativa a MCO que es menos sensible a las observaciones aberrantes y que proporciona estimaciones consistentes de los parámetros de mediana condicional.

TÉRMINOS CLAVE

Datos faltantes
Efecto parcial promedio (EPP)
Error de medición
Error de medición multiplicativo
Errores clásicos en las variables (ECV)
Especificación incorrecta de la forma funcional
Mediana condicional
Mínimas desviaciones absolutas (MDA)
Modelo de coeficientes (pendientes) aleatorios
Modelos no anidados
Muestra no aleatoria
Muestreo estratificado
Observaciones aberrantes
Observaciones influyentes
Prueba de Davidson-MacKinnon
Prueba de error de especificación de la regresión (RESET, por sus siglas en inglés)
Residuales estudentizados
Selección muestral endógena
Selección muestral exógena
Sesgo de atenuación
Solución suplente al problema de variables omitidas
Variable dependiente rezagada
Variable explicativa endógena
Variable proxy

PROBLEMAS

9.1 En el problema 4.11, la *R*-cuadrada de la estimación del modelo

$$\log(salary) = \beta_0 + \beta_1\log(sales) + \beta_2\log(mktval) + \beta_3 profmarg + \beta_4 ceoten + \beta_5 comten + u,$$

empleando los datos del archivo CEOSAL2.RAW, fue $R^2 = .353$ ($n = 177$). Cuando se agregan $ceoten^2$ y $comten^2$, $R^2 = .375$. ¿Hay alguna evidencia de especificación incorrecta de la forma funcional?

9.2 El ejercicio para computadora C8.4 se modifica usando los resultados de la votación en 1990 para los mandatarios titulares que fueron elegidos en 1988. El candidato A fue elegido en 1988 y en 1990 buscaba la reelección; *voteA90* es la participación del candidato A en la votación entre dos partidos en 1990. La participación del candidato A en 1988 se usa como variable proxy para la calidad del candidato. Todas las demás variables son de la elección de 1990. Empleando los datos del archivo VOTE2.RAW se estimaron las ecuaciones siguientes:

$$\begin{aligned}\widehat{voteA90} = {}& \underset{(9.25)}{75.71} + \underset{(.046)}{.312}\, prtystrA + \underset{(1.01)}{4.93}\, democA \\ & - \underset{(.684)}{.929}\log(expendA) - \underset{(.281)}{1.950}\log(expendB) \\ n = {}& 186,\ R^2 = .495,\ \bar{R}^2 = .483,\end{aligned}$$

y

$$\begin{aligned}\widehat{voteA90} = {}& \underset{(10.01)}{70.81} + \underset{(.052)}{.282}\, prtystrA + \underset{(1.06)}{4.52}\, democA \\ & - \underset{(.687)}{.839}\log(expendA) - \underset{(.292)}{1.846}\log(expendB) + \underset{(.053)}{.067}\, voteA88 \\ n = {}& 186,\ R^2 = .499,\ \bar{R}^2 = .485.\end{aligned}$$

i) Interprete el coeficiente de *voteA88* y analice su significancia estadística.
ii) ¿Tiene la adición de *voteA88* un gran efecto sobre los demás coeficientes?

9.3 Sea *math10* el porcentaje de estudiantes de la Universidad de Michigan que obtiene una calificación de aprobado en el examen de matemáticas estandarizado (vea también el ejemplo 4.2). Interesa estimar el efecto del gasto por estudiante (*expend*) sobre el desempeño en matemáticas. Un modelo sencillo es

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 \log(enroll) + \beta_3 poverty + u,$$

donde *poverty* es el porcentaje de estudiantes que viven en la pobreza y *enroll* es el tamaño de la matrícula escolar.

i) La variable *lnchprg* es el porcentaje de estudiantes con derecho al programa federal de desayuno escolar. ¿Por qué es ésta una variable proxy adecuada para *poverty*?

ii) La tabla siguiente contiene estimaciones de MCO, con y sin *lnchprg* como variable explicativa.

Variable dependiente: *math10*

Variables independientes	(1)	(2)
log(*expend*)	11.13 (3.30)	7.75 (3.04)
log(*enroll*)	.022 (.615)	–1.26 (.58)
lnchprg	—	–.324 (.036)
intercepto	−69.24 (26.72)	−23.14 (24.99)
Observaciones *R*-cuadrada	428 .0297	428 .1893

Explique por qué el efecto de los gastos sobre *math10* es menor en la columna (2) que en la (1). En la columna (2), ¿sigue siendo el efecto mayor a cero?

iii) ¿Parece que, permaneciendo todos los demás factores constantes, el índice de aprobados sea menor en escuelas más grandes? Explique.

iv) Interprete el coeficiente de *lnchprg* en la columna (2).

v) ¿Qué piensa del importante aumento en R^2 que se observa de la columna (1) a la (2)?

9.4 La ecuación siguiente explica las horas por semana que ve televisión un niño de acuerdo con la edad del mismo (*age*), educación de la madre (*motheduc*), educación del padre (*fatheduc*) y cantidad de hermanos (*sibs*):

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

Preocupa que la variable *tvhours** se haya medido con error en la investigación realizada. Sea *tvhours* las horas semanales de televisión reportadas.

i) ¿Qué requieren en esta aplicación los supuestos de errores clásicos en las variables (ECV)?

ii) ¿Considera usted que es probable que se satisfagan los supuestos de ECV? Explique.

9.5 En el ejemplo 4.4, se estimó un modelo en el que para una muestra de universidades se relacionaba la cantidad de delitos en un campus universitario con la matrícula de estudiantes. La muestra empleada no fue una muestra aleatoria de las universidades de Estados Unidos, dado que en 1992 muchas de ellas no reportaban los delitos en el campus. ¿Considera usted que el que las universidades no hayan reportado los delitos puede verse como una selección muestral exógena? Explique.

9.6 En el modelo (9.17), muestre que MCO estima de manera consistente α y β si a_i no está correlacionada con x_i y b_i no está correlacionada con x_i y x_i^2, los cuales son supuestos más débiles que los dados en (9.19). [*Sugerencia*: exprese la ecuación como en (9.18) y recuerde, de acuerdo con el capítulo 5, que para la consistencia del intercepto y de la pendiente de MCO basta que $\mathrm{E}(u_i) = 0$ y $\mathrm{Cov}(x_i, u_i) = 0$.]

9.7 Considere el modelo de regresión simple con error de medición clásico, $y = \beta_0 + \beta_1 x^* + u$, donde se tienen m mediciones de x^*. Exprese éstas como $z_h = x^* + e_h$, $h = 1, \ldots, m$. Suponga que x^* no está correlacionada con u, e_1, ..., e_m, que los errores de medición no están correlacionados por pares y que tienen la misma varianza, σ_e^2. Sea $w = (z_1 + \ldots + z_m)/m$ el promedio de las mediciones de x^*, de manera que para cada observación i, $w_i = (z_{i1} + \ldots + z_{im})/m$ sea el promedio de las m mediciones. Sea $\bar{\beta}_1$ el estimador de MCO obtenido de la regresión simple de y_i sobre 1, w_i, $i = 1, \ldots, n$, usando una muestra aleatoria de datos.

i) Muestre que

$$\mathrm{plim}(\bar{\beta}_1) = \beta_1\left\{\frac{\sigma_{x^*}^2}{[\sigma_{x^*}^2 + (\sigma_e^2/m)]}\right\}.$$

[*Sugerencia*: el plim de $\bar{\beta}_1$ es $\mathrm{Cov}(w, y)/\mathrm{Var}(w)$.]

ii) Compare la inconsistencia en $\bar{\beta}_1$ con aquella en la que se tiene una sola medición (es decir, $m = 1$). ¿Qué ocurre cuando m aumenta? Explique.

EJERCICIOS EN COMPUTADORA

C9.1 i) Aplique la prueba RESET de acuerdo con la ecuación (9.3) para el modelo estimado en el ejercicio para computadora C7.5. ¿Hay en esta ecuación evidencia de especificación incorrecta de la forma funcional?

ii) Calcule una forma de RESET que sea robusta a la heterocedasticidad. ¿Modifica esto su conclusión del inciso i)?

C9.2 Para este ejercicio utilice la base de datos WAGE2.RAW.

i) En el ejemplo 9.3 utilice la variable *KWW* (puntuación en el examen "Conocimiento del mundo laboral") como proxy para capacidad en lugar de *IQ*. En este caso, ¿cuál es el rendimiento de la educación?

ii) Ahora, utilice, simultáneamente *IQ* y *KWW* como variables proxy. ¿Qué ocurre con el rendimiento estimado de la educación?

iii) En el inciso ii), ¿son *IQ* y *KWW* individualmente significativas? ¿Son conjuntamente significativas?

C9.3 Para este ejercicio utilice los datos del archivo.

i) Considere el modelo de regresión simple

$$\log(scrap) = \beta_0 + \beta_1 grant + u,$$

donde *scrap* es la tasa de desperdicio en una empresa y *grant* es una variable binaria que indica si la empresa recibe subvención para capacitación en el trabajo. ¿Ve alguna razón por la que los factores no observados en *u* puedan estar correlacionados con *grant*?

ii) Estime el modelo de regresión simple empleando los datos de 1988. (Tendrá 54 observaciones.) Obtener una subvención para la capacitación en el trabajo, ¿reduce significativamente la tasa de desperdicio de una empresa?

iii) Ahora, agregue como variable explicativa log($scrap_{87}$). ¿Cómo modifica esto el efecto estimado de *grant*? Interprete el coeficiente de *grant*. ¿Es este coeficiente estadísticamente significativo al nivel de 5% contra la alternativa de una cola H_1: $\beta_{grant} < 0$?

iv) Pruebe la hipótesis nula de que el parámetro de log($scrap_{87}$) es igual a uno contra la alternativa de dos colas. Dé el valor-*p* de esta prueba.

v) Repita los incisos iii) y iv), empleando errores estándar robustos a la heterocedasticidad y analice brevemente cualquier diferencia notable.

C9.4 Para este ejercicio utilice los datos de 1990 del archivo INFMRT.RAW.

i) Estime de nuevo la ecuación (9.43), pero ahora incluya una variable binaria (que se llame *DC*) para el Distrito de Columbia. Interprete el coeficiente de *DC* y analice su magnitud y significancia.

ii) Compare las estimaciones y los errores estándar del inciso i) con los de la ecuación (9.44). ¿Qué concluye acerca de la inclusión de una variable binaria para una sola observación?

C9.5 Emplee los datos del archivo RDCHEM.RAW para examinar con más detalle los efectos de las observaciones aberrantes sobre las estimaciones de MCO y para ver cómo MDA es menos sensible a las observaciones aberrantes. El modelo es

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profmarg + u,$$

donde primero deberá modificarse *sales* para que esté dada en miles de millones de dólares con objeto de hacer que las estimaciones sean más fáciles de interpretar.

i) Estime la ecuación anterior por MCO, con y sin la empresa que tiene ventas anuales por casi 40 mil millones de dólares. Analice cualquier diferencia notable en los coeficientes estimados.

ii) Estime la misma ecuación por MDA, también con y sin la empresa más grande. Analice cualquier diferencia notable en los coeficientes estimados.

iii) Con base en sus hallazgos en los incisos i) y ii), ¿cuál diría usted que es más invariante, MDA o MCO?

C9.6 Repita el ejemplo 4.10 eliminando aquellas escuelas en las que los beneficios para los profesores sean menores a 1% del salario.

i) ¿Cuántas observaciones se pierden?

ii) ¿Eliminar estas observaciones tiene alguna consecuencia importante sobre el efecto sustitución estimado?

C9.7 Para este ejercicio utilice los datos del archivo LOANAPP.RAW.

i) ¿En cuántas observaciones es $obrat > 40$, es decir, las obligaciones de endeudamiento son más de 40% del ingreso total?

ii) Estime de nuevo el modelo del inciso iii) del ejercicio para computadora C7.8, excluyendo aquellas observaciones para las que $obrat > 40$. ¿Qué ocurre con la estimación y con el estadístico *t* de *white*?

iii) ¿Parece la estimación de β_{white} demasiado sensible a la muestra empleada?

C9.8 Para este ejercicio emplee los datos del archivo TWOYEAR.RAW.

i) La variable *stotal* es una variable de una prueba estandarizada, que puede funcionar como variable proxy para capacidad no observada. Encuentre la media y la desviación estándar muestrales de *stotal*.

ii) Corra las regresiones simples de *jc* y *univ* sobre *stotal*. ¿Están relacionadas estas dos variables de educación universitaria con *stotal*? Explique.

iii) Agregue *stotal* a la ecuación (4.17) y pruebe la hipótesis de que el rendimiento de dos y cuatro años de universidad es el mismo, contra la alternativa de que el rendimiento de cuatro años de universidad es mayor. Compare sus resultados con los de la sección 4.4.

iv) Agregue $stotal^2$ a la ecuación estimada en el inciso iii). ¿Parece necesario el término cuadrático?

v) Agregue a la ecuación del inciso iii) los términos de interacción $stotal \cdot jc$ y $stotal \cdot univ$. ¿Son estos términos conjuntamente significativos?

vi) ¿Cuál sería su modelo final en el que se controle la capacidad mediante el uso de *stotal*? Justifique su respuesta.

C9.9 En este ejercicio se pide que compare las estimaciones de MCO y MDA de los efectos de la elegibilidad para el plan 40l(k) si se es elegible) sobre los activos financieros netos (*nettfa*). El modelo es

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 male + \beta_6 e401k + u.$$

i) Utilice los datos del archivo 401KSUBS.RAW para estimar la ecuación por MCO y dé los resultados en la forma usual. Interprete el coeficiente de *e401k*.

ii) Utilice los residuales de MCO para probar si hay heterocedasticidad usando la prueba Breusch-Pagan. ¿Es u independiente de las variables explicativas?

iii) Estime la ecuación empleando MDA y dé los resultados en la misma forma que con MCO. Interprete la estimación por MDA de β_6.

iv) Concilie sus resultados de los incisos i) y iii).

C9.10 Para este ejercicio debe utilizar dos bases de datos, JTRAIN2.RAW y JTRAIN3.RAW. La primera es el resultado de un experimento de capacitación para el trabajo. El archivo JTRAIN3.RAW contiene datos observacionales en que son principalmente los mismos individuos quienes determinan si participan en la capacitación para el trabajo. Las dos bases de datos cubren el mismo periodo.

i) ¿En la base de datos JTRAIN2.RAW, ¿qué fracción de los hombres recibió capacitación para el trabajo? ¿Cuál es esta fracción en JTRAIN3.RAW? ¿Por qué cree usted que exista una diferencia tan grande?

ii) Empleando JTRAIN2.RAW, corra una regresión simple de *re78* sobre *train*. ¿Cuál es el efecto estimado sobre las ganancias reales (*re78*) que tiene la participación en la capacitación para el trabajo ($train = 1$, si participó)?

iii) Ahora, como controles, agregue a la regresión del inciso ii) las variables *re74*, *re75*, *educ*, *age*, *black* e *hisp*. ¿Varía mucho el efecto estimado de la capacitación para el trabajo sobre *re78*? ¿Por qué? (*Sugerencia:* recuerde que éstos son datos experimentales.)

iv) Corra las regresiones de los incisos ii) y iii) con los datos del archivo JTRAIN3.RAW, y dé solamente los coeficientes estimados de *train*, así como su estadístico t. ¿Cuál es ahora el efecto de controlar los factores adicionales, y por qué?

v) Defina $avgre = (re74 + re75)/2$. Determine los promedios muestrales, las desviaciones estándar y los valores mínimo y máximo de las dos bases de datos. ¿Son estas bases de datos representativas de las mismas poblaciones de 1978?

vi) Para casi 96% de hombres de la base de datos JTRAIN2.RAW *avgre* es menor de 10,000 dólares. Empleando sólo estos hombres, corra la regresión

re78 sobre *train,re74,re75,educ,age,black,hisp*

y dé la estimación para capacitación (*train*) y su estadístico t. Corra la misma regresión con la base de datos JTRAIN3.RAW, usando sólo los hombres para los que $avgre \leq 10$.

Para la submuestra de hombres de bajo ingreso, compare los efectos estimados de la capacitación entre los conjuntos de datos experimental y no experimental.

vii) Ahora utilice cada base de datos para estimar la regresión simple de *re78* sobre *train*, pero sólo para los hombres sin empleo en 1974 y 1975. Compare ahora las estimaciones para capacitación.

viii) Usando los resultados de las regresiones anteriores, analice la importancia potencial de tener poblaciones comparables al contrastar estimaciones experimentales y no experimentales.

C9.11 Para esta pregunta utilice los datos del archivo MURDER.RAW del año 1993, aunque primero necesitará obtener la tasa rezagada de homicidios, llámesele $mrdrte_{-1}$.

i) Corra la regresión del *mrdrte* sobre *exec*, *unem* (ejecuciones por pena de muerte en los últimos tres años y tasa de desempleo, respectivamente). ¿Cuáles son el coeficiente y el estadístico *t* de *exec*? ¿Proporciona esta regresión evidencia de un efecto disuasivo de la pena capital?

ii) ¿Cuántas ejecuciones se reportan en Texas en 1993? (En realidad, esta es la suma de las ejecuciones en el año en curso y en los dos años anteriores.) Compare esto con los otros estados. A la regresión del inciso i) agregue una variable binaria para Texas. ¿Es su estadístico *t* inusualmente grande? De acuerdo con esto, ¿parece que Texas sea "una observación aberrante"?

iii) A la regresión del inciso i) agregue la tasa rezagada de homicidios. ¿Qué sucede con $\hat{\beta}_{exec}$ y con su significancia estadística?

iv) De acuerdo con la regresión del inciso iii), ¿parece que Texas sea una observación aberrante? ¿Qué efecto tiene sobre $\hat{\beta}_{exec}$ eliminar Texas de la regresión?

C9.12 Para este ejercicio utilice los datos del archivo ELEM94_95. Vea también el ejercicio para computadora C4.10.

i) Usando todos los datos, corra la regresión de *lavgsal* sobre *bs*, *lenrol*, *lstaff* y *lunch*. Dé el coeficiente de *bs* así como su error estándar usual y su error estándar robusto a la heterocedasticidad. ¿Qué concluye acerca de la significancia económica y estadística de $\hat{\beta}_{bs}$?

ii) Ahora elimine las cuatro observaciones en las que $bs > .5$, es decir, en las que las ganancias medias son (supuestamente) más de 50% del sueldo promedio. ¿Cuál es el coeficiente de *bs*? ¿Es estadísticamente significativo usando el error estándar robusto a la heterocedasticidad?

iii) Verifique que las cuatro observaciones en las que $bs > .5$ son las 68, 1,127, 1,508 y 1,670. Defina cuatro variables binarias para cada una de estas observaciones. (Puede llamarles *d*68, *d*1127, *d*1508, y *d*1670.) Agregue estas variables a la regresión del inciso i) y verifique que los coeficientes y los errores estándar de MCO de las otras variables son idénticos a los del inciso ii). ¿Cuál de las cuatro binarias tienen un estadístico *t* estadísticamente distinto de cero al nivel de 5%?

iv) Verifique que en esta base de datos el dato con el residual estudentizado mayor (estadístico *t* más grande para la variable binaria) tiene una gran influencia sobre las estimaciones de MCO del inciso iii). (Es decir, corra MCO usando todas las observaciones menos la que tiene el residual estudentizado mayor.) ¿Tiene efectos importantes eliminar cada una de las otras observaciones con $bs > .5$?

v) ¿Qué concluye acerca de la sensibilidad de MCO a una sola observación, incluso con un tamaño de muestra grande?

vi) Verifique que el estimador de MDA no es sensible a la inclusión de la observación identificada en el inciso iii).

PARTE 2

Análisis de regresión con datos de series de tiempo

Ahora que se tienen bases más sólidas para comprender cómo utilizar el modelo de regresión múltiple para las aplicaciones de corte transversal, podemos estudiar el análisis econométrico de los datos de series de tiempo. Dado que nos basaremos mucho en el método de mínimos cuadrados ordinarios, ya se ha realizado buena parte del trabajo concerniente a su mecánica e inferencia. No obstante, como se señaló en el capítulo 1, los datos de series de tiempo cuentan con ciertas características que no poseen aquellos de corte transversal y exigen atención especial al aplicar MCO.

El capítulo 10 cubre el análisis de regresión básico y presta atención a los problemas exclusivos de los datos de series de tiempo. Se proporciona un conjunto de supuestos de Gauss-Markov y del modelo lineal clásico para las aplicaciones de series de tiempo. También se estudian problemas sobre la forma funcional, las variables binarias, las tendencias y la estacionalidad.

En virtud de que ciertos modelos de series de tiempo violan por fuerza los supuestos de Gauss-Markov, en el capítulo 11 se describe la naturaleza de estas violaciones y se presentan las propiedades de mínimos cuadrados ordinarios para muestras grandes. Como ya no es posible suponer un muestreo aleatorio, se deben cubrir las condiciones que limitan la correlación temporal en una serie de tiempo para asegurar que el análisis asintótico usual sea válido.

En el capítulo 12 se plantea un problema nuevo e importante: la correlación serial de los términos de error en las regresiones con series de tiempo. Se analizarán las consecuencias, las formas de hacer pruebas y los métodos para tratar con la correlación serial. El capítulo 12 también contiene una explicación de cómo surge la heterocedasticidad en los modelos de series de tiempo.

CAPÍTULO 10

Análisis básico de regresión con datos de series de tiempo

Este capítulo comienza con el estudio de las propiedades de MCO para estimar los modelos de regresión lineal en los que se utilizan datos de series de tiempo. En la sección 10.1 se estudian algunas diferencias conceptuales entre los datos de series de tiempo y los datos de corte transversal. La sección 10.2 proporciona algunos ejemplos de regresiones con series de tiempo que se estiman a menudo en las ciencias sociales empíricas. Luego centramos nuestra atención en las propiedades de los estimadores de MCO para muestras finitas y se establecen los supuestos de Gauss-Markov y del modelo lineal clásico para la regresión con series de tiempo. Si bien estos supuestos cuentan con características comunes al caso de corte transversal, también tienen algunas diferencias significativas que es necesario destacar.

Además, se revisarán algunos temas que tratamos en la regresión con datos de corte transversal, tales como el uso e interpretación de la forma funcional logarítmica y las variables binarias. En la sección 10.5 se retoman los temas importantes de cómo incorporar las tendencias y representar la estacionalidad en la regresión múltiple.

10.1 Naturaleza de los datos de series de tiempo

Una característica obvia de los datos de series de tiempo que los distingue de aquellos de corte transversal es que tienen un orden temporal. Por ejemplo, en el capítulo 1 se analizó de forma breve una base de datos de series de tiempo sobre el empleo, el salario mínimo y otras variables económicas de Puerto Rico. En dicha base de datos se debe saber que los datos de 1970 preceden de inmediato a los de 1971. Para analizar los datos de series de tiempo en las ciencias sociales, es necesario reconocer que el pasado influye en el futuro, pero no a la inversa (a diferencia del universo de "Viaje a las estrellas"). Para destacar el orden adecuado de los datos de series de tiempo, en la tabla 10.1 aparece una lista parcial de datos sobre la inflación y la tasa de desempleo en Estados Unidos tomados de varias ediciones del informe *Economic Report of the President*, incluido el informe de 2004 (tablas B-42 y B-64).

Existe otra diferencia más sutil entre los datos de corte transversal y los datos de series de tiempo. En los capítulos 3 y 4 se estudiaron las propiedades estadísticas de los estimadores de MCO con base en la noción de que las muestras se extrajeron de forma aleatoria de la población adecuada. Entender por qué los datos de corte transversal deben verse como resultados aleatorios es muy sencillo: cada muestra diferente extraída de la población, por lo general, producirá valores distintos de las variables dependientes e independientes (como la educación, la experiencia, el salario, etc.). Por tanto, las estimaciones por MCO calculadas a partir de diferentes muestras aleatorias por lo general serán diferentes, razón por la cual se considera que los estimadores de MCO son variables aleatorias.

TABLA 10.1

Lista parcial de datos sobre la inflación y la tasa de desempleo en Estados Unidos, 1948-2003

Año	Inflación	Desempleo
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
. . .	. . .	. . .
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

¿Cómo se debe considerar la aleatoriedad en los datos de series de tiempo? Desde luego, las series de tiempo económicas satisfacen los requisitos intuitivos de ser resultados de variables aleatorias. Por ejemplo, hoy no se sabe cómo cerrará el promedio industrial Dow Jones al final de las operaciones de mañana. Ni tampoco cuál será el crecimiento anual de la producción en Canadá el año próximo. Dado que los resultados de estas variables no se conocen por anticipado, deben verse sin lugar a dudas como variables aleatorias.

En términos formales, a una secuencia de variables aleatorias indexadas en el tiempo se le llama **proceso estocástico** o **proceso de series de tiempo** ("estocástico" es sinónimo de aleatorio). Cuando se conforma una base de datos de series de tiempo, se obtiene un resultado posible, o *realización*, del proceso estocástico. Únicamente se puede ver una sola realización, ya que no es posible retroceder en el tiempo y empezar de nuevo el proceso. (Esto es análogo al análisis de corte transversal en el que únicamente se puede reunir una sola muestra aleatoria.) No obstante, si ciertas condiciones históricas fueran distintas, por lo general se obtendría una realización diferente para el proceso estocástico y es por ello que los datos de series de tiempo se consideran como el resultado de variables aleatorias. El conjunto de todas las realizaciones posibles de un proceso de series de tiempo desempeña el papel de la población en el análisis de corte transversal. El tamaño de muestra para una base de datos de series de tiempo es el número de periodos durante los cuales se observan las variables de interés.

10.2 Ejemplos de modelos de regresión con series de tiempo

En esta sección se estudian dos ejemplos de modelos de series de tiempo que han sido útiles en el análisis empírico y que se estiman con facilidad mediante mínimos cuadrados ordinarios. En el capítulo 11 se verán otros modelos.

Modelos estáticos

Imagine que tiene a su disposición datos de series de tiempo sobre dos variables, y y z, en las cuales y_t y z_t son contemporáneas. Un **modelo estático** que relaciona a y con z es

$$y_t = \beta_0 + \beta_1 z_t + u_t, t = 1, 2, \dots, n. \tag{10.1}$$

El nombre de "modelo estático" proviene del hecho de que se está representando una relación contemporánea entre y y z. Por lo común, un modelo de esta índole se postula cuando se considera que un cambio en z en el momento t ejerce un efecto inmediato sobre y: $\Delta y_t = \beta_1 \Delta z_t$, cuando $\Delta u_t = 0$. Los modelos de regresión estáticos también se usan cuando se tiene interés en conocer el intercambio entre y y z.

Un ejemplo de modelo estático es la *curva estática de Phillips*, dada por

$$inf_t = \beta_0 + \beta_1 unem_t + u_t, \tag{10.2}$$

donde inf_t es la tasa de inflación anual y $unem_t$ es la tasa de desempleo. Esta forma de la curva de Phillips supone una *tasa natural de desempleo* constante y expectativas inflacionarias constantes, y se emplea para estudiar el intercambio contemporáneo entre la inflación y el desempleo. [Vea, por ejemplo, Mankiw (1994, sección 11.2)].

Como es natural, podemos tener diversas variables explicativas en un modelo de regresión estático. Sean $mrdrte_t$ los homicidios por cada 10,000 personas en una determinada ciudad durante el año t, $convrte_t$ la tasa de condena por homicidio, $unem_t$ la tasa de desempleo local y $yngmle_t$ la fracción de la población que se compone de hombres entre 18 y 25 años de edad. De esta manera, un modelo estático de regresión múltiple que explica las tasas de homicidio es

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t. \tag{10.3}$$

Con ayuda de un modelo como éste esperamos estimar, por ejemplo, el efecto *ceteris paribus* de un incremento en la tasa de condenas para una actividad delictiva en particular.

Modelos de rezagos distribuidos finitos

En un **modelo de rezagos distribuidos finitos (RDF)** se permite que una o más variables influyan en y en forma rezagada. Por ejemplo, para las observaciones anuales, considere el modelo

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t, \tag{10.4}$$

donde gfr_t es la tasa de fertilidad general (niños nacidos por cada 1,000 mujeres en edad de concebir) y pe_t es el valor en dólares reales de la exención personal de impuestos. La idea es ver si, en conjunto, la decisión de tener hijos se asocia con el valor impositivo de tenerlos. La ecuación (10.4) reconoce que, tanto por motivos biológicos como conductuales, las decisiones de procrear no serían un resultado inmediato de los cambios en la exención personal.

La ecuación (10.4) es un ejemplo del modelo

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \tag{10.5}$$

el cual es un modelo de RDF *de orden dos*. Para interpretar los coeficientes en la ecuación (10.5), suponga que z es una constante, igual a c, en todos los periodos antes del periodo t. En t, z aumenta una unidad a $c + 1$, y después vuelve a su nivel anterior en el periodo $t + 1$. (Es decir, el incremento en z es temporal.) De manera más concreta,

$$\ldots, z_{t-2} = c, z_{t-1} = c, z_t = c + 1, z_{t+1} = c, z_{t+2} = c, \ldots.$$

Para concentrarse en el efecto *ceteris paribus* que z tiene sobre y, se fija el término de error de cada periodo en cero. Por consiguiente,

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,\\
y_t &= \alpha_0 + \delta_0(c + 1) + \delta_1 c + \delta_2 c,\\
y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1(c + 1) + \delta_2 c,\\
y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2(c + 1),\\
y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,
\end{aligned}$$

y así sucesivamente. De las dos primeras ecuaciones se obtiene, $y_t - y_{t-1} = \delta_0$, que demuestra que δ_0 es el cambio inmediato en y debido al aumento de una unidad en z en el periodo t. Por lo común, a δ_0 se le conoce como **propensión de impacto** o **multiplicador de impacto**.

Asimismo, $\delta_1 = y_{t+1} - y_{t-1}$ es el cambio en y un periodo después de la modificación temporal, y $\delta_2 = y_{t+2} - y_{t-1}$ es el cambio en y dos periodos después de la modificación. En el momento $t + 3$, y se ha revertido a su nivel inicial: $y_{t+3} = y_{t-1}$. Esto se debe a que se ha supuesto que sólo dos rezagos de z aparecen en la ecuación (10.5). Cuando se traza la gráfica de δ_j como una función de j, se obtiene la **distribución de rezagos**, la cual resume el efecto dinámico que un aumento temporal en z tiene sobre y. En la figura 10.1 se aprecia una posible distribución de rezagos para el modelo de RDF de orden dos. (Desde luego, nunca se conocerán los parámetros de δ_j; así que se estimará δ_j y luego se graficará la distribución de rezagos estimada.)

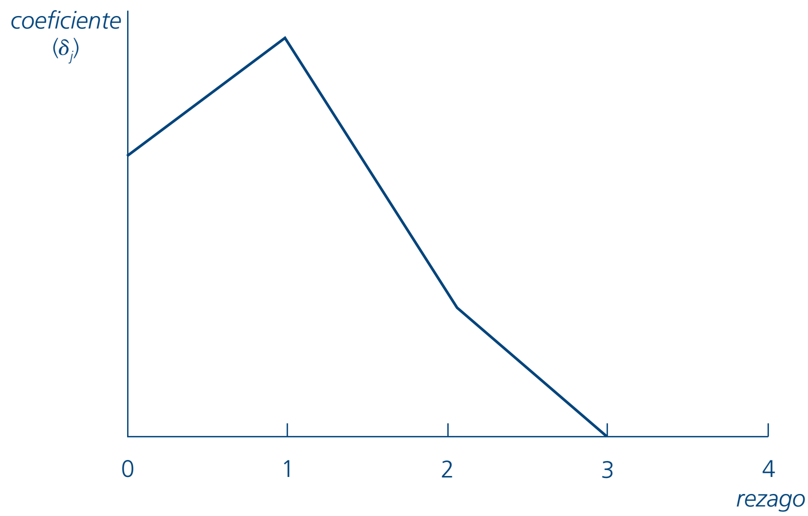
La distribución de rezagos de la figura 10.1 implica que el efecto más grande está en el primer rezago. Dicha distribución tiene una interpretación útil. Si se estandariza el valor inicial de y en $y_{t-1} = 0$, la distribución de rezagos determina todos los valores subsiguientes de y debido a un incremento temporal de una unidad en z.

También nos interesa el cambio en y debido a un incremento *permanente* en z. Antes del periodo t, z es igual a la constante c. En t, z aumenta de manera permanente a $c + 1$: $z_s = c$, $s < t$ y $z_s = c + 1$, $s \geq t$. Una vez más, al igualar los errores a cero, se tiene

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,\\
y_t &= \alpha_0 + \delta_0(c + 1) + \delta_1 c + \delta_2 c,\\
y_{t+1} &= \alpha_0 + \delta_0(c + 1) + \delta_1(c + 1) + \delta_2 c,\\
y_{t+2} &= \alpha_0 + \delta_0(c + 1) + \delta_1(c + 1) + \delta_2(c + 1),
\end{aligned}$$

FIGURA 10.1

Distribución de rezagos con dos rezagos distintos de cero. El efecto máximo está en el primer rezago.



y así sucesivamente. Con el aumento permanente en z, luego de un periodo, y ha aumentado en $\delta_0 + \delta_1$, y después de dos periodos, y se ha incrementado en $\delta_0 + \delta_1 + \delta_2$. No hay modificaciones adicionales en y al cabo de dos periodos. Esto muestra que la suma de los coeficientes de la z actual y las z rezagadas, $\delta_0 + \delta_1 + \delta_2$, es el cambio a *largo plazo* en y dado un incremento permanente en z y se denomina **propensión de largo plazo (PLP)** o **multiplicador de largo plazo**. Con frecuencia, la PLP es de interés en los modelos de rezagos distribuidos.

A manera de ejemplo, en la ecuación (10.4), δ_0 mide el cambio inmediato en la fertilidad debido a un aumento de un dólar en *pe*. Como se mencionó antes, hay razones para pensar que δ_0 es pequeña, si no es que cero. Sin embargo δ_1 o δ_2, o ambas, podrían ser positivas. Si *pe* aumenta un dólar *de manera permanente*, entonces al cabo de dos años, *gfr* habrá cambiado en $\delta_0 + \delta_1 + \delta_2$. Este modelo supone que no hay más modificaciones luego de dos años. Si esto es de este modo o no, se trata de un asunto empírico.

Un modelo de rezagos distribuidos finitos de orden q se escribe como

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \ldots + \delta_q z_{t-q} + u_t. \tag{10.6}$$

Éste contiene el modelo estático como un caso especial, al establecer que $\delta_1, \delta_2, \ldots, \delta_q$ son iguales a cero. En ocasiones, el propósito primordial de la estimación de un modelo de rezagos distribuidos es probar si z ejerce un efecto rezagado sobre y. El multiplicador de impacto siempre es el coeficiente de la z actual, δ_0. En ocasiones, se omite la z_t en la ecuación (10.6), en cuyo caso el multiplicador de impacto es cero. La distribución de rezagos es de nuevo la δ_j graficada como función de j. El multiplicador de largo plazo es la suma de todos los coeficientes de las variables z_{t-j}:

$$\text{PLP} = \delta_0 + \delta_1 + \ldots + \delta_q. \tag{10.7}$$

Por la correlación, a menudo sustancial, entre z y sus diferentes rezagos, es decir, debido a la multicolinealidad en (10.6), resulta difícil obtener estimadores precisos de cada δ_j individual. Curiosamente, aun cuando la δ_j no puede estimarse con precisión, a menudo se pueden obtener buenas estimaciones de la PLP. Más adelante se verá un ejemplo.

Es posible tener más de una variable explicativa que aparezca con rezagos o agregar variables contemporáneas a un modelo de RDF. Por ejemplo, el nivel de educación promedio para las mujeres en edad fértil se podría sumar a (10.4), lo que permite dar cuenta de los cambios en los grados de escolaridad de las mujeres.

Pregunta 10.1

En una ecuación para datos anuales, suponga que

$$int_t = 1.6 + .48\ inf_t - .15\ inf_{t-1} + .32\ inf_{t-2} + u_t,$$

donde *int* es una tasa de interés e *inf* es la tasa de inflación. ¿Cuáles son las propensiones de impacto y de largo plazo?

Una convención sobre el índice de tiempo

Cuando los modelos cuentan con variables explicativas rezagadas (y, como se verá en el siguiente capítulo, con y rezagadas), puede surgir la confusión respecto al tratamiento de las observaciones iniciales. Por ejemplo, si se supone que la ecuación (10.5) se mantiene comenzando en $t = 1$, entonces las variables explicativas del primer periodo son z_1, z_0 y z_{-1}. Por convención, éstos son los valores iniciales de nuestra muestra, de modo que siempre comienza el índice temporal en $t = 1$. En la práctica, esto no es muy importante, ya que los paquetes de regresión dan seguimiento de forma automática a las observaciones disponibles para la estimación de modelos con rezagos. Sin embargo, para éste y algunos de los capítulos siguientes, se necesitan algunas convenciones sobre el primer periodo que puede ser representado por la ecuación de la regresión.

10.3 Propiedades en muestras finitas de MCO bajo los supuestos clásicos

En esta sección se proporciona una lista completa de las propiedades en muestras finitas o muestras pequeñas de los MCO de acuerdo con los supuestos estándar. Se presta particular atención a la forma en que deben modificarse las premisas de nuestro análisis de corte transversal con el fin de cubrir las regresiones con series de tiempo.

Insesgamiento de MCO

El primer supuesto plantea sencillamente que el proceso de series de tiempo sigue un modelo que es lineal en sus parámetros.

Supuesto ST.1 (Lineal en los parámetros)

El proceso estocástico $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t): t = 1, 2, \ldots, n\}$ sigue el modelo lineal

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \quad (10.8)$$

donde $\{u_t: t = 1, 2, \ldots, n\}$ es una secuencia de errores o perturbaciones. Aquí, n es el número de observaciones (periodos).

En la notación x_{tj}, t denota el periodo y j es, como de costumbre, una etiqueta para señalar una de las k variables explicativas. En este punto se aplica la terminología empleada en la regresión de corte transversal: y_t es la variable dependiente, la variable explicada o el regresando; las x_{tj} son las variables independientes, variables explicativas o los regresores.

Debemos pensar en el supuesto ST.1 como si fuera idéntico en esencia a la premisa RLM.1 (el primer supuesto del corte transversal), sólo que ahora se está especificando un modelo lineal para datos de series de tiempo. Los ejemplos tratados en la sección 10.2 se pueden expresar en la forma de la ecuación (10.8) al definir x_{tj} de manera adecuada. Por ejemplo, la ecuación (10.5) se obtiene al establecer $x_{t1} = z_t$, $x_{t2} = z_{t-1}$ y $x_{t3} = z_{t-2}$.

Para plantear y analizar varios de los supuestos restantes, suponga que $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ denota el conjunto de todas las variables independientes en la ecuación en el momento t. Además, $\mathbf{X}$ denota la colección de todas las variables independientes para todos los periodos. Es útil considerar a $\mathbf{X}$ como una matriz, con n filas y k columnas, lo que refleja la manera en que se almacenan los datos de series de tiempo en el software de econometría: la t-ésima fila de $\mathbf{X}$ es $\mathbf{x}_t$, que consta de todas las variables independientes en el periodo t. Por tanto, la primera fila de $\mathbf{X}$ corresponde a $t = 1$, la segunda fila a $t = 2$ y la última a $t = n$. En la tabla 10.2 se proporciona un ejemplo que utiliza $n = 8$ y las variables explicativas de la ecuación (10.3).

Como es natural, al igual que con la regresión de corte transversal, es necesario descartar la colinealidad perfecta entre los regresores.

Supuesto ST.2 (No hay colinealidad perfecta)

En la muestra (y, por ende, en los procesos de series de tiempo subyacentes) no hay variables independientes que sean constantes ni que sean una combinación lineal perfecta de las otras.

TABLA 10.2

Ejemplo de X para las variables explicativas de la ecuación (10.3)

t	*convrte*	*unem*	*yngmle*
1	.46	.074	.12
2	.42	.071	.12
3	.42	.063	.11
4	.47	.062	.09
5	.48	.060	.10
6	.50	.059	.11
7	.55	.058	.12
8	.56	.059	.13

Este supuesto se estudió con detenimiento en el contexto de los datos de corte transversal en el capítulo 3. Los problemas son en esencia los mismos con los datos de series de tiempo.

Recuerde, el supuesto ST.2 permite que las variables explicativas se correlacionen, pero descarta la correlación *perfecta* en la muestra.

El supuesto final requerido para el insesgamiento de MCO es el equivalente en series de tiempo a la premisa RLM.4, y también obvia la necesidad de un muestreo aleatorio de la premisa RLM.2.

Supuesto ST.3 (Media condicional cero)

Para cada t, dadas las variables explicativas para *todos* los periodos, el valor esperado del error u_t es cero. Matemáticamente,

$$\mathrm{E}(u_t|\mathbf{X}) = 0,\ t = 1, 2, \ldots, n. \tag{10.9}$$

Este supuesto es crucial y es necesario tener una idea de su significado. Como en el caso del corte transversal, resulta mucho más sencillo ver este supuesto en función de la ausencia de correlación: el supuesto ST.3 implica que el error en el periodo t, u_t, no se correlaciona con ninguna variable explicativa en *cada* uno de los periodos. El hecho de que esto se plantee en términos de la esperanza condicional implica que se debe especificar de forma adecuada la relación funcional entre y_t y las variables explicativas. Si u_t es independiente de $\mathbf{X}$ y $\mathrm{E}(u_t) = 0$, el supuesto ST.3 es válido de manera automática.

Dado el análisis de corte transversal del capítulo 3, no es de sorprender que se necesite que u_t no esté correlacionada con las variables explicativas fechadas en el mismo periodo t: en términos de la media condicional,

$$\mathrm{E}(u_t|x_{t1}, \ldots, x_{tk}) = \mathrm{E}(u_t|\mathbf{x}_t) = 0. \tag{10.10}$$

Cuando la ecuación (10.10) es válida, se dice que las x_{tj} son **contemporáneamente exógenas**. La ecuación (10.10) implica que u_t y las variables explicativas no se correlacionan de manera contemporánea: $\mathrm{Corr}(x_{tj},u_t) = 0$, para toda j.

El supuesto ST.3 exige algo más que exogeneidad contemporánea: u_t no debe correlacionarse con x_{sj}, ni siquiera cuando $s \neq t$. Este es el sentido estricto en el que las variables explicativas deben ser exógenas, y cuando ST.3 es válido decimos que las variables explicativas son **estrictamente exógenas**. En el capítulo 11 se mostrará que la ecuación (10.10) es suficiente para probar la consistencia de los estimadores de MCO. Pero para demostrar que los estimadores de MCO son insesgados, se necesita el supuesto de exogeneidad estricta.

En el caso del corte transversal, no se estipula de manera explícita la forma en que el término de error, por ejemplo para la persona i, u_i, se correlaciona con las variables explicativas para *otras* personas de la muestra. Esto no fue necesario debido a que con el muestreo aleatorio (premisa RLM.2), u_i *automáticamente* es independiente de las variables explicativas en las observaciones distintas de i. En un contexto de series de tiempo, el muestreo aleatorio casi nunca es adecuado, de manera que se debe suponer de forma explícita que el valor esperado de u_t no se relaciona con las variables explicativas en ningún periodo.

Es importante tomar en cuenta que el supuesto ST.3 no pone restricciones a la correlación de las variables independientes o en u_t a lo largo del tiempo. El supuesto ST.3 sólo indica que el valor esperado de u_t no se relaciona con las variables independientes en todos los periodos.

Todo lo que provoque que los factores inobservables en el periodo t se correlacionen con alguna de las variables explicativas en cualquier periodo invalida el supuesto ST.3. Dos candidatos para esta invalidación son las variables omitidas y el error de medición en alguno de los regresores. Sin embargo, el supuesto de exogeneidad estricta también puede invalidarse por otras razones menos obvias. En el modelo de regresión estática simple

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

El supuesto ST.3 requiere no sólo que u_t y z_t no estén correlacionadas, sino que u_t tampoco se correlacione con valores pasados ni futuros de z. Esto tiene dos implicaciones. En primer lugar, z no puede tener un efecto rezagado sobre y. Si z tiene un efecto rezagado sobre y, entonces se debe estimar un modelo de rezagos distribuidos. Un punto más sutil es que la exogeneidad estricta excluye la posibilidad de que cambios en el término de error de hoy ocasionen modificaciones futuras en z. Esto descarta en forma eficaz la retroalimentación de y para valores futuros de z. Por ejemplo, considere un modelo estático simple para explicar la tasa de homicidios de una ciudad en términos de oficiales de policía *per cápita*:

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t.$$

Tal vez sea razonable suponer que u_t no se correlaciona con $polpc_t$ e incluso con valores pasados de $polpc_t$; suponga que lo anterior es cierto, pero que la ciudad ajusta el tamaño de su fuerza policiaca con base en los valores pasados de la tasa de homicidios. Esto significa que, $polpc_{t+1}$ se correlacionaría con u_t (ya que una variable u_t mayor conduce a una $mrdrte_t$ mayor). Si éste fuera el caso, por lo general se violaría el supuesto ST.3.

Hay consideraciones semejantes en los modelos de rezagos distribuidos. Por lo general, no preocupa que u_t se correlacione con la z anterior debido a que estamos controlando las z anteriores en el modelo. Pero la retroalimentación de u sobre las z futuras siempre representa un problema.

Las variables explicativas que son estrictamente exógenas no reaccionan a lo que ha sucedido con y en el pasado. Un factor como la cantidad de precipitación pluvial en una función de producción agrícola satisface este requisito: la precipitación en cualquier año en el futuro no se ve influida por la producción del año en curso o de los años anteriores. Sin embargo, una variable como la cantidad de trabajo podría no ser estrictamente exógena, ya que el agricultor es quien la elige y éste puede ajustarla con base en la producción del último año. Las variables de política, como el crecimiento en la masa monetaria, los gastos de prestaciones sociales y los límites de velocidad en las autopistas, a menudo se ven influidos por lo que ha sucedido en el pasado. En las ciencias sociales, muchas variables explicativas bien pueden infringir la premisa de exogeneidad estricta.

Aun cuando el supuesto ST.3 sea poco realista, partimos de éste para llegar a la conclusión de que los estimadores de MCO son insesgados. La mayor parte de los tratamientos de los modelos estáticos y de los rezagos distribuidos dan por hecho ST.3 al considerar el supuesto más estricto de que las variables explicativas no son aleatorias o son fijas en muestras repetidas. El supuesto de no aleatoriedad desde luego es falso para las observaciones de series de tiempo; el supuesto ST.3 cuenta con la ventaja de ser más realista respecto a la naturaleza aleatoria de las x_{tj}, mientras que aísla la premisa necesaria respecto a cómo se relacionan u_t y las variables explicativas para que los estimadores de MCO sean insesgados.

Teorema 10.1 (Insesgamiento de los estimadores de MCO)

Bajo los supuestos ST.1, ST.2 y ST.3, los estimadores de MCO son insesgados condicionales sobre **X**, y por tanto también incondicionalmente: $E(\hat{\beta}_j) = \beta_j$, $j = 0, 1, \ldots, k$.

Pregunta 10.2

En el modelo $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$ de RDF, ¿qué necesitamos suponer sobre la secuencia $\{z_0, z_1, \ldots, z_n\}$ para que el supuesto ST.3 sea válido?

La demostración de este teorema es, en esencia, la misma que la del teorema 3.1 del capítulo 3, y por ende se omite aquí. Cuando se comparó el teorema 10.1 con el 3.1, pudimos eliminar la premisa del muestreo aleatorio al suponer que, para cada t, u_t tiene una media de cero dadas las variables explicativas en todos los periodos. Si esta suposición no es válida, no se puede demostrar que los MCO son insesgados.

El análisis del sesgo de las variables omitidas que se vio en la sección 3.3 es en esencia el mismo en el caso de las series de tiempo. En concreto, la tabla 3.2 y el análisis que la acompaña se pueden emplear como antes se hizo para determinar las direcciones del sesgo debidas a las variables omitidas.

Las varianzas de los estimadores de MCO y el teorema de Gauss-Markov

Es necesario agregar dos supuestos para complementar las premisas de Gauss-Markov para las regresiones con series de tiempo. La primera resulta familiar gracias al análisis de corte transversal.

Supuesto ST.4 (Homocedasticidad)

La varianza de u_t condicional en $\mathbf{X}$, es la misma para cualquier t: $\text{Var}(u_t|\mathbf{X}) = \text{Var}(u_t) = \sigma^2$, $t = 1, 2, \ldots, n$.

Este supuesto significa que $\text{Var}(u_t|\mathbf{X})$ no puede depender de $\mathbf{X}$ —es suficiente con que u_t y $\mathbf{X}$ sean independientes— *además* que $\text{Var}(u_t)$ debe ser constante en el tiempo. Cuando el supuesto ST.4 no es válido, se dice que los errores son *heterocedásticos*, como en el caso del corte transversal. Por ejemplo, considere una ecuación para determinar las tasas de las letras del Tesoro estadounidense a tres meses ($i3_t$) con base en la tasa de inflación (inf_t) y en el déficit federal como porcentaje del producto interno bruto (def_t):

$$i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t. \tag{10.11}$$

Entre otras cosas, el supuesto ST.4 exige que los factores inobservables que influyen en las tasas de interés tengan una varianza constante respecto al tiempo. Como se sabe que los cambios en el régimen político influyen en la variabilidad de las tasas de interés, este supuesto bien puede ser falso. Además, tal vez la variabilidad de las tasas de interés dependa del nivel de la inflación o del tamaño relativo del déficit. Esto también violaría la premisa de la homocedasticidad.

Cuando $\text{Var}(u_t|\mathbf{X})$ depende de $\mathbf{X}$, a menudo está subordinada a las variables explicativas en el periodo t, $\mathbf{x}_t$. En el capítulo 12 se verá que las pruebas de heterocedasticidad del capítulo 8 también se emplean para las regresiones con series de tiempo, por lo menos bajo ciertos supuestos.

El supuesto final de Gauss-Markov para el análisis de series de tiempo es nuevo.

Supuesto ST.5 (No hay correlación serial)

Los errores, condicionales sobre $\mathbf{X}$, en dos periodos distintos, no están correlacionados: $\text{Corr}(u_t, u_s|\mathbf{X}) = 0$, para cualquier $t \neq s$.

La forma más sencilla de considerar este supuesto es ignorar el condicionamiento sobre $\mathbf{X}$. Así, el supuesto ST.5 es sencillamente

$$\text{Corr}(u_t,u_s) = 0, \text{ para cualquier } t \neq s. \tag{10.12}$$

(Así es como se plantea el supuesto de no correlación serial cuando $\mathbf{X}$ se trata como no aleatoria.) Cuando se considera que el supuesto ST.5 tiene posibilidades de ser válido, nos concentramos en la ecuación (10.12) debido a que es fácil interpretarla.

Cuando (10.12) es falsa, se dice que los errores en (10.8) tienen **correlación serial** o **autocorrelación**, debido a que se correlacionan a lo largo del tiempo. Considere el caso de los errores de periodos adyacentes. Suponga que cuando $u_{t-1} > 0$ entonces, en promedio, el error en el siguiente periodo, u_t, también es positivo. Por tanto, $\text{Corr}(u_t,u_{t-1}) > 0$, y los errores sufren de correlación serial. Esto quiere decir, en la ecuación (10.11), que si las tasas de interés aumentan de manera abrupta en este periodo, es probable que en el siguiente estén por encima del promedio (para los niveles dados de inflación y déficit). Esta caracterización resulta ser razonable para los términos de error en muchas aplicaciones de series de tiempo, las cuales se verán en el capítulo 12. Por ahora, damos por sentado el supuesto ST.5.

Es importante destacar que el ST.5 no supone nada sobre la correlación temporal de las variables *independientes*. Por ejemplo, en la ecuación (10.11) es casi seguro que inf_t esté correlacionada en el tiempo. Pero esto no tiene nada que ver con la validez del ST.5.

Una pregunta natural es: ¿por qué en los capítulos 3 y 4 no supusimos que los errores para diferentes observaciones de corte transversal no están correlacionados? La respuesta proviene del supuesto de muestreo aleatorio: en el muestreo aleatorio, u_i y u_h son independientes para cualesquiera dos observaciones i y h. También se demuestra que bajo el muestreo aleatorio los errores para las diferentes observaciones, condicionales sobre las variables explicativas de la muestra, son independientes. Así, para nuestros propósitos, consideramos que la correlación serial sólo es un problema en las regresiones con series de tiempo. (En los capítulos 13 y 14, el problema de la correlación serial se presentará en el análisis de datos de panel.)

Los supuestos ST.1 a ST.5 son los adecuados de Gauss-Markov para las aplicaciones de series de tiempo, aunque también tienen otros usos. A veces los supuestos ST.1 a ST.5 se satisfacen en las aplicaciones de corte transversal, aun cuando el muestreo aleatorio no sea un supuesto razonable, como cuando las unidades de corte transversal son grandes en relación con la población. Imagine que tiene una base de datos de corte transversal en el ámbito citadino. Podría suceder que exista una correlación en algunas de las variables explicativas entre las ciudades de un mismo estado, como el impuesto predial o los pagos de seguridad social *per cápita*. La correlación de las variables explicativas entre las observaciones no genera problemas para verificar los supuestos de Gauss-Markov, siempre y cuando los errores no se correlacionen entre estas ciudades. Sin embargo, en este capítulo nuestro principal interés es la aplicación de los supuestos de Gauss-Markov a los modelos de regresión con datos de series de tiempo.

Teorema 10.2 (Varianzas de muestreo de los estimadores de MCO)

Con base en los supuestos ST.1 a ST.5 de Gauss-Markov para las series de tiempo, la varianza de $\hat{\beta}_j$, condicional sobre **X**, es

$$\text{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2/[\text{STC}_j(1 - R_j^2)], \; j = 1, \ldots, k, \qquad \textbf{10.13}$$

donde STC_j es la suma total de cuadrados de x_{tj} y R_j^2 es la R-cuadrada de la regresión de x_j sobre las otras variables independientes.

La ecuación (10.13) es la misma varianza que se determinó en el capítulo 3 bajo los supuestos de Gauss-Markov para corte transversal. Como la demostración es muy parecida a la del teorema 3.2, aquí se omite. El planteamiento del capítulo 3 sobre los factores que generan grandes varianzas, incluida la multicolinealidad entre las variables explicativas, se aplica de inmediato al caso de las series de tiempo.

El estimador usual de la varianza del error también es insesgado bajo los supuestos ST.1 a ST.5 y el teorema de Gauss-Markov es válido.

Teorema 10.3 (Estimación insesgada de σ^2)

Bajo los supuestos ST.1 a ST.5, el estimador $\hat{\sigma}^2 = \text{SRC}/gl$ es un estimador insesgado de σ^2, donde $gl = n - k - 1$.

Teorema 10.4 (Teorema de Gauss-Markov)

Bajo los supuestos ST.1 a ST.5, los estimadores de MCO son los mejores estimadores lineales insesgados condicionales sobre **X**.

Aquí lo esencial es que MCO tienen las mismas propiedades en muestras finitas bajo ST.1 a ST.5 que aquellas que posee bajo RLM.1 a RLM.5.

Pregunta 10.3

En el modelo de RDF $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, explique la naturaleza de cualquier clase de multicolinealidad en las variables explicativas.

Inferencia bajo los supuestos del modelo lineal clásico

Para utilizar los errores estándar usuales de MCO, los estadísticos t y F, se necesita agregar un último supuesto que es análogo a la premisa de normalidad empleada en el análisis de corte transversal.

Supuesto ST.6 (Normalidad)

Los errores u_t son independientes de **X** y son independientes e idénticamente distribuidos como Normal$(0,\sigma^2)$.

El supuesto ST.6 comprende los supuestos ST.3, ST.4 y ST.5, pero es más fuerte debido a los supuestos de independencia y normalidad.

Teorema 10.5 (Distribuciones de muestreo normales)

Bajo los supuestos ST.1 a ST.6, los supuestos del MCL para series de tiempo, los estimadores de MCO se distribuyen de forma normal, condicionales sobre **X**. Además, bajo la hipótesis nula, cada estadístico t tiene una distribución t y cada estadístico F tiene una distribución F. También es válida la construcción usual de los intervalos de confianza.

Las consecuencias del teorema 10.5 son de suma importancia. Implican que, cuando los supuestos ST.1 a ST.6 son válidos, todo lo aprendido acerca de la estimación y la inferencia para las regresiones con cortes transversales se aplica de manera directa a las regresiones con series de tiempo. Así pues, los estadísticos t pueden emplearse para probar la significancia estadística de las variables explicativas individuales y los estadísticos F pueden utilizarse para probar la significancia conjunta.

Como en el caso del corte transversal, los usuales procedimientos de inferencia están supeditados a los supuestos subyacentes. Los supuestos del modelo lineal clásico para los datos de series de tiempo son mucho más restrictivos que aquellos de corte transversal —en particular, los supuestos, de exogeneidad estricta y no correlación serial pueden resultar poco realistas—. Con todo, el esquema del MLC es un buen punto de partida para muchas aplicaciones.

Ejemplo 10.1

[Curva de Phillips estática]

Para determinar si hay un efecto de sustitución, en promedio, entre el desempleo y la inflación, probamos H_0: $\beta_1 = 0$ contra H_1: $\beta_1 < 0$ en la ecuación (10.2). Si los supuestos del modelo lineal clásico son válidos, podemos utilizar el estadístico *t* usual de MCO.

Se emplea la base de datos PHILLIPS.RAW para estimar la ecuación (10.2), limitándose a los datos hasta 1996. (En ejercicios posteriores, por ejemplo en los ejercicios para computadora C10.12 y C11.10, se le pide que utilice todos los datos hasta 2003. En el capítulo 18, se usan los años 1997 a 2003 en varios ejercicios de pronóstico.) Las estimaciones de la regresión simple son

$$\begin{aligned} \widehat{inf}_t &= 1.42 + .468\, unem_t \\ &\quad\;(1.72)\;\;(.289) \\ n &= 49,\; R^2 = .053,\; \bar{R}^2 = .033. \end{aligned} \tag{10.14}$$

Esta ecuación no sugiere un efecto de sustitución entre *unem* (desempleo) e *inf* (inflación): $\hat{\beta}_1 > 0$. El estadístico *t* para $\hat{\beta}_1$ es cerca de 1.62, lo que da un valor-*p* en contra de la alternativa de dos colas de alrededor de .11. Así, en todo caso, hay una relación positiva entre la inflación y el desempleo.

Existen ciertos problemas con este análisis que no se pueden abordar ahora con detalle. En el capítulo 12 se verá que los supuestos del MLC no son válidos. Además, la curva de Phillips estática quizá no sea el mejor modelo para determinar si hay un efecto sustitución a corto plazo entre la inflación y el desempleo. Los macroeconomistas prefieren en general la curva de Phillips aumentada por las expectativas, de la cual se ofrece un sencillo ejemplo en el capítulo 11.

A manera de segundo ejemplo, se estima la ecuación (10.11) con ayuda de datos anuales de la economía estadounidense.

Ejemplo 10.2

[Efectos de la inflación y el déficit en las tasas de interés]

Los datos de INTDEF.RAW provienen del informe *Economic Report of the President* (tablas B-73 y B-79) y abarcan los años desde 1948 hasta 2003. La variable *i3* es la tasa de las letras del Tesoro estadounidense a tres meses, *inf* es la tasa de inflación anual basada en el índice de precios al consumidor (IPC) y *def* es el déficit en el presupuesto federal como porcentaje del producto interno bruto. La ecuación estimada es

$$\begin{aligned} \widehat{i3}_t &= 1.73 + .606\, inf_t + .513\, def_t \\ &\quad\;(0.43)\;\;(.082)\qquad\;(.118) \\ n &= 56,\; R^2 = .602,\; \bar{R}^2 = .587. \end{aligned} \tag{10.15}$$

Estas estimaciones muestran que los incrementos en la inflación o el tamaño relativo del déficit aumentan las tasas de interés a corto plazo, lo cual se espera a partir de la economía básica. Por ejemplo, un aumento *ceteris paribus* de un punto porcentual en la tasa de inflación aumenta *i3* en .606 puntos. Tanto *inf* como *def* son muy significativas estadísticamente, suponiendo desde luego que los supuestos del MLC son válidos.

10.4 Forma funcional, variables binarias y números índice

Todas las formas funcionales aprendidas en capítulos anteriores se utilizan en las regresiones con series de tiempo. La más importante es el logaritmo natural: en el trabajo aplicado con frecuencia aparecen regresiones con series de tiempo que tienen efectos porcentuales constantes.

Ejemplo 10.3

[Empleo y salario mínimo en Puerto Rico]

Castillo-Freeman y Freeman (1992) utilizan datos anuales sobre la tasa de empleo en Puerto Rico, el salario mínimo y otras variables para estudiar los efectos que tiene el salario mínimo estadounidense en el empleo en Puerto Rico. Una versión simplificada de su modelo es

$$\log(prepop_t) = \beta_0 + \beta_1 \log(mincov_t) + \beta_2 \log(usgnp_t) + u_t, \tag{10.16}$$

en donde $prepop_t$ es la tasa de empleo en Puerto Rico durante el año t (proporción de las personas que trabajan respecto a la población total), $usgnp_t$ es el producto nacional bruto real de Estados Unidos (en miles de millones de dólares) y *mincov* mide la importancia del salario mínimo en relación con los salarios promedio. En particular, $mincov = (avgmin/avgwage) \cdot avgcov$, donde *avgmin* es el salario mínimo promedio, *avgwage* es el salario general promedio y *avgcov* es la tasa de cobertura promedio (la proporción real de trabajadores cubiertos bajo la ley del salario mínimo).

Usando la base de datos PRMINWGE.RAW para los años de 1950 a 1987 se obtiene

$$\begin{aligned} \widehat{\log(prepop_t)} &= \underset{(0.77)}{-1.05} - \underset{(.065)}{.154} \log(mincov_t) - \underset{(.089)}{.012} \log(usgnp_t) \\ n &= 38,\ R^2 = .661,\ \bar{R}^2 = .641. \end{aligned} \tag{10.17}$$

La elasticidad estimada de *prepop* respecto a *mincov* es $-.154$, y es estadísticamente significativa con $t = -2.37$. Por consiguiente, un salario mínimo mayor disminuye la tasa de empleo, hecho que la economía clásica predice. La variable del PNB no resulta estadísticamente significativa, pero esto cambiará cuando se considere una tendencia en el tiempo en la siguiente sección.

Podemos emplear también formas logarítmicas funcionales en modelos de rezagos distribuidos. Por ejemplo, para los datos trimestrales, suponga que la demanda de dinero (M_t) se relaciona con el producto interno bruto (GDP_t) mediante

$$\begin{aligned} \log(M_t) = \alpha_0 &+ \delta_0 \log(GDP_t) + \delta_1 \log(GDP_{t-1}) + \delta_2 \log(GDP_{t-2}) \\ &+ \delta_3 \log(GDP_{t-3}) + \delta_4 \log(GDP_{t-4}) + u_t. \end{aligned}$$

La propensión de impacto en esta ecuación, δ_0, también se conoce como la **elasticidad de corto plazo**: mide el cambio porcentual inmediato en la demanda de dinero cuando hay un aumento de 1% en el *GDP*. La propensión de largo plazo, $\delta_0 + \delta_1 + \ldots + \delta_4$, con frecuencia se denomina la **elasticidad de largo plazo**: mide el incremento porcentual en la demanda de dinero luego de cuatro trimestres cuando hay un aumento permanente de 1% en el *GDP*.

Las variables binarias también son muy útiles en las aplicaciones de series de tiempo. Dado que la unidad de observación es el tiempo, una variable binaria representa si, en cada periodo, ha ocurrido un evento determinado. Por ejemplo, para los datos anuales podemos indicar en cada año si el presidente de Estados Unidos es un demócrata o un republicano con sólo definir una variable $democ_t$, igual a uno cuando el presidente sea demócrata e igual a cero cuando no lo sea. O bien, al considerar los efectos de la pena de muerte en las tasas de homicidio en Texas, podemos definir una variable binaria para cada año que sea igual a uno si se aplicó la pena de muerte en Texas durante ese año y a cero en caso contrario.

Con frecuencia, las variables binarias se emplean para aislar ciertos periodos que pueden ser sistemáticamente distintos de otros lapsos comprendidos en una base de datos.

Ejemplo 10.4

[Efectos de la exención personal en las tasas de fertilidad]

La tasa de fertilidad general (*gfr*) es el número de niños nacidos por cada 1,000 mujeres en edad de concebir. Para los años de 1913 a 1984, la ecuación,

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t,$$

explica *gfr* en términos del valor en dólares reales de la exención personal de impuestos promedio (*pe*) y dos variables binarias. La variable *ww2* asume el valor de uno para los años de 1941 a 1945, cuando Estados Unidos se involucró en la Segunda Guerra Mundial. La variable *pill* es uno de 1963 en adelante, cuando salió al mercado la píldora anticonceptiva para el control de la natalidad.

Al usar los datos de FERTIL3.RAW, que se obtuvieron del artículo de Whittington, Alm y Peters (1990), se obtiene

$$\begin{aligned} \widehat{gfr}_t &= \underset{(3.21)}{98.68} + \underset{(.030)}{.083}\, pe_t - \underset{(7.46)}{24.24}\, ww2_t - \underset{(4.08)}{31.59}\, pill_t \\ n &= 72,\ R^2 = .473,\ \bar{R}^2 = .450. \end{aligned} \tag{10.18}$$

Cada variable es estadísticamente significativa al nivel de 1% contra la alternativa de dos colas. Se ve que la tasa de fertilidad fue menor durante la Segunda Guerra Mundial: dada *pe*, hubo alrededor de 24 nacimientos menos por cada 1,000 mujeres en edad de procrear, lo cual es una gran reducción. (De 1913 a 1984, *gfr* osciló entre 65 y 127.) De igual modo, la tasa de fertilidad disminuyó de manera sustancial a partir de la introducción de la píldora anticonceptiva.

La variable de interés económico es *pe*. La *pe* media en este periodo es de \$100.40 dólares y oscila entre cero y \$243.83 dólares. El coeficiente de *pe* implica que un aumento de 12 dólares en *pe* provoca un incremento en *gfr* cercano a un nacimiento por cada 1,000 mujeres en edad de concebir. Este efecto no es poca cosa.

En la sección 10.2 se advirtió que la tasa de fertilidad reacciona a los cambios en *pe* en forma rezagada. La estimación de un modelo de rezagos distribuidos con dos rezagos nos da

$$\begin{aligned} \widehat{gfr}_t &= \underset{(3.28)}{95.87} + \underset{(.126)}{.073}\, pe_t - \underset{(.1557)}{.0058}\, pe_{t-1} + \underset{(.126)}{.034}\, pe_{t-2} \\ &\quad - \underset{(10.73)}{22.13}\, ww2_t - \underset{(3.98)}{31.30}\, pill_t \\ n &= 70,\ R^2 = .499,\ \bar{R}^2 = .459. \end{aligned} \tag{10.19}$$

En esta regresión sólo se cuenta con 70 observaciones, ya que se perdieron dos al rezagar *pe* dos veces. Los coeficientes en las variables *pe* se estiman de forma muy imprecisa y cada uno por separado es insignificante. Resulta que hay una alta correlación entre pe_t, pe_{t-1} y pe_{t-2}, y esta multicolinealidad dificulta la estimación del efecto de cada rezago. Sin embargo, pe_t, pe_{t-1} y pe_{t-2} en conjunto son significativas: el estadístico F tiene un valor-p = .012. Así, *pe* ejerce un efecto en *gfr* [como se vio en la ecuación (10.18)], pero no contamos con buenas estimaciones para determinar si es contemporáneo o tiene un rezago de uno o dos años (o algo de todo). En realidad, pe_{t-1} y pe_{t-2} son conjuntamente insignificantes en esta ecuación (valor-p = .95), de modo que en este momento se justificaría que utilizáramos el modelo estático. Pero para ilustrarlo mejor, obtenga un intervalo de confianza para la propensión de largo plazo de este modelo.

La PLP estimada en la ecuación (10.19) es .073 − .0058 + .034 ≈ .101. Sin embargo, no contamos con información suficiente en (10.19) para obtener el error estándar de esta estimación. Con el fin de obtener el error estándar de la PLP estimada, se usa el truco propuesto en la sección 4.4. Sea $\theta_0 = \delta_0 + \delta_1 + \delta_2$ la PLP y escriba δ_0 en términos de θ_0, δ_1 y δ_2 como $\delta_0 = \theta_0 - \delta_1 - \delta_2$. Ahora sustituya δ_0 en el modelo

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \ldots$$

para obtener

$$\begin{aligned} gfr_t &= \alpha_0 + (\theta_0 - \delta_1 - \delta_2)pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \ldots \\ &= \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + \ldots . \end{aligned}$$

De esta última ecuación se obtiene $\hat{\theta}_0$ y su error estándar al hacer la regresión de gfr_t sobre pe_t, $(pe_{t-1} - pe_t)$, $(pe_{t-2} - pe_t)$, $ww2_t$ y $pill_t$. El coeficiente y el error estándar asociado con pe_t son lo que necesitamos. Al realizar esta regresión se obtiene $\hat{\theta}_0$ = .101 como el coeficiente de pe_t (lo que ya se conocía) y ee($\hat{\theta}_0$) = .030 [que no podíamos calcular a partir de (10.19)]. Por tanto, el estadístico t para $\hat{\theta}_0$ es cerca de 3.37, de manera que $\hat{\theta}_0$ es estadísticamente distinto de cero a niveles de significancia pequeños. Aun cuando ninguna de las $\hat{\delta}_j$ tiene significancia por separado, la PLP es muy significativa. El intervalo de confianza de 95% para la PLP es de entre .041 y .160 aproximadamente.

Whittington, Alm y Peters (1990) permiten más rezagos, pero limitan los coeficientes para ayudar a paliar el problema de multicolinealidad que dificulta la estimación de las δ_j individuales. (El problema 10.6 proporciona un ejemplo de cómo llevarlo a cabo.) Para estimar la PLP, lo cual parecería ser de interés primordial aquí, tales restricciones son innecesarias. Whittington, Alm y Peters también controlan variables adicionales, como el salario promedio de las mujeres y la tasa de desempleo.

Las variables binarias explicativas son el componente clave en lo que se denomina **estudio de evento**. En un estudio de este tipo, el objetivo es ver si un determinado evento influye en algún resultado. Los economistas que estudian organización industrial han investigado los efectos de ciertos eventos sobre los precios de las acciones de las empresas. Por ejemplo, Rose (1985) estudió los efectos que tienen las nuevas reglamentaciones del transporte por carretera sobre los precios de las acciones de las compañías del ramo.

Una versión sencilla de una ecuación utilizada para tales estudios de eventos es

$$R_t^f = \beta_0 + \beta_1 R_t^m + \beta_2 d_t + u_t,$$

donde R_t^f es el rendimiento de las acciones de la empresa f durante el periodo t (por lo común una semana o un mes), R_t^m es el rendimiento del mercado (que por lo general se calcula para un amplio índice bursátil), y d_t es una variable binaria que indica cuándo ocurrió el evento. Por ejemplo, si la empresa es una aerolínea, d_t podría denotar si la empresa sufrió un accidente públicamente conocido, o estuvo próxima a uno, durante la semana t. Incluir R_t^m en la ecuación controla la posibilidad de que amplios movimientos del mercado pudieran coincidir con los accidentes de la aerolínea. A veces se emplean múltiples variables binarias. Por ejemplo, si el evento es la imposición de una nueva reglamentación que pudiese influir en una empresa determinada, se incluiría una variable binaria que abarcara unas cuantas semanas antes del anuncio público de la reglamentación y una segunda variable para unas semanas después de dicho anuncio. La primera variable podría detectar la presencia de información interna.

Antes de dar un ejemplo de un estudio de evento, es necesario formular la noción de un **número índice** y la diferencia entre las variables económicas nominales y las reales. Un número índice por lo general agrega un gran volumen de información a una sola cantidad. Las cifras de esta índole se emplean por lo general en el análisis de series de tiempo, en especial en aplicaciones macroeconómicas. Un ejemplo de número índice es el índice de la producción industrial (IPI), calculado cada mes por el consejo directivo de la Reserva Federal. El IPI es una medida de producción de un amplio espectro de industrias y, como tal, su magnitud en un determinado año no tiene un sentido cuantitativo. Para interpretar la magnitud del IPI se debe conocer el **periodo base** y el **valor base**. En el informe 1997 *Economic Report of the President* (*ERP*) de 1997, el año base es 1987 y el valor base es 100. (Fijar el valor del IPI en 100 en el periodo base es sólo una convención; tiene tanto sentido como establecer IPI = 1 en 1987, y algunos índices se definen con un valor base de 1.) Como el IPI fue de 107.7 en 1992, se puede decir que la producción industrial fue 7.7% mayor en 1992 que en 1987. Se puede utilizar el IPI de cualquier bienio para calcular la diferencia porcentual en la producción industrial durante ese lapso. Por ejemplo, como IPI = 61.4 en 1970 e IPI = 85.7 en 1979, la producción industrial creció cerca de 39.6% durante la década de los setenta.

Es fácil modificar el periodo base de cualquier número índice y, a veces, se debe hacer para dar un año base común a los números índices reportados con años base distintos. Por ejemplo, si se desea cambiar el año base del IPI de 1987 a 1982, sólo se divide el IPI para cada año entre el valor de 1982 y luego se multiplica por 100 para asignar el valor de 100 al periodo base. En general, la fórmula es

$$nuevoíndice_t = 100(viejoíndice_t/viejoíndice_{\text{nuevabase}}), \tag{10.20}$$

donde $viejoíndice_{\text{nuevabase}}$ es el valor original del índice en el año base nuevo. Por ejemplo, con el año base 1987, el IPI en 1992 es de 107.7; si se modifica el año base a 1982, el IPI en 1992 se vuelve 100(107.7/81.9) = 131.5 (ya que el IPI en 1982 fue 81.9).

Otro ejemplo importante de un número índice es el *índice de precios*, como el Índice de Precios al Consumidor (IPC). Ya se ha usado el IPC para calcular las tasas de inflación anuales en el ejemplo 10.1. Al igual que con el índice de producción industrial, el IPC sólo es importante cuando se compara a lo largo de varios años (o meses, si se tienen datos mensuales). En el *ERP* de 1997, se tiene IPC = 38.8 en 1970 y en 1990 IPC = 130.7. De esta manera, el nivel general de precios creció cerca de 237% en este periodo de 20 años. (En el *ERP* de 1997, el IPC se define de tal manera que su promedio en 1982, 1983 y 1984 es igual a 100; por tanto, el periodo base figura como 1982-1984.)

Además de utilizarse para calcular las tasas de inflación, los índices de precios son necesarios para convertir una serie de tiempo medida en *dólares nominales* (o *dólares corrientes*) en *dólares reales* (o *dólares constantes*). Se supone que la mayor parte del comportamiento económico se ve influido por las variables reales y no por las nominales. Por ejemplo, la economía laboral clásica supone que la oferta de trabajo se basa en el salario real por hora y no en el nominal. Obtener el salario real a partir del nominal resulta sencillo si contamos con un índice de precios como el IPC. Se debe procurar dividir primero el IPC entre 100, para que el valor en el año base sea 1. Así, si w denota el salario medio por hora en dólares nominales y p = IPC/100, el *salario real* es simplemente w/p. Este salario se mide en dólares del periodo base del IPC.

Por ejemplo, en la tabla B-45 del *ERP* de 1997, los ingresos por hora promedio se reportan en términos nominales y en dólares de 1982 (lo que significa que el IPC utilizado para calcular el salario real tuvo como año base 1982). Esta tabla informa que el salario nominal por hora en 1960 fue de $2.09, pero medido en dólares de 1982, el salario fue de $6.79. El salario por hora real alcanzó su punto máximo en 1973, en $8.55 de 1982, y disminuyó a $7.40 en 1995. Así pues, hubo una disminución importante en los salarios reales durante esos 20 años. (Si se comparan los salarios nominales de 1973 y 1995, se obtiene un panorama muy engañoso: $3.94 en 1973 y $11.44 en 1995. En virtud de que el salario real bajó, el aumento en el salario nominal se debe por completo a la inflación.)

Las medidas estándar de las variables económicas se proporcionan en términos reales. La más importante de estas medidas es el *producto interno bruto* o *GDP*. Cuando la prensa popular informa el crecimiento del *GDP*, casi siempre es un crecimiento del *GDP real*. En el *ERP* de 1997 (tabla B-9), se reporta el *GDP* en miles de millones de dólares de 1992. En el ejemplo 10.3 se usó una medida similar de producción, el producto nacional bruto real.

Suceden cosas interesantes cuando las variables en dólares reales se utilizan en combinación con los logaritmos naturales. Suponga, por ejemplo, que el promedio de horas semanales trabajadas se relaciona con el salario real mediante

$$\log(hours) = \beta_0 + \beta_1 \log(w/p) + u.$$

Considerando el hecho de que $\log(w/p) = \log(w) - \log(p)$, se puede escribir esto como

$$\log(hours) = \beta_0 + \beta_1 \log(w) + \beta_2 \log(p) + u, \tag{10.21}$$

con la restricción de que $\beta_2 = -\beta_1$. Por tanto, el supuesto de que sólo el salario real influye en la oferta de trabajo impone una restricción en los parámetros del modelo (10.21). Si $\beta_2 \neq -\beta_1$, entonces el nivel de precios ejerce un efecto sobre la oferta de trabajo, algo que puede suceder si los trabajadores no entienden muy bien la distinción entre salarios reales y nominales.

Hay muchos aspectos prácticos para el cálculo de los números índice, pero llevaría mucho explicarlos aquí. La mayor parte de los libros de macroeconomía intermedia, como el de Mankiw (1994, capítulo 2), contienen análisis detallados sobre los índices de precios. Lo que aquí importa es saber aplicar los números índice en el análisis de regresión. Como se mencionó, dado que las magnitudes de los números índice no son en especial informativas, a menudo aparecen en forma logarítmica, de modo que los coeficientes de regresión se interpretan en términos de cambio porcentual.

Ahora se dará un ejemplo de un estudio de evento en el que también se aplican números índice.

Ejemplo 10.5

[Demandas *antidumping* e importaciones químicas]

Krupp y Pollard (1996) analizaron los efectos de las demandas *antidumping* de las industrias químicas estadounidenses sobre la importación de diversas sustancias químicas. Aquí se considera en particular una sustancia química industrial, el cloruro de bario, un agente limpiador utilizado en diversos procesos químicos y en la producción de gasolina. Los datos provienen del archivo BARIUM.RAW. A principios de la década de los ochenta, los productores estadounidenses de cloruro de bario consideraron que China estaba ofreciendo sus exportaciones a Estados Unidos a un precio injustamente bajo (acción conocida como *dumping*) y la industria del cloruro de bario interpuso una demanda ante la Comisión Estadounidense de Comercio Internacional (CCI) en octubre de 1983. La CCI falló a favor de la industria del cloruro de bario estadounidense en octubre de 1984. Hay varias interrogantes que resultan interesantes en este caso, pero sólo se mencionarán algunas de ellas. En primer lugar, ¿las importaciones son insólitamente altas en el periodo inmediato anterior a la demanda inicial? En segundo lugar, ¿las importaciones cambian de manera notable después de una demanda *antidumping*? Y por último, ¿cuál es la reducción en las importaciones luego del fallo a favor de la industria estadounidense?

Para responder a estas preguntas, seguimos a Krupp y Pollard y se definen tres variables binarias: *befile6* es igual a 1 durante los seis meses anteriores a la demanda, *affile6* indica los seis meses posteriores a la misma, y *afdec6* denota los seis meses que siguieron al fallo positivo. La variable dependiente es el volumen de importaciones de cloruro de bario de China, *chnimp*, que se emplea de forma logarítmica. Se incluyen como variables explicativas, todas en forma logarítmica, un índice de la producción química, *chempi* (para controlar la demanda general de cloruro de bario), el volumen de la producción de gasolina, *gas* (otra variable de demanda) y un índice del tipo de cambio, *rtwex*, que mide la solidez del dólar en comparación con otras divisas. El índice de producción química se definió en 100 en junio de 1977. Aquí el análisis difiere ligeramente del de Krupp y Pollard en cuanto a que aquí se utilizan logaritmos naturales de todas las variables (salvo las binarias, desde luego) y se incluyen las tres variables binarias en la misma regresión.

Los datos mensuales que se utilizan de febrero de 1978 a diciembre de 1988 proporcionan lo siguiente:

$$\begin{aligned}\widehat{\log(chnimp)} = {} & \underset{(21.05)}{-17.80} + \underset{(.48)}{3.12}\log(chempi) + \underset{(.907)}{.196}\log(gas) \\ & + \underset{(.400)}{.983}\log(rtwex) + \underset{(.261)}{.060}\, befile6 - \underset{(.264)}{.032}\, affile6 - \underset{(.286)}{.565}\, afdec6 \\ & n = 131,\ R^2 = .305,\ \bar{R}^2 = .271.\end{aligned} \tag{10.22}$$

La ecuación muestra que *befile6* es insignificante en términos estadísticos, de modo que no hay evidencia de que las importaciones chinas fueran de manera inusitada altas en los seis meses anteriores a la interposición de la demanda. Además, aun cuando el estimador del coeficiente de *affile6* es negativo, es pequeño (lo que indica una disminución de alrededor de 3.2% en las importaciones chinas) y es muy insignificante en términos estadísticos. El coeficiente de *afdec6* muestra una disminución sustancial en las importaciones chinas de cloruro de bario luego del fallo a favor de la industria estadounidense, que no es sorpresivo. Como el efecto es tan grande, se calcula el cambio porcentual exacto: $100[\exp(-.565) - 1] \approx -43.2\%$. El coeficiente es estadísticamente significativo a un nivel de 5% en contra de una alternativa de dos colas.

Los signos de los coeficientes de las variables de control son los esperados: un incremento en la producción general de químicos aumenta la demanda del agente limpiador. La producción de gasolina no afecta de forma significativa las importaciones chinas. El coeficiente sobre log(*rtwex*) muestra que un incremento en el valor del dólar respecto a otras monedas incrementa la demanda de las importaciones chinas, como lo predice la teoría económica. (De hecho, la elasticidad no es estadísticamente diferente de uno. ¿Por qué?)

Las interacciones entre las variables cualitativas y cuantitativas se emplean también en el análisis de series de tiempo. A continuación se da un ejemplo de importancia práctica.

Ejemplo 10.6

[Resultados de la elección y desempeño económico]

Fair (1996) resume su trabajo explicando los resultados de la elección presidencial en términos del desempeño económico. Explica la proporción del voto que en una elección bipartidista obtiene el candidato demócrata con datos de 1916 a 1992 (cada cuatro años), para un total de 20 observaciones. Se estima una versión simplificada del modelo de Fair (utilizando nombres de variables que son más descriptivos que los suyos):

$$demvote = \beta_0 + \beta_1 partyWH + \beta_2 incum + \beta_3 partyWH \cdot gnews + \beta_4 partyWH \cdot inf + u,$$

donde *demvote* es la proporción del voto que corresponde al candidato demócrata en una elección bipartidista. La variable explicativa *partyWH* es similar a una variable binaria, pero adopta el valor 1 si un demócrata está en la Casa Blanca y -1 si está un republicano. Fair utiliza esta variable para imponer la restricción de que el efecto de un republicano en la Casa Blanca tiene la misma magnitud, pero con signo opuesto, que el de un demócrata en la Casa Blanca. Se trata de una restricción natural ya que, por definición, las participaciones de los partidos deben sumar uno. También conserva dos grados de libertad, lo que es importante con tan pocas observaciones. De igual modo, la variable *incum* se define como 1 si el mandatario titular que está buscando la reelección es un demócrata, -1 si es republicano, y 0 en otros casos. La variable *gnews* es el número de trimestres durante los primeros 15 (de un total de 16) en la administración actual, en los que el crecimiento trimestral de la producción real *per cápita* estuvo por encima de 2.9% (a una tasa anual), e *inf* es la tasa de inflación anual promedio de los primeros 15 trimestres de la administración. Vea Fair (1996) para obtener definiciones precisas.

Los economistas están más interesados en los términos de interacción *partyWH·gnews* y *partyWH·inf*. Como *partyWH* es igual a uno cuando un demócrata está en la Casa Blanca, β_3 mide el efecto de las buenas noticias económicas sobre el partido en el poder; se espera que $\beta_3 > 0$. Asimismo, β_4 mide el efecto que la inflación tiene sobre el partido en el poder. Como la inflación durante una administración se considera una mala noticia, esperamos que $\beta_4 < 0$.

La ecuación estimada con ayuda de los datos de FAIR.RAW es

$$\begin{aligned}\widehat{demvote} = {} & \underset{(.012)}{.481} - \underset{(.0405)}{.0435}\, partyWH + \underset{(.0234)}{.0544}\, incum \\ & + \underset{(.0041)}{.0108}\, partyWH \cdot gnews - \underset{(.0033)}{.0077}\, partyWH \cdot inf \\ & n = 20,\ R^2 = .663,\ \bar{R}^2 = .573.\end{aligned} \tag{10.23}$$

Todos los coeficientes, salvo el de *partyWH*, son estadísticamente significativos a un nivel de 5%. Buscar la reelección vale alrededor de 5.4 puntos porcentuales en la participación del voto (recuerde que *demvote* se mide como proporción). Además, la variable de las noticias económicas ejerce un efecto positivo: un trimestre más de buenas noticias vale alrededor de 1.1 puntos porcentuales. La inflación, como se esperaba, tiene un efecto negativo: si la inflación anual promedio es, por decir, dos puntos porcentuales mayor, el partido en el poder pierde cerca de 1.5 puntos porcentuales de los votos en la elección bipartidista.

Se pudo haber utilizado esta ecuación para predecir el resultado de la elección presidencial de 1996 entre Bill Clinton, demócrata, y Bob Dole, republicano (el candidato independiente, Ross Perot se excluye porque la ecuación de Fair es sólo para una elección bipartidista). Dado que Clinton buscaba la reelección, $partyWH = 1$ e $incum = 1$. Para pronosticar el resultado de la elección se necesitan las variables *gnews* e *inf*. Durante los primeros 15 trimestres del mandato de Clinton, el PIB real *per cápita* excedió 2.9% en tres ocasiones, de modo que $gnews = 3$. Además, utilizando el deflactor del PIB reportado en la tabla B-4 del *ERP*, de 1997, la tasa de inflación media anual (calculada usando la fórmula de Fair) del cuarto trimestre de 1991 al tercer trimestre de 1996 fue de 3.019. Al sustituir estos valores en la ecuación (10.23) se obtiene

$$\widehat{demvote} = .481 - .0435 + .0544 + .0108(3) - .0077(3.019) \approx .5011.$$

Por tanto, con base en la información conocida antes de la elección de noviembre, se pronosticaba que Clinton recibiría una mayoría muy escasa del voto bipartidista: cerca de 50.1%. De hecho, Clinton ganó con más facilidad: su participación en el voto bipartidista fue de 54.65%.

10.5 Tendencias y estacionalidad

Caracterización de la tendencia en las series de tiempo

Muchas series de tiempo económicas tienen una tendencia común de crecimiento a lo largo del tiempo. Se debe reconocer que ciertas series contienen una **tendencia en el tiempo** para hacer inferencias causales con los datos de las series de tiempo. Ignorar el hecho de que dos secuencias tienen tendencias en la misma dirección o en direcciones opuestas nos conduce a la conclusión falsa de que los cambios en una variable en realidad son ocasionados por modificaciones en otra variable. En muchos casos, parece que dos procesos de series de tiempo están correlacionados sólo porque ambos tienen una tendencia en el tiempo por razones relacionadas con otros factores inobservables.

La figura 10.2 contiene una gráfica de productividad laboral (producción por hora de trabajo) para los años de 1947 a 1987 en Estados Unidos. Esta serie muestra una clara tendencia hacia arriba, que refleja el hecho de que los trabajadores se han vuelto más productivos con el tiempo.

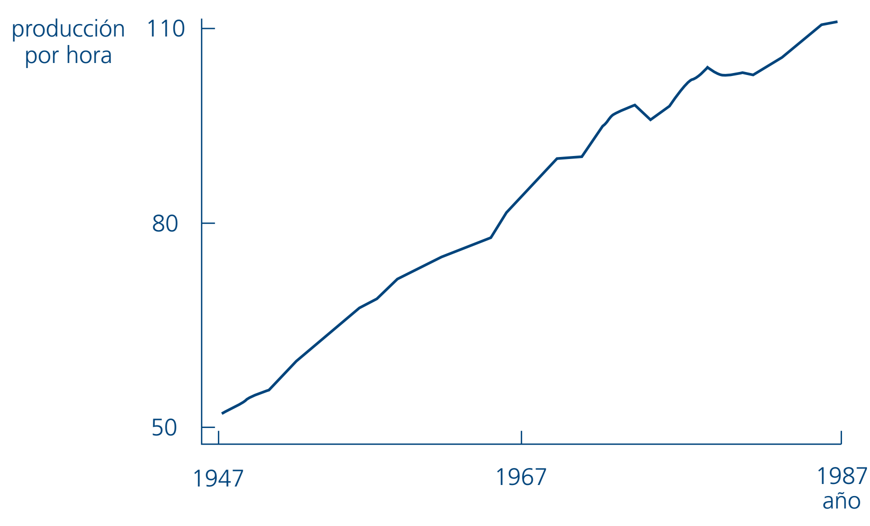
Otras series, al menos sobre ciertos periodos, tienen claras tendencias hacia abajo. Como las tendencias positivas son más comunes, en ellas se centrará el siguiente análisis.

¿Qué clase de modelos estadísticos representan de manera adecuada el comportamiento de la tendencia? Una formulación popular es escribir la serie $\{y_t\}$ como

$$y_t = \alpha_0 + \alpha_1 t + e_t,\ t = 1, 2, \ldots, \qquad \boxed{10.24}$$

FIGURA 10.2

Producción por hora de trabajo en Estados Unidos, durante 1947-1987; 1977 = 100.



en donde, en el caso más sencillo, $\{e_t\}$ es una secuencia independiente e idénticamente distribuida (i.i.d.) con $E(e_t) = 0$ y $Var(e_t) = \sigma_e^2$. Advierta cómo el parámetro α_1 multiplica al tiempo, t, lo que da como resultado una **tendencia lineal en el tiempo**. La interpretación de α_1 en la ecuación (10.24) es sencilla: si se mantienen fijos todos los demás factores (los de e_t), α_1 mide el cambio en y_t de un periodo al siguiente debido al transcurso del tiempo: cuando $\Delta e_t = 0$,

$$\Delta y_t = y_t - y_{t-1} = \alpha_1.$$

Otra manera de considerar una secuencia que tiene una tendencia lineal en el tiempo es que su valor promedio es una función lineal del tiempo:

$$E(y_t) = \alpha_0 + \alpha_1 t. \tag{10.25}$$

Si $\alpha_1 > 0$, entonces, en promedio, y_t está creciendo en el tiempo y por consiguiente tiene una tendencia hacia arriba. Si $\alpha_1 < 0$, entonces y_t tiene una tendencia hacia abajo. Los valores de y_t no caen de manera exacta sobre la línea en la ecuación (10.25) debido a la aleatoriedad, pero los valores esperados sí están sobre ella. A diferencia de la media, la varianza de y_t es constante en el tiempo: $Var(y_t) = Var(e_t) = \sigma_e^2$.

Si $\{e_t\}$ es una secuencia i.i.d., entonces $\{y_t\}$ es una secuencia independiente, pero no es idénticamente distribuida. Una caracterización más realista de la tendencia en las series de tiempo permite que $\{e_t\}$ esté correlacionada a través del tiempo, pero esto no modifica la cualidad de una tendencia lineal en el tiempo. De hecho, lo que es importante para el análisis de regresión, bajo los supuestos del modelo lineal clásico, es que $E(y_t)$ es lineal en t. Cuando estudie las propiedades de muestra grandes de MCO en el capítulo 11, tendrá que plantearse cuánta correlación temporal se permite en $\{e_t\}$.

Pregunta 10.4

En el ejemplo 10.4 se utilizó la tasa de fertilidad general como variable dependiente en un modelo de rezagos distribuidos finito. De 1950 a mediados de la década de los ochenta, la tasa de fertilidad general (*gfr*) muestra una clara tendencia hacia abajo. ¿Puede una tendencia lineal con $\alpha_1 < 0$ ser realista para todos los periodos futuros? Explique por qué.

Muchas series de tiempo económicas se aproximan mejor por medio de una **tendencia exponencial**, la cual se sigue cuando una serie tiene la misma tasa de crecimiento promedio de un periodo a otro. En la figura 10.3 se trazan los datos de las importaciones nominales anuales de Estados Unidos para los años de 1948 a 1995 (*ERP* de 1997, tabla B-101).

En los primeros años, se ve que el cambio en las importaciones anuales es relativamente pequeño, pero aumenta con el transcurso del tiempo. Esto es consistente con una *tasa de crecimiento promedio constante*: el cambio porcentual es casi el mismo en cada periodo.

En la práctica, una tendencia exponencial en una serie de tiempo se consigue modelando el logaritmo natural de la serie como una tendencia lineal (suponiendo que $y_t > 0$):

$$\log(y_t) = \beta_0 + \beta_1 t + e_t, t = 1, 2, \ldots. \tag{10.26}$$

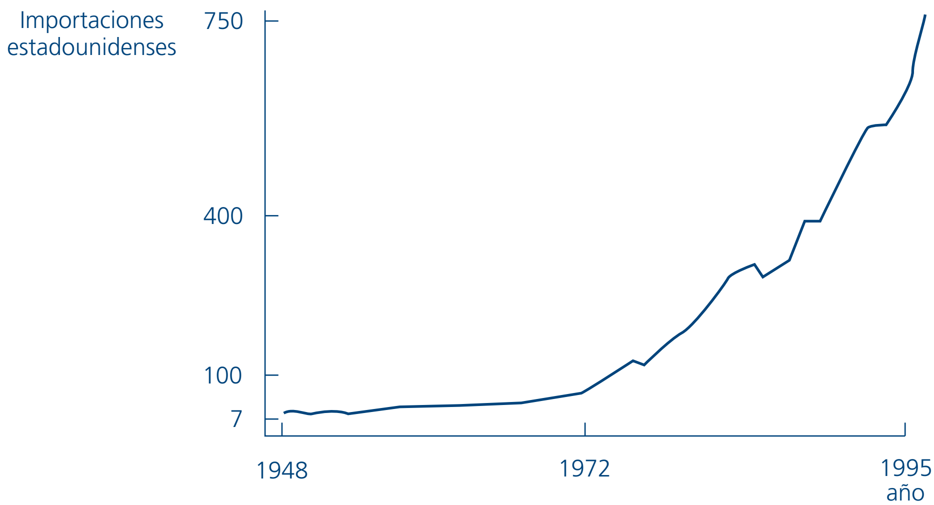
La exponenciación muestra que y_t por sí misma tiene una tendencia exponencial: $y_t = \exp(\beta_0 + \beta_1 t + e_t)$. Como se deseará utilizar series de tiempo de tendencia exponencial en los modelos de regresión lineal, la ecuación (10.26) resulta ser la forma más conveniente para representar estas series.

¿Cómo se interpreta β_1 en la ecuación (10.26)? Recuerde que, para cambios pequeños, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$ es aproximadamente el cambio proporcional en y_t:

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}. \tag{10.27}$$

FIGURA 10.3

Importaciones nominales de Estados Unidos durante los años de 1948 a 1995 (en miles de millones de dólares).



Al lado derecho de la ecuación (10.27) también se le llama **tasa de crecimiento** de y del periodo $t - 1$ al periodo t. Para convertir la tasa de crecimiento en un porcentaje, sólo se multiplica por 100. Si y_t cumple con (10.26), entonces, tomando cambios y estableciendo $\Delta e_t = 0$,

$$\Delta \log(y_t) = \beta_1, \text{ para toda } t. \tag{10.28}$$

En otras palabras, β_1 es aproximadamente la tasa de crecimiento promedio por periodo de y_t. Por ejemplo, si t denota un año y $\beta_1 = .027$, entonces y_t crece alrededor de 2.7% al año en promedio.

Aunque las tendencias lineales y exponenciales son las más comunes, las tendencias en el tiempo pueden ser más complicadas. Por ejemplo, en vez del modelo de tendencia lineal de la ecuación (10.24), se podría tener una tendencia cuadrática en el tiempo:

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t. \tag{10.29}$$

Si α_1 y α_2 son positivas, entonces la pendiente de la tendencia está aumentando, como se ve con facilidad al calcular la pendiente aproximada (manteniendo a e_t fijo):

$$\frac{\Delta y_t}{\Delta t} \approx \alpha_1 + 2\alpha_2 t. \tag{10.30}$$

[Si está usted familiarizado con el cálculo, reconocerá el lado derecho de la ecuación (10.30) como la derivada de $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ respecto a t.] Si $\alpha_1 > 0$, pero $\alpha_2 < 0$, la tendencia tiene una forma de U invertida. Tal vez ésta no sea una muy buena descripción de ciertas series que muestran tendencia debido a que requieren que una tendencia decreciente siga después de una tendencia creciente. Sin embargo, durante un periodo dado, puede ser una manera flexible de modelar series de tiempo que tienen tendencias más complicadas que las de las ecuaciones (10.24) o (10.26).

Uso de variables con tendencia en el análisis de regresión

En el análisis de regresión resulta muy fácil llevar la cuenta de las variables explicadas o explicativas que muestran tendencia. En primer lugar, nada respecto a las variables con tendencia viola necesariamente los supuestos ST.1 a ST.6 del modelo lineal clásico. Sin embargo, debe procurar tener en cuenta el hecho de que factores inobservables con tendencia que afectan a y_t podrían estar correlacionados con las variables explicativas. Si ignora esta posibilidad, podemos encontrar una relación falsa entre y_t y una o más variables explicativas. El fenómeno de encontrar una relación falsa entre dos o más variables con tendencia, sencillamente porque cada una está creciendo con el tiempo, es un **problema de regresión espuria**. Por fortuna, la adición de una tendencia en el tiempo elimina este problema.

En concreto, considere un modelo donde dos factores inobservables, x_{t1} y x_{t2}, afectan a y_t. Además, hay factores inobservables que están creciendo o decreciendo de manera sistemática con el tiempo. Un modelo que captura esto es

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + u_t. \tag{10.31}$$

Este modelo encaja en el esquema de la regresión lineal múltiple con $x_{t3} = t$. Tomar en cuenta la tendencia en esta ecuación reconoce de manera explícita que y_t puede estar creciendo ($\beta_3 > 0$) o decreciendo ($\beta_3 < 0$) en el tiempo por motivos que en esencia no están relacionados con x_{t1} o x_{t2}. Si la ecuación (10.31) satisface los supuestos ST.1, ST.2 y ST.3, entonces al omitir t y hacer la regresión de y_t sobre x_{t1}, x_{t2} por lo común generará estimadores sesgados de β_1 y β_2: se ha omitido, en efecto, una variable importante, t, de la regresión. Esto es válido en particular si x_{t1} y x_{t2} poseen una tendencia, porque entonces pueden estar muy correlacionadas con t. El siguiente ejemplo muestra cómo la omisión de una tendencia en el tiempo puede dar como resultado una regresión espuria.

Ejemplo 10.7

[Inversión y precios de vivienda]

Los datos de HSEINV.RAW son observaciones anuales sobre la inversión y el índice de precios de la vivienda en Estados Unidos de 1947 a 1988. Sea *invpc* la inversión real en vivienda *per cápita* (en miles de dólares) y *price* un índice del precio de vivienda (igual a 1 en 1982). Una regresión simple en la forma de elasticidad constante, que puede pensarse como una ecuación de oferta para la cantidad de viviendas disponibles, da

$$\begin{gathered}\widehat{\log(invpc)} = -.550 + 1.241 \log(price) \\ (.043) \quad (.382) \\ n = 42,\ R^2 = .208,\ \bar{R}^2 = .189.\end{gathered} \tag{10.32}$$

La elasticidad de la inversión *per cápita* respecto al precio es muy grande y significativa; estadísticamente no es diferente de 1. Debe ser cuidadoso aquí. Tanto *invpc* como *price* muestran sesgos hacia arriba. En particular, si hace la regresión de log(*invpc*) se obtiene un coeficiente sobre la tendencia igual a .0081 (error estándar = .0018); la regresión de log(*price*) en t arroja un coeficiente de tendencia igual a .0044 (error estándar = .0004). Aunque los errores estándar en los coeficientes de tendencia no son necesariamente confiables, ya que estas regresiones tienden a contener una correlación serial sustancial, las estimaciones de los coeficientes revelan sesgos hacia arriba.

Para representar el comportamiento de la tendencia de las variables, se añade una tendencia en el tiempo:

$$
\begin{aligned}
\widehat{\log(invpc)} &= -.913 - .381\, log(price) + .0098\, t \\
&\ \ (.136)\ \ (.679) \qquad\qquad\quad (.0035) \\
n &= 42,\ R^2 = .341,\ \bar{R}^2 = .307.
\end{aligned}
\tag{10.33}
$$

La historia es muy distinta ahora: la elasticidad estimada del precio es negativa y en términos estadísticos no es diferente de cero. La tendencia en el tiempo es estadísticamente significativa y su coeficiente implica un incremento aproximado de 1% anual en *invpc*, en promedio. A partir de este análisis, no se puede concluir que la inversión real *per cápita* en vivienda esté afectada en lo absoluto por el precio. Hay otros factores, capturados en la tendencia en el tiempo, que influyen en *invpc*, pero no los hemos modelado. Los resultados en la ecuación (10.32) muestran una relación espuria entre *invpc* y *price* debido al hecho de que el precio también tiene una tendencia ascendente en el tiempo.

En algunos casos, la adición de una tendencia en el tiempo hace que una variable explicativa clave resulte *más* significativa. Esto puede suceder si las variables dependiente e independiente tienen tendencias distintas (por ejemplo, una hacia arriba y otra hacia abajo), pero el movimiento en la variable independiente *alrededor de* su línea de tendencia provoca un movimiento de la variable dependiente alejándose de su línea de tendencia.

Ejemplo 10.8

[Ecuación de fertilidad]

Si se añade una tendencia lineal en el tiempo a la ecuación (10.18), se obtiene

$$
\begin{aligned}
\widehat{gfr}_t &= 111.77 + .279\, pe_t - 35.59\, ww2_t + .997\, pill_t - 1.15\, t \\
&\ \ (3.36)\ \ (.040) \qquad\ (6.30) \qquad\ (6.626) \qquad\ (.19) \\
n &= 72,\ R^2 = .662,\ \bar{R}^2 = .642.
\end{aligned}
\tag{10.34}
$$

El coeficiente de *pe* es más de tres veces la estimación de la ecuación (10.18) y es mucho más significativa en términos estadísticos. Resulta interesante que *pill* no sea significativa una vez que se permite una tendencia lineal. Como se ve por la estimación, *gfr* estaba descendiendo, en promedio, durante este periodo, manteniendo constantes los demás factores.

Como la tasa general de fertilidad mostró tendencias tanto hacia arriba como hacia abajo durante el periodo de 1913 a 1984, se puede ver la intensidad que tiene el efecto estimado de *pe* cuando se utiliza una tendencia cuadrática:

$$
\begin{aligned}
\widehat{gfr}_t &= 124.09 + .348\, pe_t - 35.88\, ww2_t - 10.12\, pill_t \\
&\ \ (4.36)\ \ (.040) \qquad\ (5.71) \qquad\quad (6.34) \\
&\ - 2.53\, t + .0196\, t^2 \\
&\ \ \ (.39) \quad\ (.0050) \\
n &= 72,\ R^2 = .727,\ \bar{R}^2 = .706.
\end{aligned}
\tag{10.35}
$$

El coeficiente de *pe* es incluso más grande y más significativo en términos estadísticos. Ahora, *pill* tiene el efecto negativo esperado y es poco significativa, y ambos términos de tendencia son estadísticamente significativos. La tendencia cuadrática es una forma flexible de representar el comportamiento inusual de la tendencia de *gfr*.

Tal vez se pregunte respecto al ejemplo 10.8: ¿por qué detenerse en una tendencia cuadrática? Nada nos impide agregar, por ejemplo, t^3 como variable independiente y, de hecho, esto podría justificarse (vea el ejercicio para computadora C10.6). Pero debemos procurar no entusiasmarnos cuando se incluyan términos de tendencia en un modelo. Son recomendables las tendencias más o menos simples que registran los amplios movimientos en la variable dependiente que no se explican por medio de las variables independientes en el modelo. Si se incluyen los suficientes términos polinomiales en t, entonces se puede rastrear muy bien cualquier serie. Pero esto nos ofrece poca ayuda para determinar qué variables explicativas influyen en y_t.

Interpretación de las regresiones con tendencia en el tiempo mediante la eliminación de la tendencia

Incluir una tendencia en el tiempo en un modelo de regresión permite hacer una interpretación atractiva en términos de la **eliminación de la tendencia** en las series de datos originales, antes de utilizarlas en el análisis de regresión. Para concretar, el análisis se concentrará en el modelo (10.31), pero las conclusiones son mucho más generales.

Cuando se hace la regresión de y_t sobre x_{t1}, x_{t2} y t, se obtiene la ecuación ajustada

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t. \qquad \boxed{10.36}$$

Podemos ampliar la interpretación de MCO como descuento de efectos parciales que se cubrieron en el capítulo 3, para mostrar que $\hat{\beta}_1$ y $\hat{\beta}_2$ se obtienen como sigue.

i) Haga la regresión de cada una de las variables y_t, x_{t1} y x_{t2} sobre una constante y la tendencia en el tiempo t y guarde los residuales, digamos, $\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, $t = 1, 2, \ldots, n$. Por ejemplo,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t.$$

Así, puede pensarse en $\ddot{y}_t$ como el resultado de haber *eliminado la tendencia lineal.* Para eliminar la tendencia de y_t, se ha estimado por MCO el modelo

$$y_t = \alpha_0 + \alpha_1 t + e_t$$

los residuales de esta regresión, $\hat{e}_t = \ddot{y}_t$, tienen la tendencia en el tiempo eliminada (al menos en la muestra). Una interpretación similar es válida para $\ddot{x}_{t1}$ y $\ddot{x}_{t2}$.

ii) Realice la regresión de

$$\ddot{y}_t \text{ sobre } \ddot{x}_{t1}, \ddot{x}_{t2}. \qquad \boxed{10.37}$$

(El intercepto no es necesario, pero si se incluye no afecta en nada; se estimará como cero.) Esta regresión produce exactamente $\hat{\beta}_1$ y $\hat{\beta}_2$ de la ecuación (10.36).

Esto significa que las estimaciones de interés primordial, $\hat{\beta}_1$ y $\hat{\beta}_2$, pueden interpretarse como si provinieran de una regresión *sin* una tendencia en el tiempo, pero en la que se eliminó primero la tendencia de la variable dependiente y de todas las independientes. La misma conclusión es válida para cualquier número de variables independientes y si la tendencia es cuadrática o de algún otro grado polinomial.

Si t se omite de (10.36), entonces no ocurre la eliminación de la tendencia y parecería que y_t se relaciona con una o más de las x_{tj} sencillamente porque cada una contiene una tendencia; lo vimos en el ejemplo 10.7. Si el término de tendencia es estadísticamente significativo, y los resultados se modifican de manera importante cuando se agrega una tendencia en el tiempo a una regresión, entonces los resultados iniciales sin una tendencia deben tratarse con reserva.

La interpretación de $\hat{\beta}_1$ y $\hat{\beta}_2$ muestra que es una buena idea incluir una tendencia en la regresión si alguna variable independiente tiene tendencia, incluso si y_t no la tiene. Si y_t no cuenta con una tendencia observable, pero, digamos, x_{t1} está creciendo con el tiempo, entonces la exclusión de una tendencia de la regresión haría parecer como si x_{t1} no tuviera ningún efecto sobre y_t, aun cuando los movimientos de x_{t1} en torno a su tendencia sí influyan en y_t. Esto se captará si t se incluye en la regresión.

Ejemplo 10.9

[Empleo en Puerto Rico]

Cuando se añadió una tendencia lineal a la ecuación (10.17), las estimaciones son

$$\begin{aligned} \widehat{\log(prepop_t)} = {} & \underset{(1.30)}{-8.70} - \underset{(.044)}{.169} \log(mincov_t) + \underset{(0.18)}{1.06} \log(usgnp_t) \\ & - \underset{(.005)}{.032}\, t \\ n = 38,\ & R^2 = .847,\ \bar{R}^2 = .834. \end{aligned} \tag{10.38}$$

El coeficiente de log(*usgnp*) ha cambiado en forma drástica: de −.012 a 1.06, y de ser insignificante a ser muy significativo. El coeficiente del salario mínimo se ha modificado sólo un poco, aunque el error estándar es de manera notable más pequeño, volviendo a log(*mincov*) más significativo que antes.

La variable $prepop_t$ no muestra una tendencia clara hacia arriba o hacia abajo, pero log(*usgnp*) posee una tendencia lineal ascendente. [Una regresión de log(*usgnp*) sobre t da una estimación de alrededor de .03, de modo que *usgnp* está creciendo cerca de 3% al año durante el periodo.] Puede considerarse la estimación 1.06 como sigue: cuando *usgnp* aumenta 1% por *encima* de su tendencia a largo plazo, *prepop* se incrementa alrededor de 1.06 por ciento.

Cálculo de la *R*-cuadrada cuando la variable dependiente tiene tendencia

Las *R*-cuadradas en las regresiones con series de tiempo a menudo son muy altas, en especial cuando se comparan con las *R*-cuadradas características de los datos de corte transversal. ¿Esto significa que aprendemos más sobre los factores que influyen en y a partir de los datos de las series de tiempo? No necesariamente. Por una parte, los datos de series de tiempo con frecuencia vienen como agregados (por ejemplo, el salario medio por hora en la economía estadounidense) y los agregados a menudo son más fáciles de explicar que los resultados sobre personas, familias o empresas, datos estos últimos de corte transversal. Sin embargo, las *R*-cuadradas usuales y ajustadas de las regresiones con series de tiempo pueden ser elevadas de manera artificial, cuando la variable dependiente muestra tendencia. Recuerde que la R^2 es una medida de la dimensión de la varianza del error en relación con la varianza de y. La fórmula para la *R*-cuadrada ajustada muestra esto de forma directa:

$$\bar{R}^2 = 1 - (\hat{\sigma}_u^2/\hat{\sigma}_y^2),$$

donde $\hat{\sigma}_u^2$ es el estimador insesgado de la varianza del error, $\hat{\sigma}_y^2 = \text{STC}/(n-1)$, and $\text{STC} = \sum_{t=1}^{n}(y_t - \bar{y})^2$. Ahora, la estimación de la varianza del error no es un problema cuando y_t tiene tendencia, siempre y cuando se incluya en la regresión una tendencia en el tiempo. Sin embargo, cuando $\text{E}(y_t)$ sigue, por ejemplo, una tendencia lineal en el tiempo [vea (10.24)], $\text{STC}/(n-1)$ ya no es un estimador insesgado o consistente de Var(y_t). De hecho, STC/($n - 1$) puede sobrestimar de manera sustancial la varianza de y_t, porque no toma en cuenta su tendencia y_t.

Cuando la variable dependiente tiene tendencia lineal, cuadrática o cualquier otra tendencia polinomial, resulta sencillo calcular una medida de bondad de ajuste que primero descuente el efecto de cualquier tendencia en el tiempo sobre y_t. El método más sencillo es calcular la R-cuadrada usual de una regresión en la que ya se ha eliminado la tendencia de la variable dependiente. Por ejemplo, si el modelo es la ecuación (10.31), entonces primero se hace la regresión de y_t sobre t y se obtienen los residuales $\ddot{y}_t$. Luego, se efectúa la regresión

$$\ddot{y}_t \text{ sobre } x_{t1}, x_{t2} \text{ y } t. \tag{10.39}$$

La R-cuadrada de esta regresión es

$$1 - \frac{\text{SRC}}{\sum_{t=1}^{n} \ddot{y}_t^2}, \tag{10.40}$$

donde SRC es idéntica a la suma de residuales cuadrados de la ecuación (10.36). Como $\sum_{t=1}^{n} \ddot{y}_t^2 \leq \sum_{t=1}^{n} (y_t - \bar{y})^2$ (y por lo común la desigualdad es estricta), la R-cuadrada de la ecuación (10.40) no es mayor (por lo general es menor) que la R-cuadrada de (10.36). (La suma de residuales cuadrados es idéntica en ambas regresiones.) Cuando y_t contiene una marcada tendencia lineal en el tiempo, la ecuación (10.40) puede ser mucho menor que la R-cuadrada usual.

La R-cuadrada en (10.40) refleja mejor qué tan adecuadas son x_{t1} y x_{t2} para explicar y_t porque compensan el efecto de la tendencia en el tiempo. Después de todo, una variable con tendencia siempre se puede explicar mediante algún tipo de tendencia, pero esto no significa que hayamos descubierto un factor que genere movimientos en y_t. Una R-cuadrada ajustada también puede calcularse a partir de la ecuación (10.40): divida SRC entre ($n - 4$) ya que éstos son los *gl* de la ecuación (10.36), y divida $\sum_{t=1}^{n} \ddot{y}_t^2$ entre ($n - 2$), ya que hay dos parámetros de tendencia estimados en la eliminación de la tendencia de y_t. En general, la SRC se divide entre los *gl* de la regresión usual (que comprende cualquier tendencia en el tiempo), y $\sum_{t=1}^{n} \ddot{y}_t^2$ se divide entre ($n - p$), donde p es el número de parámetros de tendencia estimados en la eliminación de la tendencia de y_t. Consulte Wooldridge (1991a) para obtener un análisis más a fondo sobre el cálculo de las medidas de bondad de ajuste para las variables con tendencia.

Ejemplo 10.10

[Inversión en vivienda]

En el ejemplo 10.7, se vio que incluir una tendencia lineal en el tiempo junto con log(*price*) en la ecuación de la inversión en vivienda produjo un efecto importante en la elasticidad del precio. Pero la R-cuadrada de la regresión (10.33) indica de manera literal que estamos "explicando" 34.1% de la variación en log(*invpc*). Esto es engañoso. Si se elimina primero la tendencia de log(*invpc*) y se hace la regresión de la variable sin tendencia sobre log(*price*) y t, la R-cuadrada se vuelve .008 y la R-cuadrada ajustada en realidad es negativa. Por consiguiente, los movimientos en log(*price*) en torno a su tendencia en la práctica no explican los movimientos en log(*invpc*) en torno a su tendencia. Esto es consistente con el hecho de que el estadístico t de log(*price*) en la ecuación (10.33) es muy pequeño.

Antes de dar por terminada esta subsección, debemos hacer una observación final. En el cálculo de la forma R-cuadrada del estadístico F para probar hipótesis múltiples, sólo se utilizan las R-cuadradas usuales sin eliminación de la tendencia. Recuerde que la forma R-cuadrada del estadístico F es sólo un recurso de cálculo y por ende la fórmula usual siempre es adecuada.

Estacionalidad

Si se observa una serie de tiempo a intervalos mensuales o trimestrales (incluso semanales o diarios), ésta puede manifestar **estacionalidad**. Por ejemplo, la venta mensual de viviendas en la región central de Estados Unidos se ve influida sobremanera por el clima. Si bien los patrones climáticos en cierto modo son aleatorios, es seguro que el clima durante enero por lo general, es más inclemente que en junio y, por tanto, la construcción de viviendas nuevas es mayor en junio que en enero. Una manera de modelar este fenómeno es permitir que el valor esperado de la serie, y_t, sea distinto cada mes. En otro ejemplo, las ventas al menudeo en el cuarto trimestre en general son más altas que en los tres trimestres anteriores debido a las vacaciones de Navidad. Una vez más, esto puede ser captado si se permite que las ventas al menudeo sean diferentes en el transcurso de un año. Esto además de permitir la posibilidad de considerar una media con tendencia. Por ejemplo, las ventas al menudeo en el primer trimestre de este año fueron mayores que las del cuarto trimestre de hace 30 años, ya que han crecido a un ritmo constante. No obstante, si se comparan las ventas medias dentro de un año cualquiera, el factor del periodo vacacional suele hacer que las ventas resulten mayores durante el cuarto trimestre.

Aun cuando muchas series de datos mensuales y trimestrales muestran patrones estacionales, no todas lo hacen. Por ejemplo, no existe un patrón estacional observable en los intereses mensuales o en las tasas de inflación. Asimismo, las series que manifiestan patrones estacionales a menudo se **ajustan estacionalmente** antes de reportarlas para su uso público. Una serie ajustada estacionalmente es aquella a la que, en principio, se le han eliminado los factores estacionales. El ajuste estacional puede realizarse de una gran variedad de maneras, pero un análisis minucioso rebasa el alcance de este libro. [Vea Harvey (1990) y Hylleberg (1992) para obtener una idea de los tratamientos pormenorizados.]

El ajuste estacional se ha vuelto tan común que en muchos casos no es posible obtener datos no ajustados estacionalmente. El PIB trimestral estadounidense es un ejemplo destacado. En el informe anual *Economic Report of the President*, muchas bases de datos macroeconómicas reportadas mensualmente (al menos en los años más recientes) y aquellas que muestran patrones estacionales se ajustan todas estacionalmente. Las principales fuentes para series de tiempo macroeconómicas, incluida *Citibase*, también ajustan estacionalmente a muchas de las series. Por esta razón, el alcance para usar nuestro propio ajuste estacional con frecuencia se ve limitado.

En ocasiones, se trabaja con datos que no están ajustados estacionalmente y resulta útil saber con cuáles métodos sencillos se cuenta para tratar la estacionalidad en los modelos de regresión. En general, se puede incluir un conjunto de **variables binarias estacionales** para representar la estacionalidad en la variable dependiente, las independientes o ambas.

El método es sencillo. Imagine que cuenta con datos mensuales y piensa que los patrones estacionales de un año son más o menos constantes en el tiempo. Por ejemplo, como la Navidad siempre se celebra en la misma época del año, podemos esperar que las ventas al menudeo sean más altas, en promedio, en los últimos meses del año que en los primeros. O bien, como los patrones climáticos son parecidos, en términos generales, a lo largo de los años, la construcción de viviendas nuevas en la región central de Estados Unidos en promedio será mayor durante el verano que en el invierno. Un modelo general para datos mensuales que representa este fenómeno es

$$y_t = \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_3 apr_t + \dots + \delta_{11} dec_t + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t, \quad \mathbf{10.41}$$

donde feb_t, mar_t, …, dec_t son variables binarias que indican si el periodo t corresponde al mes adecuado. En esta formulación, enero es el mes base y β_0 es el intercepto de enero. Si no hay estacionalidad en y_t, una vez que las x_{tj} se

Pregunta 10.5

En la ecuación (10.41), ¿cuál es el intercepto para marzo? Explique por qué las variables binarias estacionales satisfacen el supuesto de exogeneidad estricta.

han controlado, entonces δ_1 a δ_{11} son todas cero. Esto se verifica con facilidad por medio de una prueba F.

Ejemplo 10.11

[Efectos de las demandas *antidumping*]

En el ejemplo 10.5, se usaron datos mensuales que no se ajustaron estacionalmente. Por tanto, debe agregar variables binarias estacionales para asegurarse que ninguna de las condiciones importantes cambie. Podría suceder que los meses previos a la demanda sean meses en que las importaciones son mayores o menores, en promedio, que en los demás meses. Cuando se agregan las 11 variables binarias mensuales como en la ecuación (10.41) y se prueba su significancia conjunta, se obtiene el valor-p = .59, y de este modo se prueba que las variables binarias estacionales son conjuntamente insignificantes. Además, no hay un cambio importante en las estimaciones una vez que se toma en cuenta la significancia estadística. Krupp y Pollard (1996) en realidad utilizaron tres variables binarias para las estaciones (otoño, primavera y verano, con el invierno como estación base), en lugar de un conjunto completo de binarias mensuales; el resultado es esencialmente el mismo.

Si los datos son trimestrales, entonces se incluirían variables binarias para tres de los cuatro trimestres, y la categoría omitida sería el trimestre base. A veces resulta útil que las binarias estacionales interactúen con algunas de las x_{tj} para permitir que el efecto de x_{tj} sobre y_t difiera a lo largo del año.

De la misma manera que incluir una tendencia en el tiempo en una regresión se interpreta como la eliminación de la tendencia inicial de los datos, incluir binarias estacionales en una regresión puede interpretarse como la **eliminación de la estacionalidad** de los datos. En concreto, considere la ecuación (l0.41) con $k = 2$. Los coeficientes de pendiente de MCO $\hat{\beta}_1$ y $\hat{\beta}_2$ correspondientes a x_1 y x_2 se obtienen como sigue:

i) Determine las regresiones de y_t, x_{t1} y x_{t2} sobre una constante y las binarias mensuales, feb_t, mar_t, …, dec_t, y guarde los residuales, digamos, $\ddot{y}_t$, $\ddot{x}_{t1}$ y $\ddot{x}_{t2}$, para toda $t = 1, 2 \ldots, n$. Por ejemplo,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 feb_t - \hat{\alpha}_2 mar_t - \ldots - \hat{\alpha}_{11} dec_t.$$

Este es un método de eliminación de la estacionalidad de una serie de tiempo mensual. La misma interpretación es válida para $\ddot{x}_{t1}$ y $\ddot{x}_{t2}$.

ii) Realice la regresión, sin las binarias mensuales, de $\ddot{y}_t$ sobre $\ddot{x}_{t1}$ y $\ddot{x}_{t2}$ como en la ecuación (10.37)]. Esto da como resultado $\hat{\beta}_1$ y $\hat{\beta}_2$.

En algunos casos, si y_t tiene una estacionalidad pronunciada, una mejor medida de la bondad de ajuste es una R-cuadrada basada en la y_t sin estacionalidad. Esto compensa cualquier efecto estacional que no explique la x_{tj}. Los ajustes de grados de libertad específicos se estudian en Wooldridge (1991a).

Las series de tiempo que muestran patrones estacionales también pueden tener tendencia, en cuyo caso debe estimarse un modelo de regresión con una tendencia en el tiempo y variables binarias estacionales. Las regresiones pueden entonces interpretarse como regresiones que utilizan series sin tendencia ni estacionalidad. Los estadísticos de la bondad de ajuste se estudian en Wooldridge (1991a): en particular, se eliminan la tendencia y la estacionalidad de y_t al hacer la regresión sobre una tendencia en el tiempo y binarias estacionales, antes de calcular la R-cuadrada.

RESUMEN

En este capítulo se ha cubierto el análisis de regresión básico con los datos de series de tiempo. Bajo los supuestos análogos a los del análisis de corte transversal, los estimadores de MCO son insesgados (según los supuestos ST.1 a ST.3), los estimadores de MCO son MELI (de acuerdo con los supuestos ST.1 a ST.5), y los errores estándar usuales de MCO, los estadísticos t y F pueden emplearse para la inferencia estadística (bajo ST.1 a ST.6). Por la correlación temporal en la mayor parte de los datos de series de tiempo, se tienen que hacer suposiciones explícitamente sobre cómo se relacionan los errores con las variables explicativas en todos los periodos y acerca de la correlación temporal en los errores mismos. Los supuestos del modelo lineal clásico pueden ser muy restrictivos para las aplicaciones de series de tiempo, pero son un punto de partida natural. Se han aplicado tanto en la regresión estática como en los modelos de rezagos distribuidos finitos.

Los logaritmos y las variables binarias se emplean de forma regular en las aplicaciones de series de tiempo y en los estudios de eventos. También se estudiaron los números índice y las series de tiempo medidas en términos nominales y reales.

La tendencia y la estacionalidad se manejan con facilidad en un esquema de regresión múltiple cuando se incluyen variables de tiempo y binarias estacionales en las ecuaciones de regresión. Se presentaron los problemas que ocurren con la R-cuadrada usual como medida de bondad de ajuste y se propusieron algunas alternativas simples basadas en la eliminación de la tendencia y de la estacionalidad.

Supuestos del modelo clásico lineal para regresiones de series de tiempo

A continuación se encuentra un resumen de los seis supuestos del modelo clásico lineal (MCL) para regresiones de series de tiempo. Los supuestos ST.1 a ST.5 son la versión para series de tiempo de los supuestos de Gauss-Markov (que implican que los estimadores de MCO son MELI y tienen las varianzas de muestreo usuales). Sólo se necesitaron ST.1, ST.2 y ST.3 para establecer la insesgadez de MCO. Como en el caso de la regresión de corte transversal, el supuesto de normalidad, ST.6, fue empleado para que se pueda realizar inferencia estadística exacta para cualquier tamaño de muestra.

Supuesto ST.1 (Lineal en los parámetros)

El proceso estocástico $\{(x_{t1}, x_{t2}, \dots, x_{tk}, y_t): t = 1, 2, \dots, n\}$ sigue el modelo lineal

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t,$$

donde $\{u_t: t = 1, 2, \dots, n\}$ es la secuencia de errores o perturbaciones. Aquí, n es el número de observaciones (periodos).

Supuesto ST.2 (No hay colinealidad perfecta)

En la muestra (y, por tanto, en el proceso subyacente de serie de tiempo), no hay variables independientes que sean constantes ni una combinación lineal perfecta de las otras.

Supuesto ST.3 (Media condicional cero)

Para cada t, el valor esperado del error u_t, dadas las variables explicativas para *todos* los periodos, es cero. En términos matemáticos, $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1, 2, \dots, n$.

El supuesto ST.3 reemplaza al supuesto RLM.4 de corte transversal, y también significa que no es necesario hacer el supuesto de muestreo aleatorio RLM.2. Recuerde que el supuesto ST.3 implica que el error en cada periodo t no está correlacionado con ninguna variable explicativa en *ningún* periodo (incluido, por supuesto, el periodo t).

Supuesto ST.4 (Homocedasticidad)

Condicional en **X**, la varianza de u_t es la misma para todo t: $\mathrm{Var}(u_t|\mathbf{X}) = \mathrm{Var}(u_t) = \sigma^2$, $t = 1, 2, \ldots, n$.

Supuesto ST.5 (No hay correlación serial)

Condicional en **X**, los errores de dos periodos distintos no están correlacionados: $\mathrm{Corr}(u_t, u_s|\mathbf{X}) = 0$, para todo $t \neq s$.

Recuerde que se agregó el supuesto de no correlación serial, junto con el de homocedasticidad, para obtener las mismas fórmulas para las varianzas que se calcularon en las regresiones de corte transversal con muestreo aleatorio. Como se verá en el capítulo 12, el supuesto ST.5 se viola con frecuencia en modos que pueden hacer que la inferencia estadística no sea confiable.

Supuesto ST.6 (Normalidad)

Los errores u_t son independientes de **X** y son independientes e idénticamente distribuidos como Normal $(0, \sigma^2)$.

TÉRMINOS CLAVE

Ajustada estacionalmente
Autocorrelación
Contemporáneamente exógena
Correlación serial
Distribución de rezagos
Elasticidad de corto plazo
Elasticidad de largo plazo
Eliminación de la estacionalidad
Eliminación de la tendencia
Estacionalidad
Estrictamente exógena
Estudio de evento
Modelo de rezagos distribuidos finitos (RDF)
Modelo estático
Multiplicador de impacto
Multiplicador de largo plazo
Número índice
Periodo base
Problema de regresión espuria
Proceso de series de tiempo
Proceso estocástico
Propensión de impacto
Propensión de largo plazo (PLP)
Tasa de crecimiento
Tendencia en el tiempo
Tendencia exponencial
Tendencia lineal en el tiempo
Valor base
Variables binarias estacionales

PROBLEMAS

10.1 Decida si está de acuerdo o no con cada uno de los siguientes enunciados y dé una breve explicación de su decisión:

i) Al igual que con las observaciones de corte transversal, suponemos que la mayor parte de las series de tiempo se distribuyen de forma independiente.

ii) El estimador MCO en una regresión de series de tiempo es insesgado bajo los primeros tres supuestos de Gauss-Markov.

iii) Una variable con tendencia no puede usarse como variable dependiente en el análisis de regresión múltiple.

iv) La estacionalidad no es un problema cuando se usan series de tiempo con observaciones anuales.

10.2 Sea $gGDP_t$ el cambio porcentual anual en el producto interno bruto y sea int_t una tasa de interés de corto plazo. Suponga que $gGDP_t$ se relaciona con la tasa de interés mediante la ecuación

$$gGDP_t = \alpha_0 + \delta_0 int_t + \delta_1 int_{t-1} + u_t,$$

donde u_t no se correlaciona con int_t, int_{t-1} ni con todos los demás valores de las tasas de interés. Imagine que la Reserva Federal sigue la regla de política:

$$int_t = \gamma_0 + \gamma_1(gGDP_{t-1} - 3) + v_t,$$

donde $\gamma_1 > 0$. (Cuando el crecimiento del *GDP* del último año está por encima de 3%, la Reserva Federal aumenta las tasas de interés para evitar que la economía se "sobrecaliente".) Si v_t no se correlaciona con todos los valores pasados de int_t y u_t, argumente por qué int_t debe correlacionarse con u_{t-1}. (*Sugerencia*: rezague un periodo la primera ecuación y sustituya $gGDP_{t-1}$ en la segunda ecuación). ¿Qué supuesto de Gauss-Markov viola esto?

10.3 Imagine que y_t sigue un modelo de RDF de segundo orden:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

Sea z^* el *valor de equilibrio* de z_t y y^* el valor de equilibrio de y_t, de modo que

$$y^* = \alpha_0 + \delta_0 z^* + \delta_1 z^* + \delta_2 z^*.$$

Muestre que el cambio en y^*, debido a una modificación en z^*, es igual a la propensión de largo plazo multiplicada por el cambio en z^*:

$$\Delta y^* = PLP \cdot \Delta z^*.$$

Esto da una forma opcional de interpretar el PLP.

10.4 Cuando los tres indicadores de eventos *befile6*, *affile6* y *afdec6* se omiten de la ecuación (10.22), se obtiene $R^2 = .281$ y $\bar{R}^2 = .264$. ¿Los eventos son conjuntamente significativos al nivel de 10 por ciento?

10.5 Suponga que cuenta con datos trimestrales sobre la construcción de viviendas nuevas, la tasa de interés y el ingreso real *per cápita*. Especifique un modelo para la construcción de viviendas que represente la posible tendencia y estacionalidad en las variables.

10.6 En el ejemplo 10.4 se vio que las estimaciones de los coeficientes rezagados individuales en un modelo de rezagos distribuidos eran muy imprecisas. Una forma de aminorar el problema de multicolinealidad es suponer que las δ_j siguen un patrón relativamente simple. En concreto, considere un modelo con cuatro rezagos:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \delta_3 z_{t-3} + \delta_4 z_{t-4} + u_t.$$

Ahora suponga que las δ_j siguen una forma cuadrática en el rezago, j:

$$\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2,$$

para los parámetros γ_0, γ_1 y γ_2. Este es un ejemplo de un *modelo de rezagos distribuidos polinomiales (RDP).*

i) Sustituya cada δ_j en el modelo de rezagos distribuidos y escríbalo en función de los parámetros γ_h, para $h = 0,1,2$.

ii) Explique la regresión que realizaría para estimar las γ_h.

iii) El modelo de rezagos distribuidos polinomiales es una versión restringida del modelo general. ¿Cuántas restricciones se han impuesto? ¿Cómo las probaría? (*Sugerencia:* piense en la prueba F.)

10.7 En el ejemplo 10.4 escribimos el modelo que contiene explícitamente la propensión de largo plazo, θ_0, como

$$gfr_t = \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + u_t,$$

donde se omitieron, por simplicidad, las otras variables explicativas. Como ocurre siempre con el análisis de regresión múltiple, θ_0 debe tener una interpretación *ceteris paribus*. En concreto, si pe_t aumenta uno (un dólar) y $(pe_{t-1} - pe_t)$ y $(pe_{t-2} - pe_t)$ se mantienen fijas, gfr_t debe cambiar en θ_0.

i) Si $(pe_{t-1} - pe_t)$ y $(pe_{t-2} - pe_t)$ se mantienen fijas, pero pe_t está cambiando, ¿qué debe ser cierto respecto a los cambios en pe_{t-1} y pe_{t-2}?

ii) ¿Cómo le ayuda su respuesta del inciso i) a interpretar θ_0 en la ecuación anterior como la PLP?

10.8 En el modelo lineal de la ecuación (10.8), se dice que las variables explicativas $\mathbf{x}_t = (x_{t1}, ..., x_{tk})$ son *secuencialmente exógenas* (o también *débilmente exógenas*) si

$$\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = 0, \; t = 1, 2, \ldots,$$

de manera que los errores son impredecibles dados los valores actuales y todos los valores *pasados* de las variables explicativas.

i) Explique por qué la exogeneidad secuencial está implícita en la exogeneidad estricta.

ii) Explique por qué la exogeneidad contemporánea está implícita en la exogeneidad secuencial.

iii) ¿Los estimadores de MCO por lo general son insesgados bajo el supuesto de exogeneidad secuencial? Explique la razón.

iv) Considere un modelo que explique la tasa anual de infecciones de SIDA como un rezago distribuido del uso de condones *per cápita* para un estado, provincia o región:

$$\mathrm{E}(HIVrate_t|pccon_t, pccont_{t-1}, \ldots,) = \alpha_0 + \delta_0 pccon_t + \delta_1 pccon_{t-1} + \delta_2 pccon_{t-2} + \delta_3 pccon_{t-3}.$$

Explique por qué este modelo satisface el supuesto de exogeneidad secuencial. ¿Le parece probable que también haya exogeneidad estricta?

EJERCICIOS EN COMPUTADORA

C10.1 En octubre de 1979 la Reserva Federal modificó su política de orientación a la oferta monetaria y en su lugar comenzó a concentrarse de forma directa en las tasas de interés a corto plazo. Con base en los datos de INTDEF.RAW, defina una variable binaria igual a 1 para los años posteriores a 1979. Incluya esta binaria en la ecuación (10.15) para ver si hay un desplazamiento en la ecuación de la tasa de interés después de 1979. ¿A qué conclusión llega?

C10.2 Utilice los datos de BARIUM.RAW para este ejercicio.

i) Agregue una tendencia lineal en el tiempo a la ecuación (10.22). ¿Alguna variable distinta de la tendencia es estadísticamente significativa?

ii) En la ecuación estimada en el inciso i), haga una prueba de significancia conjunta de todas las variables, salvo la tendencia en el tiempo. ¿A qué conclusión llega?

iii) Agregue variables binarias mensuales a esta ecuación y determine si hay estacionalidad. ¿Incluir las binarias mensuales modifica de manera importante las otras estimaciones o sus errores estándar?

C10.3 Agregue la variable log(*prgnp*) a la ecuación del salario mínimo en la ecuación (10.38). ¿Esta variable es significativa? Interprete el coeficiente. ¿De qué manera influye el hecho de agregar log(*prgnp*) al efecto estimado del salario mínimo?

C10.4 Utilice los datos de FERTIL3.RAW para verificar que el error estándar de la PLP valga cerca de .030 en la ecuación (10.19).

C10.5 Para este ejercicio use los datos de EZANDERS.RAW. Los datos se refieren al desempleo mensual (*uclms*) en Anderson Township en Indiana, de enero de 1980 a noviembre de 1988. En 1984, una zona empresarial se ubicó en Anderson (así como en otras ciudades de Indiana). [Vea Papke (1994) para los detalles.]

i) Haga la regresión de log(*uclms*) sobre una tendencia lineal y 11 variables binarias mensuales. ¿Cuál fue la tendencia general del desempleo en este periodo? (Interprete el coeficiente de la tendencia en el tiempo.) ¿Hay evidencia de estacionalidad en el desempleo?

ii) Agregue *ez*, una variable binaria igual a 1 en los meses en los que Anderson tuvo una zona empresarial, a la regresión del inciso i). ¿La presencia de la zona empresarial disminuyó el desempleo? ¿Por cuánto? [Debe utilizar la ecuación (7.10) del capítulo 7.]

iii) ¿Qué suposiciones necesita hacer para atribuir el efecto del inciso ii) a la creación de una zona empresarial?

C10.6 Utilice los datos de FERTIL3.RAW para este ejercicio.

i) Haga la regresión de gfr_t sobre t y t^2 y guarde los residuales. Esto da una gfr_t, sin tendencia, es decir, $g\ddot{f}_t$.

ii) Haga la regresión de $g\ddot{f}_t$ sobre todas las variables de la ecuación (10.35), incluidas t y t^2. Compare la R-cuadrada con la de (10.35). ¿A qué conclusión llega?

iii) Vuelva a estimar la ecuación (10.35) pero agregue t^3 a la ecuación. ¿Este término adicional es estadísticamente significativo?

C10.7 Emplee la base de datos CONSUMP.RAW para este ejercicio.

i) Estime un modelo de regresión simple que relacione el crecimiento del consumo real *per cápita* (de no durables y servicios) con el crecimiento del ingreso disponible real *per cápita*. Utilice el cambio en los logaritmos en ambos casos. Reporte los resultados de la manera acostumbrada. Interprete la ecuación y plantee la significancia estadística. (En el archivo, *c* y *y* son consumo e ingreso disponible real per cápita, respectivamente.)

ii) Agregue un rezago del crecimiento del ingreso disponible real *per cápita* a la ecuación del inciso i). ¿A qué conclusión llega sobre los rezagos de ajuste en el crecimiento del consumo?

iii) Agregue la tasa de interés real (*r3*) a la ecuación del inciso i). ¿Influye en el crecimiento del consumo?

C10.8 Utilice los datos de FERTIL3.RAW para este ejercicio.

i) Agregue pe_{t-3} y pe_{t-4} a la ecuación (10.19). Pruebe si estos rezagos son conjuntamente significativos.

ii) Calcule la propensión de largo plazo estimada y su error estándar en el modelo del inciso i). Compárelos con los obtenidos en la ecuación (10.19).

iii) Estime el modelo de rezagos distribuidos polinomiales del problema 10.6. Calcule la PLP estimada y compárela con la que se obtiene del modelo no restringido.

C10.9 Utilice los datos de VOLAT.RAW para este ejercicio. La variable *rsp500* es el rendimiento mensual del índice del mercado accionario de Standard & Poor's 500, a una tasa anual. (Esto incluye cambios en los precios así como en los dividendos.) La variable *i3* es la tasa de rendimiento de las letras del Tesoro estadounidense a tres meses y *pcip* es el cambio porcentual en la producción industrial, éstos también a una tasa anual.

i) Considere la ecuación

$$rsp500_t = \beta_0 + \beta_1 pcip_t + \beta_2 i3_t + u_t.$$

Según usted, ¿qué signos deben tener β_1 y β_2?

ii) Estime la ecuación anterior por MCO y reporte los resultados en la forma usual. Interprete los signos y las magnitudes de los coeficientes.

iii) ¿Cuál de las variables es estadísticamente significativa?

iv) ¿Su conclusión del inciso iii) implica que el rendimiento de Standard & Poor's 500 es predecible? Explique por qué.

C10.10 Considere el modelo estimado de la ecuación (10.15); utilice los datos de INTDEF.RAW.

i) Determine la correlación entre *inf* y *def* durante el mismo periodo muestral y coméntela.

ii) Agregue un solo rezago de *inf* y *def* a la ecuación e informe los resultados como de costumbre.

iii) Compare la PLP estimada para el efecto de la inflación con aquella de la ecuación (10.15). ¿Son muy diferentes?

iv) ¿Los rezagos del modelo al nivel de 5% son conjuntamente significativos?

C10.11 La base de datos TRAFFIC2.RAW contiene 108 observaciones mensuales sobre accidentes automovilísticos, leyes de tránsito y algunas otras variables para California de enero de 1981 a diciembre de 1989. Use esta base de datos para responder a las siguientes preguntas.

i) ¿En qué mes y año entró en vigor la ley del cinturón de seguridad en California ($beltlaw = 1$)? ¿Cuándo aumentó el límite de velocidad en carretera a 65 millas por hora ($spdlaw = 1$)?

ii) Haga una regresión de la variable log(*totacc*) sobre una tendencia lineal en el tiempo y 11 variables binarias mensuales, usando enero como el mes base. Interprete la estimación del coeficiente de la tendencia. ¿Diría usted que hay estacionalidad en el total de accidentes?

iii) Añada a la regresión del inciso ii) las variables *wkends*, *unem*, *spdlaw* y *beltlaw*. Comente el coeficiente sobre la variable de desempleo (*unem*). ¿Su signo y magnitud tienen sentido?

iv) En la regresión del inciso iii) interprete los coeficientes sobre *spdlaw* y *beltlaw*. ¿Los efectos estimados son los que usted esperaba? Explique su respuesta.

v) La variable *prcfat* es el porcentaje de accidentes en los que hay cuando menos un muerto. Advierta que esta variable es un porcentaje, no una proporción. ¿Cuál es el promedio de *prcfat* sobre este periodo? ¿La magnitud parece adecuada?

vi) Realice la regresión del inciso iii) pero usando *prcfat* como la variable dependiente en lugar de log(*totacc*). Comente los efectos estimados y la significancia de las variables de la velocidad y la ley del cinturón de seguridad.

C10.12 i) Estime la ecuación (10.2) usando todos los datos de PHILLIPS.RAW y reporte los resultados como siempre. ¿Cuántas observaciones tiene ahora?

ii) Compare las estimaciones del inciso i) con aquellas de la ecuación (10.14). En particular, ¿la inclusión de los años adicionales ayuda a obtener un efecto de sustitución estimado entre la inflación y el desempleo? Explique su respuesta.

iii) Ahora haga la regresión usando sólo los años de 1997 a 2003. ¿En qué difieren estas estimaciones de aquellas de la ecuación (10.14)? ¿Las estimaciones que utilizan los siete años más recientes son lo suficientemente precisas para permitirle formular una conclusión? Explique.

iv) Considere un sistema de regresión simple en el cual se inicia con n observaciones de series de tiempo y luego se dividen en un periodo anterior y un periodo posterior. En el primer periodo se tienen n_1 observaciones y en el segundo, n_2 observaciones. Recurra a los incisos anteriores de este ejercicio para evaluar el siguiente enunciado: "Por lo general, podemos esperar que la estimación de la pendiente que utiliza las n observaciones sea aproximadamente igual a un promedio ponderado de las estimaciones de la pendiente de la primera y última submuestras, donde los ponderadores son n_1/n y n_2/n, respectivamente".

C10.13 Utilice los datos de MINWAGE.RAW para este ejercicio. En particular, use las series de empleo y salario para el sector 232 (accesorios de vestir de hombres y niños). La variable *gwage232* es el crecimiento mensual (cambio en logaritmos) del salario medio en el sector 232, *gemp232* es el crecimiento del empleo en el sector 232, *gmwage* es el crecimiento del salario mínimo federal y *gcpi* es el crecimiento del Índice de Precios al Consumidor (urbano).

i) Realice la regresión de *gwage232* sobre *gmwage*, *gcpi*. ¿El signo y la magnitud de $\hat{\beta}_{gmwage}$ tienen sentido para usted? Explique por qué. ¿Es *gmwage* estadísticamente significativo?

ii) Agregue los rezagos 1 a 12 de *gmwage* a la ecuación del inciso i). ¿Considera que sea necesario incluir estos rezagos para estimar el efecto a largo plazo del crecimiento del salario mínimo sobre el crecimiento del salario en el sector 232? Explique las razones.

iii) Realice la regresión de *gemp232* sobre *gmwage*,*gcpi*. ¿El crecimiento del salario mínimo parece tener un efecto contemporáneo sobre *gemp232*?

iv) Agregue los rezagos 1 a 12 a la ecuación del crecimiento del empleo. ¿El crecimiento en el salario mínimo tiene un efecto estadísticamente significativo sobre el crecimiento del empleo, ya sea a largo o a corto plazos? Explique por qué.

CAPÍTULO 11

Aspectos adicionales de MCO con datos de series de tiempo

En el capítulo 10 se estudiaron las propiedades en muestras finitas de MCO para datos de series de tiempo bajo un conjunto de supuestos cada vez más fuertes. Según los supuestos del modelo lineal clásico para las series de tiempo, es decir los supuestos ST.1 a ST.6, los MCO tienen *exactamente* las mismas propiedades deseables que se obtuvieron para los datos de corte transversal. Asimismo, la inferencia estadística se realizó de la misma manera que para el análisis de corte transversal.

A partir del análisis de corte transversal del capítulo 5, se sabe que existen buenas razones para estudiar las propiedades de MCO en muestras grandes; por ejemplo, si los términos de error no provienen de una distribución normal, se debe recurrir al teorema del límite central para justificar los estadísticos de prueba y los intervalos de confianza habituales de MCO.

El análisis de muestras grandes es aún más importante en contextos de series de tiempo. (A veces esto resulta un tanto irónico, ya que es difícil conseguir muestras grandes de series de tiempo, pero a menudo la única opción que se tiene es basarse en aproximaciones de muestras grandes). En la sección 10.3, se explicó cómo el supuesto de exogeneidad estricta (ST.3) se infringe en modelos estáticos y de rezagos distribuidos. Como se mostrará en la sección 11.2, los modelos con variables dependientes rezagadas deben violar el supuesto ST.3.

Por desgracia, el análisis de muestras grandes para los problemas de series de tiempo está plagado de un número mucho mayor de dificultades que el análisis de corte transversal. En el capítulo 5 se obtuvieron las propiedades de MCO en muestras grandes en el contexto del muestreo aleatorio. Las cosas se complican más si se permite que las observaciones se correlacionen en el tiempo. Sin embargo, los principales teoremas de límites son válidos para ciertos procesos de series de tiempo, aunque no para todos. La clave es si la correlación entre variables de diferentes periodos de tiempo tiende a cero con suficiente rapidez. Las series de tiempo que tienen una correlación temporal sustancial requieren una atención especial en el análisis de regresión. Este capítulo lo alertará sobre ciertos temas que atañen a tales series en el análisis de regresión.

11.1 Series de tiempo estacionarias y débilmente dependientes

En esta sección se presentan los conceptos clave que se necesitan para aplicar las aproximaciones comunes de muestras grandes en el análisis de regresión con datos de series de tiempo. Los detalles no son tan importantes como la comprensión general de los temas.

Series de tiempo estacionarias y no estacionarias

Desde siempre, la noción de **proceso estacionario** ha cumplido una función importante en el análisis de las series de tiempo. Un proceso de series de tiempo estacionario es aquel en el que sus distribuciones de probabilidad se mantienen estables con el paso del tiempo en el siguiente sentido: si se toma cualquier colección de variables aleatorias de la secuencia y se las desplaza h periodos, la distribución de probabilidad conjunta debe permanecer inalterada. A continuación se dará una definición formal de estacionariedad.

Proceso estocástico estacionario. El proceso estocástico $\{x_t: t = 1, 2, \ldots\}$ es *estacionario* si para cada conjunto de índices temporales $1 \leq t_1 < t_2 < \ldots < t_m$, la distribución conjunta de $(x_{t_1}, x_{t_2}, \ldots, x_{t_m})$ es la misma que la distribución conjunta de $(x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_m+h})$ para todos los enteros $h \geq 1$.

Esta definición es un poco abstracta, pero su significado es muy claro. Una consecuencia (al elegir $m = 1$ y $t_1 = 1$) es que x_t tiene la misma distribución que x_1 para toda $t = 2, 3, \ldots$ Esto significa que la secuencia $\{x_t: t = 1, 2, \ldots\}$ está *idénticamente distribuida*. La estacionariedad requiere aún más. Por ejemplo, la distribución conjunta de (x_1, x_2) (los primeros dos términos en la secuencia) debe ser la misma que la distribución conjunta de (x_t, x_{t+1}) para cualquier $t \geq 1$. De nuevo, esto no impone restricciones a la manera en que x_t y x_{t+1} se relacionan entre ellas; de hecho, podrían tener una correlación muy estrecha. La estacionariedad exige que la naturaleza de cualquier correlación entre términos adyacentes sea la misma en todos los periodos.

Un proceso estocástico que no es estacionario se llama **proceso no estacionario** Puesto que la estacionariedad es un aspecto del proceso estocástico subyacente, no de la única realización disponible, es muy difícil determinar si los datos reunidos fueron generados por un proceso estacionario. Sin embargo, es fácil detectar ciertas secuencias que no son estacionarias. Un proceso con una tendencia en el tiempo del tipo tratado en la sección 10.5 evidentemente no es estacionario: como mínimo, su media cambia con el tiempo.

En ocasiones, una forma más débil de estacionariedad es suficiente. Si $\{x_t: t = 1, 2, \ldots\}$ tiene un segundo momento finito, es decir, $E(x_t^2) < \infty$ para todas las t, entonces se aplica la siguiente definición.

Proceso estacionario en covarianza. Un proceso estocástico $\{x_t: t = 1, 2, \ldots\}$ con un segundo momento finito $[E(x_t^2) < \infty]$ es **estacionario en covarianza** si i) $E(x_t)$ es constante; ii) $\text{Var}(x_t)$ es constante; y iii) para cualquier t, $h \geq 1$, la $\text{Cov}(x_t, x_{t+h})$ depende sólo de h no de t.

Pregunta 11.1

Suponga que $\{y_t: t = 1, 2, \ldots\}$ es generado por $y_t = \delta_0 + \delta_1 t + e_t$, donde $\delta_1 \neq 0$ y $\{e_t: t = 1, 2, \ldots\}$ es una secuencia i.i.d. con media cero y varianza σ_e^2. i) ¿Es $\{y_t\}$ estacionaria en covarianza? ii) ¿Es $y_t - E(y_t)$ estacionaria en covarianza?

Los procesos estacionarios en covarianza se centran sólo en los dos primeros momentos del proceso estocástico: la media y la varianza del proceso son constantes en el tiempo y la covarianza entre x_t y x_{t+h} depende sólo de la distancia entre los dos términos, h, y no de la ubicación del periodo inicial, t. Se deduce de inmediato que la correlación entre x_t y x_{t+h} depende también sólo de h.

Si un proceso estacionario tiene un segundo momento finito, debe ser estacionario en covarianza, pero desde luego lo contrario no es verdadero. En ocasiones, para destacar que la estacionariedad es un requisito con mayor peso que la estacionariedad en covarianza, a la primera se le llama *estacionariedad estricta*. Como esta última simplifica los enunciados de algunos de nuestros supuestos siguientes, cuando se hable de "estacionariedad" siempre se hará referencia a la forma estricta.

¿Cómo se usa la estacionariedad en la econometría de las series de tiempo? En un nivel técnico, la estacionariedad simplifica los enunciados de la ley de los grandes números y el teorema del límite central, aun cuando en este capítulo no nos preocupemos por los enunciados formales. En un nivel práctico, si se quiere entender la relación entre dos o más variables que utilizan el análisis de regresión, se necesita dar por sentada algún tipo de estabilidad en el tiempo. Si se permite que la relación entre dos variables (por ejemplo, y_t y x_t) cambie de forma arbitraria en cada periodo, no se puede esperar aprender mucho acerca de cómo un cambio en una variable afecta a la otra si sólo se tiene acceso a una sola realización de serie de tiempo.

Al establecer un modelo de regresión múltiple para los datos de series de tiempo, se está suponiendo cierta forma de estacionariedad en la que la β_j no cambia en el tiempo. Además, los supuestos ST.4 y ST.5 implican que la varianza del proceso de error es constante en el tiempo y la correlación entre los errores en dos periodos adyacentes es igual a cero, lo cual claramente es constante en el tiempo.

Series de tiempo débilmente dependientes

La estacionariedad tiene que ver con las distribuciones conjuntas de un proceso a medida que avanza en el tiempo. Un concepto muy distinto es la dependencia débil, la cual impone restricciones a qué tan estrecha puede ser la relación entre las variables x_t y x_{t+h} a medida que aumenta la distancia temporal entre ellas, h. La noción de dependencia débil es más fácil de tratar para una serie de tiempo estacionaria: en términos generales, se dice que un proceso de serie de tiempo estacionario $\{x_t: t = 1, 2, \dots\}$ es **débilmente dependiente** si x_t y x_{t+h} son "casi independientes" a medida que h aumenta sin límite. Un enunciado parecido es verdadero si la secuencia no es estacionaria, pero entonces se debe asumir que el concepto de casi independientes no depende del punto de partida, t.

La descripción de la dependencia débil que se expuso en el párrafo anterior es, por fuerza, imprecisa. No se puede definir de manera formal la dependencia débil, porque ninguna definición cubre todos los casos que interesan. Hay muchas formas específicas de dependencia débil definidas formalmente, pero están fuera del alcance de este libro. [Vea White (1984), Hamilton (1994) y Wooldridge (1994b) para un tratamiento avanzado de estos conceptos.]

Para nuestros propósitos, una noción intuitiva del significado de la dependencia débil es suficiente. Las secuencias estacionarias en covarianza pueden describirse en términos de las correlaciones: una serie de tiempo estacionaria en covarianza es débilmente dependiente si la correlación entre x_t y x_{t+h} se vuelve cero con la "suficiente rapidez" cuando $h \to \infty$. (Debido a la estacionariedad en covarianza, las correlaciones no dependen del punto de partida, t.) En otras palabras, a medida que las variables se distancian en el tiempo, la correlación entre ellas se vuelve cada vez más pequeña. Las secuencias estacionarias en covarianza donde la $\mathrm{Corr}(x_t, x_{t+h}) \to 0$ cuando $h \to \infty$ se dice que están **no correlacionadas asintóticamente**. Así es como se define, de manera intuitiva, la dependencia débil. Desde un punto de vista técnico, es necesario dar por sentado que la correlación converge a cero con suficiente rapidez, pero se pasará esto por alto.

¿Por qué la dependencia débil es importante para el análisis de regresión? En principio, porque reemplaza el supuesto del muestreo aleatorio al dar por sentado que la ley de los grandes números (LGN) y el teorema del límite central (TLC) son válidos. El teorema del límite central más conocido para los datos de series de tiempo requiere estacionariedad y alguna forma de dependencia débil: por tanto, las series de tiempo estacionarias y débilmente dependientes son ideales para el análisis de regresión múltiple. En la sección 11.2 se expone que los MCO se pueden justificar de una manera muy general al recurrir a la LGN y el TLC. Las series de tiempo que no son débilmente dependientes —se verán ejemplos de ellas en la sección 11.3— por lo común no cumplen con el TLC, razón por la cual su uso en el análisis de regresión múltiple resulta problemático.

El ejemplo más sencillo de una serie de tiempo débilmente dependiente es una secuencia independiente idénticamente distribuida: una secuencia independiente tiene una dependencia débil insignificante. Un ejemplo más interesante de secuencia débilmente dependiente es

$$x_t = e_t + \alpha_1 e_{t-1},\ t = 1, 2, \ldots, \tag{11.1}$$

donde $\{e_t: t = 0, 1, \ldots\}$ es una secuencia i.i.d. con media cero y varianza σ_e^2. El proceso $\{x_t\}$ se llama **proceso de promedio móvil de orden uno [MA(1), del inglés *moving average*]**: x_t es un promedio ponderado de e_t y e_{t-1}; en el periodo siguiente, se omite e_{t-1} y por consiguiente x_{t+1} depende de e_{t+1} y e_t. Establecer el coeficiente de e_t como igual a 1 en la ecuación (11.1) no conlleva una pérdida de generalidad. [En la ecuación (11.1), se usan x_t y e_t como etiquetas genéricas para los procesos de series de tiempo. No necesariamente hacen referencia a las variables explicativas o a los errores en un modelo de regresión de series de tiempo, aunque tanto las variables explicativas como los errores podrían ser procesos MA(1).]

¿Por qué un proceso MA(1) es débilmente dependiente? Los términos adyacentes en la secuencia están correlacionados: como $x_{t+1} = e_{t+1} + \alpha_1 e_t$, entonces $\text{Cov}(x_t, x_{t+1}) = \alpha_1 \text{Var}(e_t) = \alpha_1 \sigma_e^2$. Puesto que $\text{Var}(x_t) = (1 + \alpha_1^2)\sigma_e^2$, entonces $\text{Corr}(x_t, x_{t+1}) = \alpha_1/(1 + \alpha_1^2)$. Por ejemplo, si $\alpha_1 = .5$, entonces $\text{Corr}(x_t, x_{t+1}) = .4$. [La máxima correlación positiva ocurre cuando $\alpha_1 = 1$, en cuyo caso, $\text{Corr}(x_t, x_{t+1}) = .5$.] Sin embargo, una vez que se consideran las variables de la secuencia que están separadas por dos periodos o más, éstas no tienen una correlación porque son independientes. Por ejemplo, $x_{t+2} = e_{t+2} + \alpha_1 e_{t+1}$ es independiente de x_t porque $\{e_t\}$ son independientes a través de t. Debido al supuesto de distribución idéntica de las e_t, $\{x_t\}$ en la ecuación (11.1) es, de hecho, estacionaria. De esta manera, un MA(1) es una secuencia estacionaria débilmente dependiente y la ley de los grandes números y el teorema del límite central pueden aplicarse a $\{x_t\}$.

Un ejemplo más conocido es el proceso

$$y_t = \rho_1 y_{t-1} + e_t,\ t = 1, 2, \ldots. \tag{11.2}$$

El punto de partida en la secuencia es y_0 (en $t = 0$) y $\{e_t: t = 1, 2, \ldots\}$ es una secuencia i.i.d. con media cero y varianza σ_e^2. También se supone que las e_t son independientes de y_0 así como que $\text{E}(y_0) = 0$. A esto se le llama **proceso autorregresivo de orden uno [AR(1)]**.

El supuesto decisivo para la dependencia débil de un proceso AR(1) es la *condición de estabilidad* $|\rho_1| < 1$. Por ende, se dice que $\{y_t\}$ es un **proceso estable AR(1)**.

Para ver que un proceso estable AR(1) es asintóticamente no correlacionado, es útil suponer que el proceso es estacionario en covarianza. (De hecho, por lo general puede mostrarse que $\{y_t\}$ es estrictamente estacionaria, pero la demostración es un tanto técnica.) De modo que se sabe que $\text{E}(y_t) = \text{E}(y_{t-1})$, y por la ecuación (11.2) con $\rho_1 \neq 1$, esto puede ocurrir sólo si $\text{E}(y_t) = 0$. Calculando la varianza de la ecuación (11.2) y basándonos en el hecho de que e_t y y_{t-1} son independientes (y por ende no tienen correlación), se llega a que $\text{Var}(y_t) = \rho_1^2 \text{Var}(y_{t-1}) + \text{Var}(e_t)$, y por tanto, bajo la estacionariedad en covarianza, se debe tener $\sigma_y^2 = \rho_1^2 \sigma_y^2 + \sigma_e^2$. Dado que $\rho_1^2 < 1$ por la condición de estabilidad, se puede encontrar con facilidad σ_y^2:

$$\sigma_y^2 = \sigma_e^2/(1 - \rho_1^2). \tag{11.3}$$

Ahora, se puede encontrar la covarianza entre y_t y y_{t+h} para $h \geq 1$. Usando la sustitución repetida,

$$\begin{aligned} y_{t+h} &= \rho_1 y_{t+h-1} + e_{t+h} = \rho_1(\rho_1 y_{t+h-2} + e_{t+h-1}) + e_{t+h} \\ &= \rho_1^2 y_{t+h-2} + \rho_1 e_{t+h-1} + e_{t+h} = \ldots \\ &= \rho_1^h y_t + \rho_1^{h-1} e_{t+1} + \ldots + \rho_1 e_{t+h-1} + e_{t+h}. \end{aligned}$$

Como $E(y_t) = 0$ para toda t, se puede multiplicar esta última ecuación por y_t y tomar las esperanzas para obtener $Cov(y_t, y_{t+h})$. Tomando en cuenta el hecho de que e_{t+j} no guarda una correlación con y_t para toda $j \geq 1$ se tiene

$$\begin{aligned} Cov(y_t, y_{t+h}) = E(y_t y_{t+h}) &= \rho_1^h E(y_t^2) + \rho_1^{h-1} E(y_t e_{t+1}) + \ldots + E(y_t e_{t+h}) \\ &= \rho_1^h E(y_t^2) = \rho_1^h \sigma_y^2. \end{aligned}$$

Como σ_y es la desviación estándar tanto de y_t como de y_{t+h}, se puede calcular con facilidad la correlación entre y_t y y_{t+h} para cualquier $h \geq 1$:

$$Corr(y_t, y_{t+h}) = Cov(y_t, y_{t+h})/(\sigma_y \sigma_y) = \rho_1^h. \quad \boxed{11.4}$$

En particular, $Corr(y_t, y_{t+1}) = \rho_1$, de manera que ρ_1 es el coeficiente de correlación entre cualesquiera dos términos adyacentes en la secuencia.

La ecuación (11.4) es importante debido a que muestra que, aun cuando y_t y y_{t+h} estén correlacionadas para cualquier $h \geq 1$, esta correlación se vuelve muy pequeña para las h grandes: como $|\rho_1| < 1$, entonces $\rho_1^h \to 0$ cuando $h \to \infty$. Aunque ρ_1 sea grande —digamos .9, lo que implica una correlación positiva muy grande entre los términos adyacentes— la correlación entre y_t y y_{t+h} tiende a cero con bastante rapidez. Por ejemplo, $Corr(y_t, y_{t+5}) = .591$, $Corr(y_t, y_{t+10}) = .349$ y $Corr(y_t, y_{t+20}) = .122$. Si t indexa los años, esto significa que la correlación entre el resultado de dos y entre las cuales hay 20 años de diferencia es cerca de .122. Cuando ρ_1 es menor, la correlación se extingue mucho más rápido. (Podría probar con $\rho_1 = .5$ para verificar esto.)

Este análisis demuestra de forma heurística que un proceso estable AR(1) es débilmente dependiente. El modelo AR(1) tiene particular importancia en el análisis de regresión múltiple con los datos de series de tiempo. En el capítulo 12 se estudiarán otras aplicaciones del mismo y en el capítulo 18 se verá su uso en la obtención de pronósticos.

Existen muchos otros tipos de series de tiempo débilmente dependientes, incluidos los híbridos de procesos autorregresivos y de promedio móvil. Pero los ejemplos anteriores son adecuados para nuestros propósitos.

Antes de concluir esta sección, hay que hacer hincapié en un punto que a menudo crea confusión en la econometría de las series de tiempo. Una serie con tendencia, aun cuando no sea estacionaria, *puede* ser débilmente dependiente. De hecho, en el modelo simple de tendencia lineal en el tiempo que se vio en el capítulo 10 [vea la ecuación (10.24)], la serie $\{y_t\}$ en realidad era independiente. Una serie que es estacionaria alrededor de su tendencia en el tiempo y que además es débilmente dependiente se conoce como **proceso estacionario con tendencia**. (Observe que el nombre no es descriptivo por completo debido a que se supone una dependencia débil junto con la estacionariedad.) Estos procesos se pueden utilizar en el análisis de regresión como se hizo en el capítulo 10, *siempre y cuando* se incluyan en el modelo las tendencias en el tiempo apropiadas.

11.2 Propiedades asintóticas de MCO

En el capítulo 10, se vieron algunos casos en los cuales los supuestos del modelo lineal clásico no se cumplían con ciertos problemas de series de tiempo. En estos casos, se debe recurrir a las propiedades en muestras grandes de MCO, del mismo modo que con el análisis de corte transversal. En esta sección se definen los supuestos y se determinan los principales resultados que justifican a los MCO de manera más general. Las demostraciones de los teoremas de este capítulo son un poco más complicadas y, por tanto se omitieron. Vea Wooldridge (1994b).

Supuesto ST.1′ (Linealidad y dependencia débil)

Se supone que el modelo es exactamente el mismo que en el supuesto ST.1, pero ahora se añade el supuesto de que $\{(\mathbf{x}_t, y_t): t = 1, 2, \ldots\}$ es estacionario y débilmente dependiente. En particular, la ley de los grandes números y el teorema del límite central pueden aplicarse a los promedios muestrales.

El requisito de linealidad en los parámetros de nuevo significa que se puede escribir el modelo como

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \tag{11.5}$$

donde las β_j son los parámetros que se van a estimar. A diferencia del capítulo 10, las x_{tj} pueden incluir rezagos de la variable dependiente. Como siempre, también se permiten rezagos en las variables explicativas.

Se ha incluido estacionariedad en el supuesto ST.1′ porque resulta conveniente para establecer e interpretar los supuestos. Si se derivaran con cuidado las propiedades asintóticas de MCO, como se hace en el apéndice E, la estacionariedad también simplificaría dichas deducciones. Pero la estacionariedad en absoluto es crítica para que los MCO tengan sus propiedades asintóticas usuales. (Como se mencionó en la sección 11.1, al suponer que las β_j son constantes en el tiempo, ya se está dando por sentada alguna forma de estabilidad en las distribuciones en el tiempo.) La restricción adicional importante en el supuesto ST.1′ al compararlo con el supuesto ST.1 es la premisa de la dependencia débil. En la sección 11.1, se invirtió una considerable cantidad de tiempo en el análisis de la dependencia débil, porque de ninguna manera es un supuesto inocuo. En la siguiente sección se presentarán los procesos de series de tiempo que violan claramente el supuesto de la dependencia débil y también se estudiará el uso de estos procesos en los modelos de regresión múltiple.

Por supuesto, se seguirá descartando la colinealidad perfecta.

Supuesto ST.2′ (No hay colinealidad perfecta)

Igual al supuesto ST.2.

Supuesto ST.3′ (Media condicional cero)

Las variables explicativas $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ son **contemporáneamente exógenas** como en la ecuación (10.10): $\mathrm{E}(u_t|\mathbf{x}_t) = 0$.

Este es el supuesto más natural concerniente a la relación entre u_t y las variables explicativas. Es mucho más débil que el supuesto ST.3 ya que no impone restricciones sobre cómo se relaciona u_t con las variables explicativas en otros periodos. En breve se verán ejemplos que satisfacen el supuesto ST.3′. Por la estacionariedad, si la exogeneidad contemporánea es válida para un periodo, es válida para todos. Si se relaja la estacionariedad, sencillamente se necesitaría suponer que la condición es válida para toda $t = 1, 2, \ldots$.

Para ciertos propósitos, es conveniente saber que el siguiente resultado de la consistencia sólo requiere que u_t tenga una media no condicional cero y no guarde una correlación con cada x_{tj}:

$$\mathrm{E}(u_t) = 0, \ \mathrm{Cov}(x_{tj}, u_t) = 0, \ j = 1, \ldots, k. \tag{11.6}$$

Se trabajará sobre todo con el supuesto de la media condicional cero, ya que conduce al análisis asintótico más sencillo.

Teorema 11.1 (Consistencia de MCO)

Bajo los supuestos ST.1′, ST.2′ y ST.3′, los estimadores de MCO son consistentes: plim $\hat{\beta}_j = \beta_j$, $j = 0, 1, \ldots, k$.

Existen algunas diferencias prácticas entre los teoremas 10.1 y 11.1. Primero, en el teorema 11.1, se concluye que los estimadores de MCO son consistentes, pero no necesariamente insesgados. Segundo, en el teorema 11.1 se ha debilitado el sentido en el cual las variables explicativas deben ser exógenas, pero la dependencia débil se requiere en las series de tiempo subyacentes. La dependencia débil también es crucial para obtener resultados de distribución aproximados, los cuales se cubrirán más adelante.

Ejemplo 11.1

[Modelo estático]

Considere un modelo estático con dos variables explicativas:

$$y_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + u_t. \tag{11.7}$$

Bajo la dependencia débil, la condición suficiente para la consistencia de MCO es

$$\mathrm{E}(u_t|z_{t1}, z_{t2}) = 0. \tag{11.8}$$

Esto excluye que haya variables omitidas contenidas en u_t que se correlacionen, ya sea con z_{t1} o z_{t2}. Además, ninguna función de z_{t1} o z_{t2} puede correlacionarse con u_t, y por consiguiente el supuesto ST.3′ excluye la especificación incorrecta de la forma funcional, como en el caso del corte transversal. Otros problemas, como el error de medición en las variables z_{t1} o z_{t2}, provocan que no se cumpla la ecuación 11.8.

En gran medida, el supuesto ST.3′ *no* descarta la correlación entre, digamos, u_{t-1} y z_{t1}. Este tipo de correlación podría surgir si z_{t1} se relaciona con las y_{t-1}, pasadas, como sigue

$$z_{t1} = \delta_0 + \delta_1 y_{t-1} + v_t. \tag{11.9}$$

Por ejemplo, z_{t1} podría ser una variable de política, como el cambio porcentual mensual en el dinero en circulación, y este cambio depende de la tasa de inflación del último mes (y_{t-1}). Un mecanismo como éste por lo general provoca que z_{t1} y u_{t-1} se correlacionen (como puede verse al sustituir y_{t-1}). Este tipo de reacción *está* permitido bajo el supuesto ST.3′.

Ejemplo 11.2

[Modelo de rezagos distribuidos finitos]

En el modelo de rezagos distribuidos finitos,

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \tag{11.10}$$

un supuesto muy natural es que el valor esperado de u_t, dados los valores actuales y *todos los valores pasados* de z, es cero:

$$\mathrm{E}(u_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = 0. \tag{11.11}$$

Esto significa que, una vez que z_t, z_{t-1} y z_{t-2} se incluyan, ningún rezago posterior de z afectará a $E(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots)$; si esto no fuera cierto, se añadirían más rezagos en la ecuación. Por ejemplo, y_t podría ser el cambio porcentual anual en la inversión y z_t una medida de las tasas de interés durante el año t. Cuando se establece $\mathbf{x}_t = (z_t, z_{t-1}, z_{t-2})$, el supuesto ST.3′ se satisface en consecuencia: los MCO serán consistentes. Al igual que en el ejemplo anterior, ST.3′ no descarta que y influya los valores futuros de z.

No es forzoso que los dos ejemplos anteriores requieran la teoría asintótica debido a que las variables explicativas *podrían* ser estrictamente exógenas. El siguiente ejemplo viola de manera clara el supuesto de exogeneidad estricta, por tanto, sólo se puede recurrir a las propiedades de MCO en muestras grandes.

Ejemplo 11.3

[El modelo AR(1)]

Considere el modelo AR(1),

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \tag{11.12}$$

donde el error u_t tiene un valor esperado de cero, dados todos los valores pasados de y:

$$E(u_t|y_{t-1}, y_{t-2}, \ldots) = 0. \tag{11.13}$$

Combinadas, estas dos ecuaciones implican que

$$E(y_t|y_{t-1}, y_{t-2}, \ldots) = E(y_t|y_{t-1}) = \beta_0 + \beta_1 y_{t-1}. \tag{11.14}$$

Este resultado es muy importante. Primero, significa que una vez que la y rezagada un periodo se ha controlado, ningún rezago posterior de y afecta el valor esperado de y_t. (Aquí es donde se origina el nombre de "primer orden".) Segundo, se supone que la relación es lineal.

Como la $\mathbf{x}_t$ contiene sólo a y_{t-1}, la ecuación (11.13) implica que el supuesto ST.3′ es válido. Por el contrario, el supuesto de exogeneidad estricta necesario para el insesgamiento, es decir el supuesto ST.3, no es válido. Puesto que el conjunto de variables explicativas para todos los periodos incluye todos los valores de y excepto el último $(y_0, y_1, \ldots, y_{n-1})$, el supuesto ST.3 requiere que para toda t, u_t no guarde una correlación con cada una de $y_0, y_1, \ldots, y_{n-1}$. Esto no puede ser cierto. De hecho, dado que u_t no tiene correlación con y_{t-1} bajo la ecuación (11.13), u_t y y_t deben estar correlacionadas. De hecho, se aprecia con facilidad que $\mathrm{Cov}(y_t, u_t) = \mathrm{Var}(u_t) > 0$. Por tanto, un modelo con una variable dependiente rezagada no puede satisfacer el supuesto de exogeneidad estricta ST.3.

Para que la condición de dependencia débil sea válida, se debe dar por sentado que $|\beta_1| < 1$, como se vio en la sección 11.1. Si esta condición se mantiene, entonces el teorema 11.1 implica que para la regresión de y_t sobre y_{t-1} MCO produce estimadores consistentes de β_0 y β_1. Por desgracia, $\hat{\beta}_1$ es sesgado y su sesgo puede ser grande si el tamaño muestral es pequeño o si β_1 está cerca de 1. (Para β_1 cercano a 1, $\hat{\beta}_1$ puede mostrar un marcado sesgo hacia abajo.) En las muestras medianas a grandes, $\hat{\beta}_1$ debe ser un buen estimador de β_1.

Para usar procedimientos de inferencia estándar, es necesario imponer versiones de los supuestos de homocedasticidad y errores sin correlación serial. Éstos son menos restrictivos que sus equivalentes en el modelo lineal clásico del capítulo 10.

Supuesto ST.4′ (Homocedasticidad)

Los errores son **contemporáneamente homocedásticos**, es decir, $\mathrm{Var}(u_t|\mathbf{x}_t) = \sigma^2$.

Supuesto ST.5′ (No hay correlación serial)

Para toda $t \neq s$, $E(u_t u_s | \mathbf{x}_t, \mathbf{x}_s) = 0$.

En el ST.4′, observe cómo se condiciona sólo en las variables explicativas en el periodo t (en comparación con ST.4). En ST.5′, sólo se condiciona en las variables explicativas en periodos que coinciden con u_t y u_s. Como se explicó antes, este supuesto es un poco difícil de interpretar, pero es la condición correcta para estudiar las propiedades en muestras grandes de MCO en una variedad de regresiones de series de tiempo. Cuando se considera el supuesto ST.5′, a menudo se ignora el condicionamiento sobre $\mathbf{x}_t$ y $\mathbf{x}_s$, y se piensa sobre si u_t y u_s no guardan una correlación, para toda $t \neq s$.

Con frecuencia la correlación serial es un problema en los modelos de regresión estáticos y con rezagos distribuidos finitos: nada garantiza que las u_t inobservables no estén correlacionadas en el tiempo. Resulta importante que el supuesto ST.5′ *sí* es válido en el modelo AR(1) definido en las ecuaciones (11.12) y (11.13). Dado que la variable explicativa en el periodo t es y_{t-1}, se debe mostrar que $E(u_t u_s | y_{t-1}, y_{s-1}) = 0$ para toda $t \neq s$. Para ver esto, imagine que $s < t$. (El otro caso se deduce por simetría.) Como $u_s = y_s - \beta_0 - \beta_1 y_{s-1}$, entonces u_s es una función de y con una fecha anterior al periodo t. Pero según la ecuación (11.13), $E(u_t | u_s, y_{t-1}, y_{s-1}) = 0$, y por consiguiente $E(u_t u_s | u_s, y_{t-1}, y_{s-1}) = u_s E(u_t | y_{t-1}, y_{s-1}) = 0$. Por la ley de las esperanzas iteradas (ver apéndice B), $E(u_t u_s | y_{t-1}, y_{s-1}) = 0$. Esto es muy importante: siempre que haya un solo rezago en la ecuación (11.12), los errores no deben estar correlacionados. Se estudiará esta característica de los modelos dinámicos de una manera más general en la sección 11.4.

Ahora se obtiene un resultado asintótico que prácticamente es idéntico al caso del corte transversal.

Teorema 11.2 (Normalidad asintótica de MCO)

Bajo los supuestos ST.1′ a ST.5′, los estimadores de MCO tienen distribuciones asintóticamente normales. Además, los errores estándar usuales de MCO, los estadísticos t, los estadísticos F y los estadísticos *ML* son asintóticamente válidos.

Este teorema proporciona una justificación adicional para cuando menos algunos de los ejemplos estimados en el capítulo 10: incluso si los supuestos del modelo lineal clásico no son válidos, los estimadores de MCO siguen siendo consistentes, y los procedimientos de inferencia usuales son válidos. Desde luego, esto depende de que los supuestos ST.1′ a ST.5′ sean verdaderos. En la siguiente sección, se analizarán casos en que el supuesto de la dependencia débil puede fallar. Los problemas de la correlación serial y la heterocedasticidad se tratan en el capítulo 12.

Ejemplo 11.4

[Hipótesis de los mercados eficientes]

Se puede utilizar el análisis asintótico para probar una versión de la *hipótesis de los mercados eficientes* (HME). Sea y_t el rendimiento porcentual semanal (del cierre del miércoles al cierre del miércoles) en el índice compuesto de la Bolsa de Nueva York (NYSE). Una forma estricta de la hipótesis de los mercados eficientes establece que la información observable para el mercado anterior a la semana t no debe ayudar a predecir el rendimiento durante la semana t. Si sólo se usa información pasada de y, la HME se define como

$$E(y_t | y_{t-1}, y_{t-2}, \ldots) = E(y_t). \tag{11.15}$$

Si la ecuación (11.15) es falsa, entonces se podría usar la información sobre los rendimientos semanales pasados para predecir el rendimiento actual. La HME supone que estas oportunidades de inversión serán advertidas y desaparecerán casi al instante.

Una forma simple de probar la ecuación (11.15) es especificar el modelo AR(1) en la ecuación (11.12) como el modelo alternativo. De esta manera, la hipótesis nula se establece con facilidad como H_0: $\beta_1 = 0$. Bajo la hipótesis nula, el supuesto ST.3′ es verdadero por la ecuación (11.15) y, como antes se vio, la correlación serial no es un problema. El supuesto de homocedasticidad es $\text{Var}(y_t|y_{t-1}) = \text{Var}(y_t) = \sigma^2$, que por el momento se supondrá que es verdadero. Bajo la hipótesis nula, los rendimientos de las acciones no están correlacionados serialmente, así que se puede dar por sentado que son débilmente dependientes. Por tanto, el teorema 11.2 establece que se puede utilizar el estadístico t usual de MCO para $\hat{\beta}_1$ para probar H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$.

Los rendimientos semanales en NYSE.RAW se calcularon usando datos de enero de 1976 a marzo de 1989. En el caso excepcional de que el miércoles fuera un día feriado, se tomó el cierre del siguiente día bursátil. El rendimiento promedio semanal (*return*) durante este periodo fue de .196 en forma porcentual, siendo el rendimiento semanal mayor de 8.45% y el menor de −15.32% (durante el crac del mercado de valores de octubre de 1987). La estimación del modelo AR(1) da

$$\begin{aligned}\widehat{return}_t &= \underset{(.081)}{.180} + \underset{(.038)}{.059}\, return_{t-1} \\ n &= 689,\ R^2 = .0035,\ \bar{R}^2 = .0020.\end{aligned} \tag{11.16}$$

El estadístico t para el coeficiente de $return_{t-1}$ es cerca de 1.55, y por ende H_0: $\beta_1 = 0$ no puede rechazarse contra la alternativa de dos colas, incluso a un nivel de significancia de 10%. La estimación sugiere una correlación ligeramente positiva en el rendimiento del NYSE de una semana a la siguiente, pero no suficientemente sólida para garantizar el rechazo de la hipótesis de los mercados eficientes.

En el ejemplo anterior, el uso de un modelo AR(1) para probar la HME tal vez no detectó la correlación entre los rendimientos semanales que están separados por más de una semana. Es fácil estimar modelos con más de un rezago. Por ejemplo, un *modelo autorregresivo de orden dos*, es decir un modelo AR(2), es

$$\begin{aligned}y_t &= \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_t \\ \text{E}(u_t|y_{t-1}, y_{t-2}, \dots) &= 0.\end{aligned} \tag{11.17}$$

Existen condiciones de estabilidad sobre β_1 y β_2 necesarias para garantizar que el proceso AR(2) sea débilmente dependiente, pero esto no constituye aquí un problema porque la hipótesis nula establece que la HME sigue siendo válida:

$$H_0\!: \beta_1 = \beta_2 = 0. \tag{11.18}$$

Si se añade el supuesto de homocedasticidad $\text{Var}(u_t|y_{t-1}, y_{t-2}) = \sigma^2$, se puede usar un estadístico F estándar para probar la ecuación (11.18). Si se estima un modelo AR(2) para $return_t$, se obtiene

$$\begin{aligned}\widehat{return}_t &= \underset{(.081)}{.186} + \underset{(.038)}{.060}\, return_{t-1} - \underset{(.038)}{.038}\, return_{t-2} \\ n &= 688,\ R^2 = .0048,\ \bar{R}^2 = .0019\end{aligned}$$

(donde se pierde una observación más debido al rezago adicional en la ecuación). Los dos rezagos son insignificantes por separado al nivel de 10%. También son conjuntamente insignificantes: al usar $R^2 = .0048$, el estadístico F es cerca de $F = 1.65$; el valor-p para este estadístico F (con 2 y 685 grados de libertad) es cercano a .193. Por tanto, no se rechaza la ecuación (11.18) ni siquiera al nivel de significancia de 15%.

Ejemplo 11.5

[Curva de Phillips aumentada por las expectativas]

Una versión lineal de la *curva de Phillips aumentada por las expectativas* puede escribirse como

$$inf_t - inf_t^e = \beta_1(unem_t - \mu_0) + e_t,$$

donde μ_0 es la *tasa natural de desempleo* e inf_t^e es la tasa de desempleo *esperada* formada en el año $t - 1$. Este modelo supone que la tasa natural es constante, algo que los macroeconomistas ponen en duda. La diferencia entre la tasa de desempleo real y la tasa natural se llama *desempleo cíclico*, mientras que la diferencia entre la inflación real y la esperada se llama *inflación no anticipada*. Los macroeconomistas llaman al término de error, e_t, *choque de oferta*. Si hay un efecto de sustitución entre la inflación no anticipada y el desempleo cíclico, entonces $\beta_1 < 0$. [Para un análisis detallado de la curva de Phillips aumentada por las expectativas consulte Mankiw (1994, sección 11.2).]

Para completar este modelo, es necesario hacer una suposición respecto a las expectativas inflacionarias. Bajo las *expectativas adaptativas*, el valor esperado de la inflación actual depende de la inflación recién observada. Una formulación particularmente sencilla es que la inflación esperada este año es la inflación del último año: $inf_t^e = inf_{t-1}$. (Consulte la sección 18.1 para obtener otra formulación de las expectativas adaptativas.) Bajo este supuesto, se puede escribir

$$inf_t - inf_{t-1} = \beta_0 + \beta_1 unem_t + e_t$$

o

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t,$$

donde $\Delta inf_t = inf_t - inf_{t-1}$ y $\beta_0 = -\beta_1\mu_0$. (Se espera que β_0 sea positivo, ya que $\beta_1 < 0$ y $\mu_0 > 0$.) Por consiguiente, bajo expectativas adaptativas, la curva de Phillips aumentada por las expectativas relaciona el *cambio* en la inflación con el nivel de desempleo y un choque de oferta, e_t. Si e_t no tiene correlación con $unem_t$, como por lo general se supone, entonces se puede estimar de manera consistente β_0 y β_1 mediante MCO. (No se tiene que dar por hecho, por ejemplo, que el choque de oferta actual no afecta las tasas de desempleo futuras.) Suponga que los ST.1′ a ST.5′ son válidos. Usando los datos hasta 1996 de la base de datos PHILLIPS.RAW se estima

$$\begin{aligned}\widehat{\Delta inf_t} &= \underset{(1.38)}{3.03} - \underset{(.230)}{.543}\, unem_t \\ n &= 48,\ R^2 = .108,\ \bar{R}^2 = .088.\end{aligned} \tag{11.19}$$

El equilibrio entre el desempleo cíclico y la inflación no anticipada es marcado en la ecuación (11.19): un incremento de un punto en *unem* reduce la inflación no anticipada más de medio punto. El efecto es estadísticamente significativo (valor-p de dos colas $\approx$ .023). Se puede contrastar esto con la curva estática de Phillips del ejemplo 10.1, donde se encontró una relación ligeramente positiva entre la inflación y el desempleo.

Como la tasa natural se puede escribir como $\mu_0 = \beta_0/(-\beta_1)$, se usa la ecuación (11.19) para obtener nuestra estimación de la tasa natural: $\hat{\mu}_0 = \hat{\beta}_0/(-\hat{\beta}_1) = 3.03/.543 \approx 5.58$. Así que se estima que la tasa natural es cerca de 5.6, lo cual entra en el rango sugerido por los macroeconomistas: de manera histórica, 5% a 6% es un rango común citado para la tasa natural de desempleo. Un error estándar de esta estimación es difícil de obtener debido a que se tiene una función no lineal de los estimadores de MCO. En Wooldridge (2002, capítulo 3) se contiene la teoría para las funciones no lineales generales. En la aplicación actual, el error estándar es .657, lo cual nos lleva a un intervalo de confianza asintótico de 95% (que se basa en la distribución normal estándar) de alrededor de 4.29 a 6.87 para la tasa natural.

Pregunta 11.2

Suponga que las expectativas se forman como $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$. ¿Qué regresión realizaría para estimar la curva de Phillips aumentada por las expectativas?

Bajo los supuestos ST.1′ a ST.5′, es posible mostrar que los estimadores de MCO son asintóticamente eficientes en la clase de los estimadores descrita en el teorema 5.3, pero se reemplaza el índice de observación de corte transversal i con el índice de series de tiempo t. Por último, los modelos con variables explicativas con tendencia satisfacen los supuestos ST.1′ a ST.5′, siempre y cuando sean estacionarios con tendencia. Mientras las tendencias en el tiempo se incluyan en las ecuaciones cuando sea necesario, los procedimientos de inferencia usuales son asintóticamente válidos.

11.3 Uso de series de tiempo altamente persistentes en el análisis de regresión

La sección anterior muestra que, siempre que las series de tiempo que se utilicen sean débilmente dependientes, los procedimientos usuales de inferencia de MCO serán válidos bajo supuestos más débiles que aquellos del modelo lineal clásico. Por desgracia, varias series de tiempo económicas no pueden caracterizarse por una dependencia débil. El uso de series de tiempo con dependencia fuerte en el análisis de regresión no plantea ningún problema, *si* los supuestos del MLC del capítulo 10 son válidos. Pero los procedimientos de inferencia usuales son muy sensibles a la violación de estos supuestos cuando los datos no son débilmente dependientes, debido a que no se puede recurrir a la ley de los grandes números ni al teorema del límite central. En esta sección se dan algunos ejemplos de series de tiempo **altamente persistentes** (o **fuertemente dependientes**) y se muestra cómo pueden transformarse para utilizarlas en el análisis de regresión.

Series de tiempo altamente persistentes

En el modelo AR(1) simple en (11.2), el supuesto $|\rho_1| < 1$ es crucial para que las series de tiempo sean débilmente dependientes. Pero resulta que muchas series de tiempo económicas se caracterizan mejor por el modelo AR(1) con $\rho_1 = 1$. En este caso se puede escribir

$$y_t = y_{t-1} + e_t, t = 1, 2, \dots, \tag{11.20}$$

donde de nuevo se supone que $\{e_t: t = 1, 2, \ldots\}$ es independiente e idénticamente distribuida con media cero y varianza σ_e^2. Asimismo, se da por hecho que el valor inicial, y_0, es independiente de e_t para toda $t \geq 1$.

El proceso de la ecuación (11.20) se llama **caminata aleatoria**. El nombre proviene del hecho que y en el periodo t se obtiene a partir del valor anterior, y_{t-1}, y al añadir una variable aleatoria de media cero que sea independiente de y_{t-1}. En ocasiones, una caminata aleatoria se define de manera distinta al suponer diferentes propiedades de las innovaciones, e_t (como la falta de correlación en vez de la independencia), pero la definición actual es suficiente para nuestros propósitos.

Primero, se calcula el valor esperado de y_t. Esto se realiza con mayor facilidad si se utiliza la sustitución repetida para obtener

$$y_t = e_t + e_{t-1} + \ldots + e_1 + y_0.$$

Al tomar el valor esperado de ambos lados se tiene

$$\begin{aligned}\mathrm{E}(y_t) &= \mathrm{E}(e_t) + \mathrm{E}(e_{t-1}) + \ldots + \mathrm{E}(e_1) + \mathrm{E}(y_0)\\ &= \mathrm{E}(y_0), \text{ para toda } t \geq 1.\end{aligned}$$

Por consiguiente, el valor esperado de una caminata aleatoria *no* depende de t. Es común suponer que $y_0 = 0$ —el proceso comienza en cero en el tiempo cero— en cuyo caso, $\mathrm{E}(y_t) = 0$ para toda t.

En comparación, la varianza de una caminata aleatoria sí cambia con t. Para calcular la varianza de una caminata aleatoria por simplicidad se supone que y_0 no es aleatoria de modo que $\mathrm{Var}(y_0) = 0$; esto no afecta ninguna conclusión importante. Por ende, según el supuesto i.i.d. para $\{e_t\}$,

$$\mathrm{Var}(y_t) = \mathrm{Var}(e_t) + \mathrm{Var}(e_{t-1}) + \ldots + \mathrm{Var}(e_1) = \sigma_e^2 t. \qquad \boxed{11.21}$$

En otras palabras, la varianza de una caminata aleatoria aumenta como una función lineal del tiempo, razón por la cual el proceso no puede ser estacionario.

Lo más importante es que una caminata aleatoria exhibe un comportamiento persistente en el sentido que el valor actual de y es importante para determinar el valor de y en un futuro muy lejano. Para entender esto, escriba la ecuación para h periodos hacia delante,

$$y_{t+h} = e_{t+h} + e_{t+h-1} + \ldots + e_{t+1} + y_t.$$

Ahora imagine que en el periodo t, se quiere calcular el valor esperado de y_{t+h} para el valor actual y_t dado. Como el valor esperado de e_{t+j}, para una y_t determinada, es cero para toda $j \geq 1$, se tiene

$$\mathrm{E}(y_{t+h}|y_t) = y_t, \text{ para toda } h \geq 1. \qquad \boxed{11.22}$$

Esto significa que, sin importar qué tan lejano se considere el futuro, nuestra mejor predicción de y_{t+h} es el valor actual, y_t. Se puede hacer una comparación con el caso del AR(1) estable, donde se usa un argumento parecido para mostrar que

$$\mathrm{E}(y_{t+h}|y_t) = \rho_1^h y_t, \text{ para toda } h \geq 1.$$

Bajo la estabilidad, $|\rho_1| < 1$, y por consiguiente $E(y_{t+h}|y_t)$ se aproxima a cero a medida que $h \rightarrow \infty$: el valor de y_t adquiere cada vez menos importancia, y $E(y_{t+h}|y_t)$ se acerca cada vez más al valor incondicional esperado, $E(y_t) = 0$.

Cuando $h = 1$, la ecuación (11.22) recuerda el supuesto de las expectativas adaptativas que se usó para la tasa de inflación en el ejemplo 11.5: si la inflación implica una caminata aleatoria, entonces el valor esperado de inf_t, dados los valores pasados de la inflación, simplemente es inf_{t-1}. Por tanto, un modelo de caminata aleatoria para la inflación justifica el uso de las expectativas adaptativas.

También se puede ver que la correlación entre y_t y y_{t+h} es cercana a 1 para valores grandes de t cuando $\{y_t\}$ implica una caminata aleatoria. Si $\text{Var}(y_0) = 0$, puede mostrarse que

$$\text{Corr}(y_t, y_{t+h}) = \sqrt{t/(t+h)}.$$

Así que la correlación depende del punto de partida, t (de modo que $\{y_t\}$ no es estacionaria en covarianza). Además, aun cuando para t fija la correlación tiende a cero conforme $h \rightarrow \infty$, no lo hace muy rápido. De hecho, cuanto más grande sea t, la correlación tiende a cero con más lentitud a medida que el valor de h aumenta. Si se elige que h sea un tanto grande, por ejemplo $h = 100$, siempre se puede escoger una t lo suficientemente grande para que la correlación entre y_t y y_{t+h} se acerque de manera arbitraria a uno. (Si $h = 100$ y se quiere que la correlación sea mayor que .95, entonces $t > 1{,}000$ logra hacerlo.) Por tanto, una caminata aleatoria no satisface el requisito de una secuencia no correlacionada asintóticamente.

La figura 11.1 traza dos realizaciones de una caminata aleatoria con un valor inicial de $y_0 = 0$ y $e_t \sim \text{Normal}(0,1)$. Por lo general no es fácil determinar si se trata de una caminata aleatoria con sólo observar un diagrama de una serie de tiempo. A continuación se estudiará un método informal para hacer la distinción entre secuencias débilmente y altamente dependientes; en el capítulo 18 se analizarán las pruebas estadísticas formales.

FIGURA 11.1

Dos realizaciones de la caminata aleatoria $y_t = y_{t-1} + e_t$, con $y_0 = 0$, $e_t \sim \text{Normal}(0,1)$ y $n = 50$.

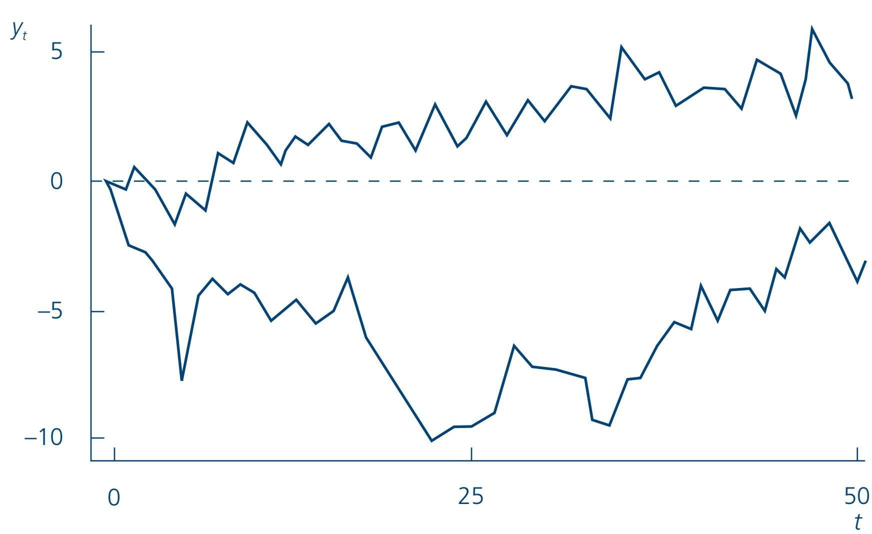
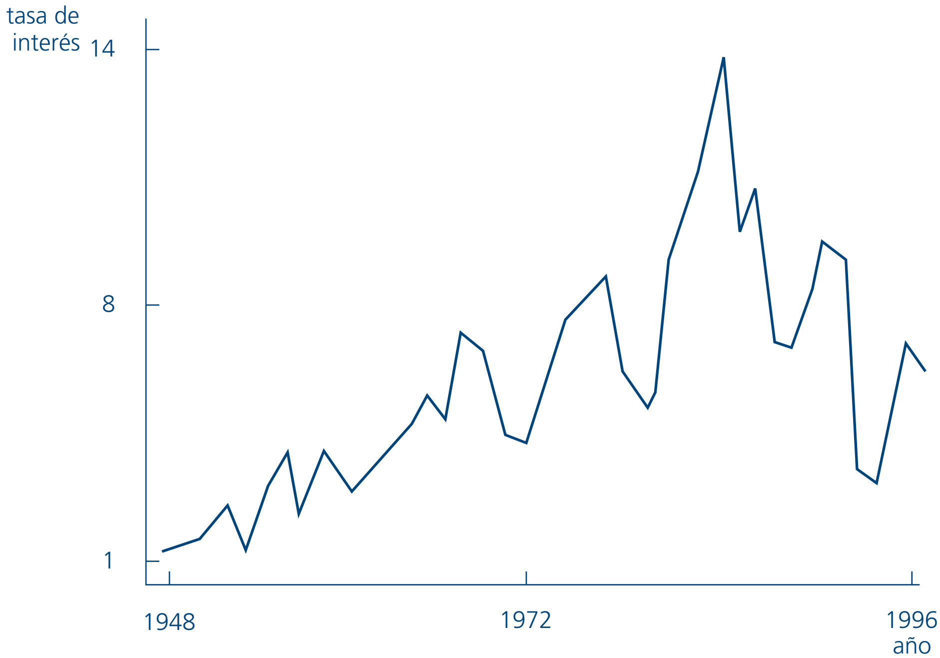


FIGURA 11.2

Tasa de los bonos del Tesoro a tres meses, para el periodo de 1948 a 1996.



Una serie que por lo general se considera bien descrita por una caminata aleatoria es la tasa de los bonos del Tesoro a tres meses. Los datos anuales se trazan en la figura 11.2 para los años de 1948 a 1996.

Una caminata aleatoria es un caso especial de lo que se conoce como **proceso de raíz unitaria**. El nombre se tomó del hecho de que $\rho_1 = 1$ en el modelo AR(1). Una clase más general de procesos de raíz unitaria se genera como en la ecuación (11.20), pero ahora se permite que $\{e_t\}$ sea una serie general débilmente dependiente. [Por ejemplo, $\{e_t\}$ podría seguir un proceso MA(1) o un proceso AR(1) estable.] Cuando $\{e_t\}$ no es una secuencia i.i.d., las propiedades de la caminata aleatoria que se obtuvieron antes ya no son válidas. Pero la característica fundamental de $\{y_t\}$ se mantiene: el valor actual de y está muy correlacionado con y incluso en el futuro distante.

Desde un punto de vista de políticas, a menudo es importante saber si una serie de tiempo económica es altamente persistente o no. Considere el caso del producto interno bruto en Estados Unidos. Si el PIB no tiene una correlación asintótica, entonces el nivel del PIB del próximo año, en el mejor de los casos, está débilmente relacionado con el nivel que el PIB tenía, digamos, hace 30 años. Esto significa que una política que afectaba al PIB hace tiempo tiene muy poco impacto perdurable. Por otro lado, si el PIB es fuertemente dependiente, el PIB del año siguiente podría estar muy correlacionado con el de hace muchos años. Así, se debe reconocer que una política que produce un cambio discreto en el PIB puede tener efectos duraderos.

Es de gran importancia no confundir los comportamientos de tendencia y persistencia elevada. Una serie puede mostrar tendencia pero no ser muy persistente, como se vio en el capítulo 10. Más aún, muchos piensan que factores tales como las tasas de interés, los índices de inflación y las tasas de desempleo son altamente persistentes, pero no tienen tendencias manifiestas hacia arriba o hacia abajo. Sin embargo, a menudo ocurre que una serie altamente persistente también muestra una clara tendencia. Un modelo que conduce a este comportamiento es la **caminata aleatoria con deriva**:

$$y_t = \alpha_0 + y_{t-1} + e_t,\ t = 1, 2, \dots, \qquad \textbf{11.23}$$

donde $\{e_t: t = 1, 2, \dots\}$ y y_0 satisfacen las mismas propiedades del modelo de caminata aleatoria. Lo nuevo es el parámetro α_0, que es el *término de deriva* o *tendencia estocástica (drift term)*. En esencia, para generar y_t, la constante α_0 se añade junto con el ruido aleatorio e_t al valor anterior y_{t-1}. Es posible demostrar, mediante sustitución repetida, que el valor esperado de y_t sigue una tendencia lineal en el tiempo:

$$y_t = \alpha_0 t + e_t + e_{t-1} + \dots + e_1 + y_0.$$

Por tanto, si $y_0 = 0$, entonces $E(y_t) = \alpha_0 t$: el valor esperado de y_t crece con el tiempo si $\alpha_0 > 0$ y decrece con el tiempo si $\alpha_0 < 0$. Siguiendo el razonamiento del caso de la caminata aleatoria pura, se puede demostrar que $E(y_{t+h}|y_t) = \alpha_0 h + y_t$, y por ende la mejor predicción de y_{t+h} en el periodo t es y_t más la deriva $\alpha_0 h$. La varianza de y_t es la misma que en el caso de la caminata aleatoria pura.

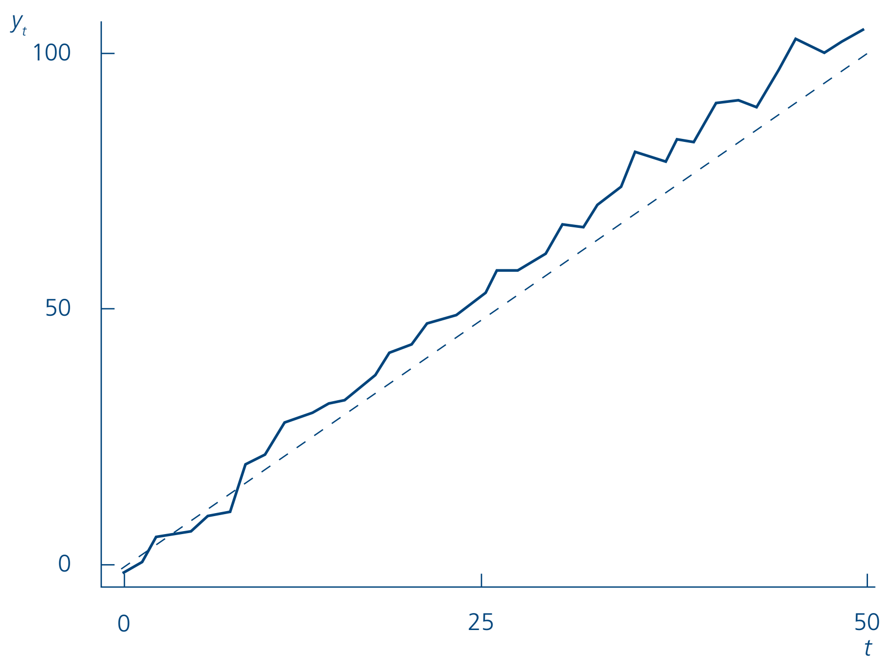
La figura 11.3 contiene una realización de una caminata aleatoria con deriva, donde $n = 50$, $y_0 = 0$, $\alpha_0 = 2$, y las e_t son variables aleatorias Normal(0,9). Como se aprecia en esta gráfica, y_t tiende a crecer con el tiempo, pero por lo común las series no regresan a la línea de tendencia.

Una caminata aleatoria con deriva es otro ejemplo de un proceso de raíz unitaria, ya que es el caso especial de $\rho_1 = 1$ en un modelo AR(1) con un intercepto:

$$y_t = \alpha_0 + \rho_1 y_{t-1} + e_t.$$

FIGURA 11.3

Una realización de la caminata aleatoria con deriva, $y_t = 2 + y_{t-1} + e_t$ con $y_0 = 0$, $e_t \sim$ Normal(0, 9) y $n = 50$. La línea punteada es el valor esperado de y_t, $E(y_t) = 2t$.



Cuando $\rho_1 = 1$ y $\{e_t\}$ es un proceso débilmente dependiente, se obtiene una clase completa de procesos de series altamente persistentes que también tienen medias con tendencia lineal.

Transformaciones de series de tiempo altamente persistentes

Cuando en una ecuación de regresión se usan series de tiempo altamente persistentes del tipo que exhibe un proceso de raíz unitaria, el resultado puede ser engañoso si se violan los supuestos del MCL. En el capítulo 18 se estudiará con más detalle el problema de la regresión espuria, pero por ahora se debe estar consciente de los problemas potenciales. Por fortuna, existen transformaciones simples que hacen que los procesos de raíz unitaria sean débilmente dependientes.

Se dice que los procesos débilmente dependientes son **integrados de orden cero** o **I(0)**. En un sentido práctico, esto significa que no es necesario hacer nada con esas series antes de utilizarlas en un análisis de regresión: los promedios de estas secuencias ya satisfacen los teoremas del límite más comunes. Se dice que los procesos de raíz unitaria, como la caminata aleatoria (con o sin deriva), son **integrados de orden uno**, o **I(1)**. Esto significa que la **primera diferencia** del proceso es débilmente dependiente (y a menudo estacionaria). Una serie de tiempo I(1) a menudo se considera un **proceso estacionario en diferencias**, aun cuando en cierto modo el nombre es engañoso respecto al énfasis que se pone en la estacionariedad después de la diferenciación, en vez de la dependencia débil en las diferencias.

Es más fácil apreciar el concepto de proceso I(1) en una caminata aleatoria. Con $\{y_t\}$ generado como en la ecuación (11.20) para $t = 1, 2, \ldots,$

$$\Delta y_t = y_t - y_{t-1} = e_t, \; t = 2, 3, \ldots; \tag{11.24}$$

por tanto, la serie en primeras diferencias $\{\Delta y_t: t = 2, 3, \ldots\}$ en realidad es una sucesión i.i.d. En un sentido más amplio, si $\{y_t\}$ es generada por la ecuación (11.24), donde $\{e_t\}$ es un proceso débilmente dependiente, entonces $\{\Delta y_t\}$ es débilmente dependiente. Por tanto, cuando se sospecha que los procesos son integrados de orden uno, a menudo se obtienen las primeras diferencias con el fin de utilizarlas en el análisis de regresión. Más adelante se verán algunos ejemplos.

Muchas series de tiempo y_t que son estrictamente positivas, son tales que $\log(y_t)$ es integrado de orden uno. En este caso, en el análisis de regresión se usa la primera diferencia de los logaritmos, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$. O bien, puesto que

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}, \tag{11.25}$$

se puede usar de manera directa el cambio proporcional o porcentual en y_t; eso fue lo que se hizo en el ejemplo 11.4, cuando en lugar de enunciar la hipótesis de los mercados eficientes en función del precio de las acciones, p_t, se usó el cambio porcentual semanal, $return_t = 100[(p_t - p_{t-1})/p_{t-1}]$.

La diferenciación de las series de tiempo antes de usarlas en el análisis de regresión conlleva otra ventaja: elimina cualquier tendencia lineal en el tiempo, como se aprecia con facilidad al escribir una variable con tendencia lineal en el tiempo como

$$y_t = \gamma_0 + \gamma_1 t + v_t,$$

en donde v_t tiene media cero. De ahí que, $\Delta y_t = \gamma_1 + \Delta v_t$, y por ende $E(\Delta y_t) = \gamma_1 + E(\Delta v_t) = \gamma_1$. En otras palabras, $E(\Delta y_t)$ es constante. El mismo argumento es válido para $\Delta\log(y_t)$ cuando $\log(y_t)$ sigue una tendencia lineal en el tiempo. Por consiguiente, en vez de incluir una tendencia en el tiempo en una regresión, se pueden diferenciar las variables que muestran tendencias claras.

Decidir si una serie de tiempo es o no I(1)

Determinar si una realización de una serie de tiempo en particular es el resultado de un proceso I(1) o I(0) puede ser muy difícil. Es por ello que se suelen utilizar las pruebas estadísticas, pero son un tema más avanzado; en el capitulo 18 se dará una introducción a las mismas.

Existen métodos informales que proporcionan lineamientos útiles para saber si un proceso de serie de tiempo se considera más o menos débilmente dependiente. Una herramienta muy simple es motivada por el modelo AR(1): si $|\rho_1| < 1$, el proceso es I(0), pero si $\rho_1 = 1$, entonces es I(1). Ya se mostró que cuando el proceso AR(1) es estable, $\rho_1 = \text{Corr}(y_t, y_{t-1})$. Por tanto, se puede estimar ρ_1 a partir de la correlación muestral entre y_t y y_{t-1}. Este coeficiente de correlación muestral se denomina **autocorrelación de primer orden** de $\{y_t\}$; lo cual se denota por $\hat{\rho}_1$. Al aplicar la ley de los grandes números, se demuestra que $\hat{\rho}_1$ es un estimador consistente de ρ_1 *siempre y cuando* $|\rho_1| < 1$. (No obstante, $\hat{\rho}_1$ no es un estimador insesgado de ρ_1.)

El valor de $\hat{\rho}_1$ se puede usar para decidir si el proceso es I(1) o I(0). Por desgracia, como $\hat{\rho}_1$ es una estimación, nunca se sabe con certeza si $\rho_1 < 1$. Lo ideal sería poder calcular un intervalo de confianza para ρ_1 con la finalidad de ver si excluye el valor $\rho_1 = 1$, pero resulta que esto es bastante difícil. Las distribuciones de muestreo del estimador $\hat{\rho}_1$ son muy diferentes cuando ρ_1 está cerca de uno que cuando ρ_1 es mucho menor que uno (de hecho, cuando ρ_1 está cerca de uno, $\hat{\rho}_1$ puede tener un sesgo hacia abajo muy grande).

En el capítulo 18 se explicará cómo probar H_0: $\rho_1 = 1$ contra H_0: $\rho_1 < 1$. Por ahora, sólo se empleará $\hat{\rho}_1$ como guía aproximada para determinar si es necesario diferenciar una serie. No existe una norma exacta para hacer esta elección. La mayoría de los economistas piensa que la diferenciación se justifica si $\hat{\rho}_1 > .9$; algunos diferenciarían cuando $\hat{\rho}_1 > .8$.

Ejemplo 11.6

[Ecuación de fertilidad]

En el ejemplo 10.4, se explicó la tasa de fertilidad general, *gfr*, en función del valor de la exención de impuestos personal, *pe*. Las autocorrelaciones de primer orden para estas series son muy grandes: $\hat{\rho}_1 = .977$ para *gfr* y $\hat{\rho}_1 = .964$ para *pe*. Estas autocorrelaciones sugieren un comportamiento de raíz unitaria, además de suscitar preguntas sobre el uso del estadístico *t* usual de MCO que se usó en ese ejemplo. Recuerde que los estadísticos *t* sólo tienen distribuciones *t* exactas bajo todo el conjunto de supuestos del modelo lineal clásico. Para de alguna manera relajar esos supuestos y aplicar los asintóticos, por lo general se necesita que las series subyacentes sean procesos I(0).

Ahora se estima la ecuación en primeras diferencias (y por simplicidad se omite la variable binaria o variable dummy):

$$
\begin{aligned}
\Delta\widehat{gfr} &= -.785 - .043\ \Delta pe \\
&\quad\ \ (.502)\quad\ (.028) \\
n &= 71,\ R^2 = .032,\ \bar{R}^2 = .018.
\end{aligned}
\tag{11.26}
$$

Ahora, se estima que un incremento en *pe* disminuye *gfr* contemporáneamente, aun cuando en términos estadísticos la estimación no sea diferente de cero al nivel de 5%. Esto da resultados muy diferentes de aquéllos cuando se estimó el modelo en niveles y origina dudas sobre el análisis anterior.

Si se añaden dos rezagos de Δpe las cosas mejoran:

$$\begin{aligned}\Delta\widehat{gfr} &= -.964 - .036\,\Delta pe - .014\,\Delta pe_{-1} + .110\,\Delta pe_{-2}\\ &\quad\;(.468)\;\;(.027)\qquad\;(.028)\qquad\;\;(.027)\\ n &= 69,\ R^2 = .233,\ \bar{R}^2 = .197.\end{aligned} \tag{11.27}$$

Aun cuando Δpe y Δpe_{-1} tienen coeficientes negativos, son pequeños y conjuntamente insignificantes (valor-p = .28). El segundo rezago es muy significativo e indica una relación positiva entre los cambios en *pe* y los cambios posteriores en *gfr* hasta dos años después. Esto tiene más sentido que obtener un efecto contemporáneo. Vea el ejercicio para computadora C11.5 para un análisis más a fondo de la ecuación en sus primeras diferencias.

Cuando las series en cuestión tienen una tendencia marcada hacia arriba o hacia abajo, es mejor obtener la autocorrelación de primer orden después de eliminar tal tendencia. Si la tendencia de los datos no se elimina, la correlación autorregresiva tiende a sobrevalorarse, lo cual introduce un sesgo hacia encontrar una raíz unitaria en un proceso con tendencia.

Ejemplo 11.7

[Salarios y productividad]

La variable *hrwage* es el salario promedio por hora en la economía estadounidense y *outphr* es la producción por hora. Una manera de estimar la elasticidad del salario por hora respecto a la producción por hora es estimar la ecuación,

$$\log(hrwage_t) = \beta_0 + \beta_1\log(outphr_t) + \beta_2 t + u_t,$$

donde se incluye la tendencia en el tiempo porque $\log(hrwage_t)$ y $\log(outphr_t)$ muestran tendencias lineales ascendentes claras. Usando los datos de EARNS.RAW para los años 1947 a 1987, se obtiene

$$\begin{aligned}\widehat{\log(hrwage_t)} &= -5.33 + 1.64\log(outphr_t) - .018\,t\\ &\qquad\;(.37)\quad\;(.09)\qquad\qquad\qquad\;(.002)\\ n &= 41,\ R^2 = .971,\ \bar{R}^2 = .970.\end{aligned} \tag{11.28}$$

(Aquí se dan las medidas de bondad de ajuste usuales; pero sería mejor informar aquellas que se basan en la variable dependiente cuya tendencia se eliminó, como en la sección 10.5.) La elasticidad estimada parece demasiado grande: un incremento de 1% en la productividad genera un aumento de alrededor de 1.64% en los salarios reales. Dado que el error estándar es muy pequeño, el intervalo de confianza de 95% excluye con facilidad una elasticidad unitaria. Los trabajadores estadounidenses tal vez no creerían que su salario aumenta 1.5% por cada incremento de 1% en la productividad.

Los resultados de la regresión en la ecuación (11.28) deben verse con cautela. Incluso después de eliminar la tendencia lineal de log(*hrwage*), la autocorrelación de primer orden es .967 y para log(*outphr*), con tendencia eliminada $\hat{\rho}_1$ = .945. Esto sugiere que las dos series tienen raíces unitarias, así que se estima de nuevo la ecuación en primeras diferencias (y ya no es necesaria una tendencia en el tiempo):

$$\begin{aligned}\Delta\widehat{\log(hrwage_t)} &= -.0036 + .809\,\Delta\log(outphr)\\ &\qquad\;(.0042)\quad(.173)\\ n &= 40,\ R^2 = .364,\ \bar{R}^2 = .348.\end{aligned} \tag{11.29}$$

Ahora, se estima que un incremento de 1% en la productividad aumenta los salarios reales en alrededor de .81%, y la estimación no es estadísticamente diferente de uno. La R-cuadrada ajustada muestra que el crecimiento en la producción explica cerca de 35% del aumento en los salarios reales. Vea el ejercicio para computadora C11.2 para una versión sencilla de rezagos distribuidos del modelo en primeras diferencias.

En los dos ejemplos anteriores, las variables dependientes e independientes parecen tener raíces unitarias. En otros casos, se podría tener una combinación de procesos con raíces unitarias y procesos débilmente dependientes (aunque posiblemente con tendencia). En el ejercicio para computadora C11.1 se proporciona un ejemplo.

11.4 Modelos dinámicamente completos y ausencia de correlación serial

En el modelo AR(1) de la ecuación (11.12), se mostró que bajo el supuesto (11.13) los errores $\{u_t\}$ *deben* ser **no correlacionados serialmente ya que el supuesto ST**.5′ se satisface: suponer que no existe una correlación serial es casi lo mismo que suponer que sólo un rezago de y aparece en $E(y_t|y_{t-1}, y_{t-2}, \ldots)$.

¿Un enunciado parecido se aplica a otros modelos de regresión? La respuesta es afirmativa. Considere el modelo simple de regresión estática

$$y_t = \beta_0 + \beta_1 z_t + u_t, \tag{11.30}$$

donde y_t y z_t tienen la misma fecha. Por consistencia de MCO, sólo se necesita que $E(u_t|z_t) = 0$. Por lo general, las $\{u_t\}$ estarán serialmente correlacionadas. Sin embargo, si se *supone* que

$$E(u_t|z_t, y_{t-1}, z_{t-1}, \ldots) = 0, \tag{11.31}$$

entonces (como se mostrará de manera general más adelante) el supuesto ST.5′ es válido. En particular, las $\{u_t\}$ no están correlacionadas serialmente. De manera natural, suponer (11.31) implica que z_t es contemporáneamente exógena, es decir, $E(u_t|z_t) = 0$.

Para entender el significado de la ecuación (11.31), se pueden escribir las ecuaciones (11.30) y (11.31) de manera equivalente como

$$E(y_t|z_t, y_{t-1}, z_{t-1}, \ldots) = E(y_t|z_t) = \beta_0 + \beta_1 z_t, \tag{11.32}$$

donde la primera igualdad es la que nos interesa ahora. Indica que, una vez que z_t se toma en consideración, ningún rezago de y ni de z explica la y actual. Es un requisito fuerte; si no se cumple, entonces es de esperar que los errores estén serialmente correlacionados.

Ahora considere un modelo de rezagos distribuidos finitos con dos rezagos:

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 z_{t-1} + \beta_3 z_{t-2} + u_t. \tag{11.33}$$

Como se espera captar los efectos rezagados que z tiene sobre y, sería natural suponer que la ecuación (11.33) reproduce la *dinámica de rezagos distribuidos*:

$$E(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = E(y_t|z_t, z_{t-1}, z_{t-2}); \tag{11.34}$$

es decir, a lo sumo importan dos rezagos de z. Si la ecuación (11.31) es válida, es posible hacer otros enunciados: una vez que z y sus dos rezagos se toman en consideración, ningún rezago de y o rezagos adicionales de z afectan la y actual:

$$\mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, \ldots) = \mathrm{E}(y_t|z_t, z_{t-1}, z_{t-2}). \tag{11.35}$$

La ecuación (11.35) es más viable que la ecuación (11.32), pero aún descarta que la y rezagada influya sobre la y actual.

En seguida considere un modelo con un rezago tanto de y como de z:

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1} + u_t.$$

Puesto que este modelo incluye una variable dependiente con rezago, la ecuación (11.31) es una suposición natural en tanto implica que

$$\mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) = \mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1});$$

en otras palabras, una vez que z_t, y_{t-1} y z_{t-1} han sido controladas, rezagos adicionales de y o de z no afectan a la y actual.

En el modelo general

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \tag{11.36}$$

donde las variables explicativas $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ pueden o no contener rezagos de y o z, la ecuación (11.31) se convierte en

$$\mathrm{E}(u_t|\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots) = 0. \tag{11.37}$$

Expresada en función de y_t,

$$\mathrm{E}(y_t|\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots) = \mathrm{E}(y_t|\mathbf{x}_t). \tag{11.38}$$

En otras palabras, sin importar las variables que forman parte de $\mathbf{x}_t$, se han incluido suficientes rezagos de y para que rezagos adicionales de y y de las variables explicativas no tengan importancia en la explicación de y_t. Cuando esta condición se cumple, se tiene un **modelo dinámicamente completo**. Como se vio antes, la completitud dinámica llega a ser un supuesto muy fuerte en los modelos estáticos y de rezagos distribuidos finitos.

En cuanto se incluyen y rezagadas como variables explicativas, se suele pensar que el modelo debe ser dinámicamente completo. En el capítulo 18 se tratarán algunas excepciones a esta afirmación.

Puesto que la ecuación (11.37) es equivalente a

$$\mathrm{E}(u_t|\mathbf{x}_t, u_{t-1}, \mathbf{x}_{t-1}, u_{t-2}, \ldots) = 0, \tag{11.39}$$

se puede mostrar que un modelo dinámicamente completo *debe* satisfacer el supuesto ST.5′. (Esta demostración no es crucial y puede omitirse sin perder continuidad.) Para concretar, tome $s < t$. Entonces, por la ley de las esperanzas iteradas (vea el apéndice B),

$$\begin{aligned}\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s) &= \mathrm{E}[\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s, u_s)|\mathbf{x}_t, \mathbf{x}_s]\\ &= \mathrm{E}[u_s \mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s)|\mathbf{x}_t, \mathbf{x}_s],\end{aligned}$$

en donde la segunda igualdad se deduce de $\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s, u_s) = u_s\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s)$. Ahora, puesto que $s < t$, $(\mathbf{x}_t, \mathbf{x}_s, u_s)$ es un subconjunto del conjunto de condicionamiento de la ecuación (11.39). Por tanto, (11.39) implica que $\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s) = 0$, y así

$$\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s) = \mathrm{E}(u_s \cdot 0|\mathbf{x}_t, \mathbf{x}_s) = 0,$$

lo que indica que el supuesto ST.5′ se mantiene.

A partir del hecho de que especificar un modelo dinámicamente completo significa que no existe una correlación serial, ¿se puede deducir que todos los modelos deben ser dinámicamente completos? Como se verá en el capítulo 18, con fines de pronóstico, la respuesta es sí. Algunos piensan que todos los modelos deben ser dinámicamente completos y que la correlación serial en los errores de un modelo es un signo de especificación incorrecta, pero esta postura es demasiado rígida. A veces nos interesa en realidad un modelo estático (como una curva de Phillips) o uno de rezagos distribuidos finitos (como la medición del cambio porcentual a largo plazo de los salarios, dado un incremento de un punto porcentual en la productividad). En el siguiente capítulo se mostrará cómo detectar y corregir la correlación serial de estos modelos.

Pregunta 11.3

Si la ecuación (11.33) es válida con $u_t = e_t + \alpha_1 e_{t-1}$ donde $\{e_t\}$ es una secuencia i.i.d. con media cero y varianza σ_e^2, ¿la ecuación (11.33) puede ser dinámicamente completa?

Ejemplo 11.8

[Ecuación de fertilidad]

En la ecuación (11.27) se estimó un modelo de rezagos distribuidos de Δgfr sobre Δpe, y se permitieron dos rezagos de esta última. Para que este modelo sea dinámicamente completo en el sentido de la ecuación (11.38), no deben aparecer rezagos de Δgfr ni nuevos rezagos de Δpe en la ecuación. Es fácil darse cuenta de que esto es falso si se añade Δgfr_{-1}: el coeficiente estimado es .300 y su estadístico t es de 2.84. Así, el modelo no es dinámicamente completo en el sentido de la ecuación (11.38).

¿Qué se hace en este caso? Se pospondrá hasta el capítulo 18 la interpretación de los modelos generales con variables dependientes rezagadas. Pero el hecho de que la ecuación (11.27) no sea dinámicamente completa indica que quizás exista una correlación serial en los errores. En el capítulo 12 se verá cómo probarlo y corregirlo.

La noción de completitud dinámica no debe confundirse con un supuesto más débil que se refiera a la inclusión de los rezagos apropiados en un modelo. En el modelo (11.36), se dice que las variables explicativas $\mathbf{x}_t$ son **secuencialmente exógenas** si

$$\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = \mathrm{E}(u_t) = 0, \ t = 1, 2, \ldots. \tag{11.40}$$

Como se vio en el problema 10.8, la exogeneidad secuencial está implícita en la exogeneidad estricta, y la exogeneidad secuencial implica exogeneidad contemporánea. Además, como $(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots)$ es un subconjunto de $(\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots)$, la exogeneidad secuencial está implícita en la completitud dinámica. Si $\mathbf{x}_t$ contiene a y_{t-1}, la completitud dinámica y la exogeneidad secuencial son la misma condición. El punto clave es que, cuando $\mathbf{x}_t$ no contiene a y_{t-1}, la exogeneidad secuencial permite la posibilidad de que la dinámica no esté completa en el sentido de captar la relación entre y_t y todos los valores pasados de y y de otras variables explicativas. Pero en los modelos de rezagos distribuidos finitos, como el estimado en la ecuación (11.27), tal vez no importe si la

y pasada tiene la capacidad de predecir la y actual. El interés principal es si se han incluido suficientes rezagos de las variables explicativas para captar la dinámica de los rezagos distribuidos. Por ejemplo, si se supone que $E(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = E(y_t|z_t, z_{t-1}, z_{t-2}) = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2}$, entonces los regresores $\mathbf{x}_t = (z_t, z_{t-1}, z_{t-2})$ son secuencialmente exógenos debido a que se ha supuesto que dos rezagos son suficientes para la dinámica de rezagos distribuidos. Pero por lo general el modelo no sería completo dinámicamente en el sentido que $E(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, z_{t-2}, \ldots) = E(y_t|z_t, z_{t-1}, z_{t-2})$, y tal vez esto no nos interese. Además, las variables explicativas en el modelo RDF pueden o no ser estrictamente exógenas.

11.5 El supuesto de homocedasticidad en los modelos de series de tiempo

El supuesto de homocedasticidad en las regresiones de series de tiempo, en particular ST.4′, es muy parecido al supuesto de las regresiones de corte transversal. Sin embargo, puesto que $\mathbf{x}_t$ puede contener y rezagadas así como variables explicativas rezagadas, se expondrá en forma breve el significado del supuesto de la homocedasticidad para diferentes regresiones de series de tiempo.

En el modelo estático simple,

$$y_t = \beta_0 + \beta_1 z_t + u_t, \tag{11.41}$$

El supuesto ST.4′ exige que

$$\mathrm{Var}(u_t|z_t) = \sigma^2.$$

Por tanto, aunque $E(y_t|z_t)$ sea una función lineal de z_t, $\mathrm{Var}(y_t|z_t)$ debe ser constante. Esto es muy sencillo.

En el ejemplo 11.4 se vio que, para el modelo AR(1) de la ecuación (11.12), el supuesto de homocedasticidad es

$$\mathrm{Var}(u_t|y_{t-1}) = \mathrm{Var}(y_t|y_{t-1}) = \sigma^2;$$

aunque $E(y_t|y_{t-1})$ dependa de y_{t-1}, $\mathrm{Var}(y_t|y_{t-1})$ no. Así, la variación en la distribución de y_t no puede depender de y_{t-1}.

Con suerte, el esquema es evidente ahora. Si se tiene el modelo

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1} + u_t,$$

el supuesto de homocedasticidad es

$$\mathrm{Var}(u_t|z_t, y_{t-1}, z_{t-1}) = \mathrm{Var}(y_t|z_t, y_{t-1}, z_{t-1}) = \sigma^2,$$

de modo que la varianza de u_t no puede depender de z_t, y_{t-1} o z_{t-1} (ni de ninguna otra función del tiempo). En general, cualquiera que sea la variable explicativa que aparezca en el modelo, se debe suponer que la varianza de y_t, dadas estas variables, es constante. Si el modelo contiene y rezagadas o variables explicativas rezagadas, entonces se descartan explícitamente las formas dinámicas de heterocedasticidad (tema que se estudiará en el capítulo 12). Pero en un modelo estático, sólo es de interés $\mathrm{Var}(y_t|z_t)$. En la ecuación (11.41) no se imponen restricciones directas sobre, por ejemplo, $\mathrm{Var}(y_t|y_{t-1})$.

RESUMEN

En este capítulo, se argumentó que los MCO se justifican mediante el análisis asintótico, siempre que se cumplan ciertas condiciones. En términos idóneos, los procesos de series de tiempo son estacionarios y débilmente dependientes, aunque la estacionariedad no es crucial. La dependencia débil es necesaria para aplicar los habituales resultados de muestras grandes, en particular el teorema del límite central.

Los procesos con tendencias determinísticas que son débilmente dependientes se usan de manera directa en el análisis de regresión, siempre que se incluyan tendencias temporales en el modelo (como en la sección 10.5). La misma afirmación es válida para los procesos con estacionariedad.

Cuando las series de tiempo son altamente persistentes (tienen raíces unitarias), se debe tener gran cautela al emplearlas de manera directa en los modelos de regresión (salvo que se esté seguro de que se aplican los supuestos del MLC del capítulo 10). Una alternativa al uso de niveles en las variables es la utilización de las primeras diferencias. En la mayoría de las series de tiempo económicas altamente persistentes la primera diferencia las hace débilmente dependientes. El uso de primeras diferencias cambia la naturaleza del modelo, pero este método suele proporcionar tanta información como el modelo en niveles. Cuando los datos son altamente persistentes, se debe tener más confianza en los resultados de la primera diferencia. En el capítulo 18 se cubrirán algunos métodos recientes y más avanzados para utilizar variables I(1) en el análisis de regresión múltiple.

Se vio que cuando los modelos son dinámicamente completos en el sentido de que en la ecuación no se requieren más rezagos de las variables, los errores no estarán correlacionados serialmente. Esto es útil porque se supone que ciertos modelos, como los autorregresivos, son dinámicamente completos. En los modelos estáticos y de rezagos distribuidos, el supuesto de completitud dinámica suele ser falso, lo que por lo regular significa que los errores están serialmente correlacionados. En el capítulo 12 se verá cómo resolver este problema.

Los supuestos "asintóticos" de Gauss-Markov para la regresión de series de tiempo

A continuación se presenta un resumen de cinco supuestos que se usaron en este capítulo para realizar una inferencia en muestras grandes para las regresiones de series de tiempo. Recuerde que este nuevo conjunto de supuestos se introdujo porque a menudo se violan las versiones de series de tiempo de los supuestos del modelo lineal clásico, en particular los supuestos de exogeneidad estricta, no correlación serial y normalidad. Un punto clave de este capítulo es que cierto tipo de dependencia débil se requiere para asegurar que el teorema del límite central se aplique. Sólo se usaron los supuestos ST.1′ a ST.3′ para la consistencia (insesgamiento) de MCO. Cuando se añaden los supuestos ST.4′ y ST.5′, es posible usar los intervalos de confianza usuales, los estadísticos t y los estadísticos F como si fueran aproximadamente válidos en las muestras grandes. A diferencia de los supuestos de Gauss-Markov y del modelo lineal clásico, históricamente no se ha asignado un nombre significativo a los supuestos ST.1′ a ST.5′. Sin embargo, estos supuestos son análogos a los de Gauss-Markov que nos permiten usar la inferencia estándar. Como suele ocurrir con el análisis de muestras grandes, se prescinde por completo del supuesto de normalidad.

Supuesto ST.1′ (Linealidad y dependencia débil)

El proceso estocástico $\{(x_{t1}, x_{t2}, \dots, x_{tk}, y_t): t = 1, 2, \dots, n\}$ sigue el modelo lineal

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t,$$

donde $\{u_t: t = 1, 2, \dots, n\}$ es la secuencia de errores o perturbaciones. Aquí, n es el número de observaciones (periodos).

Supuesto ST.2′ (No hay colinealidad perfecta)

En la muestra (y por tanto en los procesos de series de tiempo subyacentes), ninguna variable dependiente es constante ni es una combinación perfecta de las otras.

Supuesto ST.3′ (Media condicional cero)

Las variables explicativas son *contemporáneamente exógenas,* es decir, $\mathrm{E}(u_t|x_{t1}, \ldots, x_{tk}) = 0$. Recuerde, ST.3′ es notablemente más débil que el supuesto de exogeneidad estricta ST.3′.

Supuesto ST.4′ (Homocedasticidad)

Los errores son *contemporáneamente homocedásticos,* es decir, $\mathrm{Var}(u_t|\mathbf{x}_t) = \sigma^2$, donde $\mathbf{x}_t$ es la abreviatura de $(x_{t1}, x_{t2}, \ldots, x_{tk})$.

Supuesto ST.5 (No hay correlación serial)

Para toda $t \neq s$, $\mathrm{E}(u_t, u_s|\mathbf{x}_t, \mathbf{x}_s) = 0$.

TÉRMINOS CLAVE

Altamente persistente
Autocorrelación de primer orden
Caminata aleatoria
Caminata aleatoria con deriva
Contemporáneamente exógeno
Contemporáneamente homocedástico
Débilmente dependiente
Estacionario en covarianza
Fuertemente dependiente
Integrado de orden cero [I(0)]
Integrado de orden uno [I(1)]
Modelo dinámicamente completo
No correlacionado serialmente
No correlacionados asintóticamente
Primera diferencia
Proceso AR(1) estable
Proceso autorregresivo de orden uno [AR(1)]
Proceso de promedio móvil de orden uno [MA(1)]
Proceso de raíz unitaria
Proceso estacionario
Proceso estacionario con tendencia
Proceso estacionario en diferencias
Proceso no estacionario
Secuencialmente exógeno

PROBLEMAS

11.1 Sea $\{x_t: t = 1, 2, \ldots\}$ un proceso estacionario en covarianza y defínase $\gamma_h = \mathrm{Cov}(x_t, x_{t+h})$ para $h \geq 0$. [Por consiguiente, $\gamma_0 = \mathrm{Var}(x_t)$.] Demuestre que $\mathrm{Corr}(x_t, x_{t+h}) = \gamma_h/\gamma_0$.

11.2 Sea $\{e_t: t = -1, 0, 1, \ldots\}$ una secuencia de variables aleatorias independientes, idénticamente distribuidas, con media cero y varianza uno. Defina un proceso estocástico mediante

$$x_t = e_t - (1/2)e_{t-1} + (1/2)e_{t-2},\ t = 1, 2, \ldots.$$

i) Encuentre $\mathrm{E}(x_t)$ y $\mathrm{Var}(x_t)$. ¿Alguna de éstas depende de t?
ii) Demuestre que $\mathrm{Corr}(x_t, x_{t+1}) = -1/2$ y que $\mathrm{Corr}(x_t, x_{t+2}) = 1/3$. (*Sugerencia:* es más fácil usar la fórmula del problema 11.1.)
iii) Calcule $\mathrm{Corr}(x_t, x_{t+h})$ para $h > 2$.
iv) ¿Es $\{x_t\}$ un proceso no correlacionado asintóticamente?

11.3 Suponga que un proceso de series de tiempo $\{y_t\}$ es generado por $y_t = z + e_t$, para toda $t = 1, 2, \ldots,$ en el que $\{e_t\}$ es una secuencia i.i.d. con media cero y varianza σ_e^2. La variable aleatoria

z no cambia con el tiempo; tiene media cero y varianza σ_z^2. Suponga que ninguna e_t guarda una correlación serial con z.

i) Encuentre el valor esperado y la varianza de y_t. ¿Dependen sus respuestas de t?

ii) Encuentre la $\text{Cov}(y_t, y_{t+h})$ para cualquier t y h. ¿Es $\{y_t\}$ estacionaria en covarianza?

iii) Con los incisos i) y ii) demuestre que $\text{Corr}(y_t, y_{t+h}) = \sigma_z^2/(\sigma_z^2 + \sigma_e^2)$ para toda t y h.

iv) ¿Satisface y_t el requisito intuitivo de no estar correlacionada asintóticamente? Explique por qué.

11.4 Sea $\{y_t: t = 1, 2, \ldots\}$ una caminata aleatoria, como la de la ecuación (11.20), con $y_0 = 0$. Muestre que $\text{Corr}(y_t, y_{t+h}) = \sqrt{t/(t+h)}$ para $t \geq 1$, $h > 0$.

11.5 En la economía estadounidense, sea *gprice* el crecimiento mensual del índice general de precios y *gwage* el incremento mensual del salario por hora. [Estas dos variables se obtienen como diferencias de logaritmos: $gprice = \Delta\log(price)$ y $gwage = \Delta\log(wage)$.] Con los datos mensuales en WAGEPRC.RAW, se estima el siguiente modelo de rezagos distribuidos:

$$
\begin{aligned}
\widehat{gprice} = \underset{(.00057)}{-.00093} &+ \underset{(.052)}{.119}\, gwage + \underset{(.039)}{.097}\, gwage_{-1} + \underset{(.039)}{.040}\, gwage_{-2} \\
&+ \underset{(.039)}{.038}\, gwage_{-3} + \underset{(.039)}{.081}\, gwage_{-4} + \underset{(.039)}{.107}\, gwage_{-5} + \underset{(.039)}{.095}\, gwage_{-6} \\
&+ \underset{(.039)}{.104}\, gwage_{-7} + \underset{(.039)}{.103}\, gwage_{-8} + \underset{(.039)}{.159}\, gwage_{-9} + \underset{(.039)}{.110}\, gwage_{-10} \\
&+ \underset{(.039)}{.103}\, gwage_{-11} + \underset{(.052)}{.016}\, gwage_{-12} \\
n = 273,\ R^2 &= .317,\ \bar{R}^2 = .283.
\end{aligned}
$$

i) Esboce la distribución de rezagos estimados. ¿En qué rezago es mayor el efecto de *gwage* sobre *gprice*. ¿Qué rezago tiene el menor coeficiente?

ii) ¿Para qué rezagos los estadísticos t son menores que dos?

iii) ¿Cuál es la propensión de largo plazo (PLP) estimada? ¿Difiere mucho de uno? Explique lo que significa la PLP en este ejemplo.

iv) ¿Qué regresión efectuaría para obtener de manera directa el error estándar de la PLP?

v) ¿Cómo probaría la significancia conjunta de seis nuevos rezagos de *gwage*? ¿Cuáles serían los grados de libertad de la distribución F? (Tenga cuidado; aquí se pierden otras seis observaciones.)

11.6 Sea $hy6_t$ el rendimiento trimestral (en porcentaje) de comprar un bono del Tesoro a seis meses en el periodo $(t-1)$ y venderlo en el periodo t (es decir, tres meses después, debido a que se tienen datos trimestrales) como un bono de tres meses. Sea $hy3_{t-1}$ el rendimiento trimestral de comprar un bono del Tesoro a tres meses en el periodo $(t-1)$. En dicho periodo, se conoce $hy3_{t-1}$ mientras que $hy6_t$ no, porque se desconoce $p3_t$ (el precio de los bonos a tres meses) en el periodo $(t-1)$. La *hipótesis de las expectativas* (HE) afirma que estas dos inversiones a tres meses deberían ser, en promedio, iguales. En términos matemáticos, lo anterior se escribe como una esperanza condicional:

$$\text{E}(hy6_t | I_{t-1}) = hy3_{t-1},$$

donde I_{t-1} denota toda la información observada hasta el periodo $t-1$. Esto sugiere la estimación del modelo

$$hy6_t = \beta_0 + \beta_1 hy3_{t-1} + u_t,$$

y realizar la prueba de H_0: $\beta_1 = 1$. (También se podría probar H_0: $\beta_0 = 0$, pero en general se permite una *prima por plazo* por la compra de valores con diferentes vencimientos, de modo que $\beta_0 \neq 0$.)

i) La estimación de la ecuación anterior por MCO con los datos trimestrales de INTQRT.RAW (en realidad, tiene datos espaciados cada tres meses) da

$$\begin{aligned}\widehat{hy6_t} &= -.058 + 1.104\, hy3_{t-1}\\ &\quad\;\; (.070) \quad\; (.039)\\ n &= 123,\ R^2 = .866.\end{aligned}$$

¿Rechaza usted H_0: $\beta_1 = 1$ contra H_0: $\beta_1 \neq 1$ al nivel de significancia del 1%? ¿El estimador parece en términos prácticos distinto de uno?

ii) Otra implicación de la HE es que ninguna otra variable fechada en $t - 1$ o antes, explicaría $hy6_t$, una vez que $hy3_{t-1}$ se haya considerado. Al incluir un rezago de la *diferencia* entre las tasas de los bonos a seis y tres meses se tiene

$$\begin{aligned}\widehat{hy6_t} &= -.123 + 1.053\, hy3_{t-1} + .480\,(r6_{t-1} - r3_{t-1})\\ &\quad\;\; (.067) \quad\; (.039) \qquad\qquad (.109)\\ n &= 123,\ R^2 = .885.\end{aligned}$$

¿Ahora el coeficiente de $hy3_{t-1}$ es estadísticamente diferente de uno? ¿Es significativo el coeficiente de la diferencia de tasas? De acuerdo con esta ecuación, si en el periodo $t - 1$, $r6$ es superior a $r3$, ¿invertiría usted en bonos a seis o a tres meses?

iii) La correlación muestral entre $hy3_t$ y $hy3_{t-1}$ es .914. ¿Por qué esto despertaría inquietudes sobre al análisis anterior?

iv) ¿Cómo probaría la estacionalidad en la ecuación estimada en el inciso ii)?

11.7 Un *modelo de ajuste parcial* es

$$y_t^* = \gamma_0 + \gamma_1 x_t + e_t$$

$$y_t - y_{t-1} = \lambda(y_t^* - y_{t-1}) + a_t,$$

donde y_t^* es el nivel óptimo o deseado de y, mientras que y_t es el nivel real (el observado). Por ejemplo, y_t^* es el crecimiento deseado de los inventarios de la empresa y x_t es el crecimiento de las ventas. El parámetro γ_1 mide el efecto de x_t sobre y_t^*. La segunda ecuación describe cómo se ajusta la y real dependiendo de la relación entre la y deseada en el periodo t y la y observada en el periodo $t - 1$. El parámetro λ mide la velocidad del ajuste y satisface $0 < \lambda < 1$.

i) Sustituya la primera ecuación de y_t^* en la segunda ecuación y muestre que se puede escribir

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t.$$

En particular, calcule las β_j en función de las γ_j y λ y calcule u_t en términos de e_t y a_t. Por consiguiente, el modelo de ajuste parcial lleva a un modelo con una variable dependiente rezagada y una x contemporánea.

ii) Si $E(e_t|x_t, y_{t-1}, x_{t-1}, \ldots) = E(a_t|x_t, y_{t-1}, x_{t-1}, \ldots) = 0$ y todas las series son débilmente dependientes, ¿cómo estimaría las β_j?

iii) Si $\hat{\beta}_1 = .7$ y $\hat{\beta}_2 = .2$, ¿cuáles son las estimaciones de γ_1 y λ?

11.8 Suponga que la ecuación

$$y_t = \alpha + \delta t + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$$

satisface el supuesto de exogeneidad secuencial de la ecuación (11.40).

i) Imagine que diferencia la ecuación para obtener

$$\Delta y_t = \delta + \beta_1 \Delta x_{t1} + \ldots + \beta_k \Delta x_{tk} + \Delta u_t.$$

¿Por qué la aplicación de MCO en la ecuación diferenciada por lo general no da como resultado estimadores consistentes de las β_j?

ii) ¿Qué supuesto sobre las variables explicativas en la ecuación original aseguraría que aplicar MCO a la ecuación en diferencias estime consistentemente las β_j?

iii) Sea $z_{t1}, \ldots, z_{tk}$ un conjunto de variables explicativas contemporáneas a y_t. Si se especifica el modelo de regresión estático $y_t = \beta_0 + \beta_1 z_{t1} + \ldots + \beta_k z_{tk} + u_t$, describa qué se necesita suponer para que $\mathbf{x}_t = \mathbf{z}_t$ sean secuencialmente exógenas. ¿Considera probable que se mantengan estos supuestos en las aplicaciones económicas?

EJERCICIOS EN COMPUTADORA

C11.1 Use la base de datos HSEINV.RAW para este ejercicio.

i) Calcule la autocorrelación de primer orden de log(*invpc*). Ahora, calcule la autocorrelación *después* de eliminar la tendencia lineal de log(*invpc*). Realice el mismo procedimiento para log(*price*). ¿Cuál de las dos series podría tener raíz unitaria?

ii) A partir de los resultados del inciso i), estime la ecuación

$$\log(invpc_t) = \beta_0 + \beta_1 \Delta\log(price_t) + \beta_2 t + u_t$$

y escriba los resultados en la forma usual. Interprete el coeficiente $\hat{\beta}_1$ y determine si en términos estadísticos es significativo.

iii) Elimine la tendencia lineal de log($invpc_t$) y use la versión sin tendencia como la variable dependiente en la regresión del inciso ii) (véase la sección 10.5). ¿Qué ocurre con la R^2?

iv) Ahora utilice Δlog($invpc_t$) como variable dependiente. ¿En qué difieren sus resultados respecto a los resultados del inciso ii)? ¿La tendencia temporal sigue siendo significativa? ¿Por qué?

C11.2 En el ejemplo 11.7, defina el crecimiento en el salario y la producción por hora como el cambio en el logaritmo natural log: *ghrwage* = Δlog(*hrwage*) y *goutphr* = Δlog(*outphr*). Considere una extensión simple del modelo estimado en la ecuación (11.29):

$$ghrwage_t = \beta_0 + \beta_1 goutphr_t + \beta_2 goutphr_{t-1} + u_t.$$

Esto permite que un mayor crecimiento de la productividad tenga un efecto actual y rezagado en el crecimiento salarial.

i) Estime la ecuación con los datos de EARNS.RAW y reporte los resultados en la forma usual. ¿El valor rezagado de *goutphr* es estadísticamente significativo?

ii) Si $\beta_1 + \beta_2 = 1$, un incremento permanente en el crecimiento de la productividad se transforma en un mayor crecimiento salarial después de un año. Pruebe H_0: $\beta_1 + \beta_2 = 1$ contra la alternativa de dos colas. Recuerde que la forma más sencilla de hacerlo es escribir la ecuación de modo que $\theta = \beta_1 + \beta_2$ aparezca de manera directa en el modelo, como en el ejemplo 10.4 del capítulo 10.

iii) ¿Es necesario que $goutphr_{t-2}$ esté en el modelo? Explique.

C11.3 i) En el ejemplo 11.4, quizás el valor esperado del rendimiento en el periodo t, dados los rendimientos anteriores, sea una función cuadrática de $return_{t-1}$. Para verificar esta posibilidad, haga una estimación con los datos de NYSE.RAW

$$return_t = \beta_0 + \beta_1 return_{t-1} + \beta_2 return_{t-1}^2 + u_t;$$

informe los resultados como acostumbra.

ii) Enuncie y pruebe la hipótesis nula de que $\mathrm{E}(return_t|return_{t-1})$ no depende de $return_{t-1}$. (*Sugerencia:* existen dos restricciones para esta prueba.) ¿Cuál es su conclusión?

iii) Suprima $return_{t-1}^2$ del modelo pero añada el término de interacción $return_{t-1} \cdot return_{t-2}$. Ahora pruebe la hipótesis de los mercados eficientes.

iv) ¿Cuál es su conclusión en cuanto a la predicción de los rendimientos accionarios semanales a partir de los rendimientos anteriores?

C11.4 Para este ejercicio utilice los datos de PHILLIPS.RAW, pero sólo hasta 1996.

i) En el ejemplo 11.5 se supuso que la tasa natural de desempleo es constante. Una forma alterna de la curva de Phillips aumentada por las expectativas hace que dicha tasa dependa de los niveles de desempleo anteriores. En el caso más simple, la tasa natural en el periodo t es igual a $unem_{t-1}$. Si se suponen expectativas adaptativas, se obtiene una curva de Phillips donde la inflación y el desempleo están en primeras diferencias:

$$\Delta inf = \beta_0 + \beta_1 \Delta unem + u.$$

Estime el modelo, muestre los resultados en la forma usual y analice el signo, magnitud y significado estadístico de $\hat{\beta}_1$.

ii) ¿Qué modelo se ajusta mejor a los datos, la ecuación (11.19) o el modelo del inciso i)? Explíquelo.

C11.5 i) Añada una tendencia lineal en el tiempo a la ecuación (11.27). ¿Es necesaria una tendencia en el tiempo en la ecuación de primeras diferencias?

ii) Suprima la tendencia en el tiempo y agregue las variables *ww2* y *pill* a la ecuación (11.27) (no diferencie estas variables binarias). ¿Son conjuntamente significativas estas variables al nivel de 5%?

iii) Con el modelo del inciso ii), estime la PLP y obtenga su error estándar. Compare esto con la ecuación (10.19), donde *gfr* y *pe* aparecen en niveles en vez de aparecer en primeras diferencias.

C11.6 Sean $inven_t$ el valor real de los inventarios en Estados Unidos durante el año t, GDP_t el producto interno bruto y $r3_t$ la tasa de interés real (*ex post*) de los bonos del Tesoro a tres meses. La tasa de interés real *ex post* es (aproximadamente) $r3_t = i3_t - inf_t$, donde $i3_t$ es la tasa a tres meses de los bonos del Tesoro e inf_t es la tasa de inflación anual [vea Mankiw (1994, sección 6.4)]. El cambio en inventarios, $\Delta inven_t$, es la *inversión en inventarios* para el año. El *modelo del acelerador* de la inversión en inventarios es

$$\Delta inven_t = \beta_0 + \beta_1 \Delta GDP_t + u_t,$$

donde $\beta_1 > 0$. [Vea, por ejemplo, Mankiw (1994, capítulo 17).]

i) Use los datos de INVEN.RAW para estimar el modelo del acelerador. Reporte los resultados en la forma usual e interprete la ecuación. ¿En términos estadísticos $\hat{\beta}_1$ es mayor que cero?

ii) Si la tasa de interés real aumenta, crece también el costo de oportunidad de mantener inventarios, y por tanto un incremento en la tasa de interés los reduciría. Añada esta tasa de interés real al modelo del acelerador y analice los resultados.

iii) ¿Funciona mejor el nivel de la tasa de interés real que su primera diferencia, $\Delta r3_t$?

C11.7 Utilice la base de datos CONSUMP.RAW para este ejercicio. Una versión de la *hipótesis del ingreso permanente* (HIP) del consumo es que el *crecimiento* en el consumo es impredecible. [Otra versión es que el cambio en el consumo mismo es impredecible; consulte Mankiw (1994, capítulo 15) para el análisis de la HIP.] Sea $gc_t = \log(c_t) - \log(c_{t-1})$ el crecimiento en el consumo real *per cápita* (de bienes no durables y servicios). Por tanto, la HIP implica que $E(gc_t|I_{t-1}) = E(gc_t)$, donde I_{t-1} denota la información conocida en el periodo $(t - 1)$; en este caso, t indica un año.

i) Pruebe la HIP mediante la estimación de $gc_t = \beta_0 + \beta_1 gc_{t-1} + u_t$. Enuncie claramente las hipótesis nula y alternativa. ¿Qué concluye?

ii) A la regresión del inciso i), añada gy_{t-1} e $i3_{t-1}$, donde gy_t es el crecimiento del ingreso real disponible *per cápita* e $i3_t$ es la tasa de interés sobre los bonos del Tesoro a tres meses; observe que cada uno debe rezagarse en la regresión. ¿Estas dos variables adicionales son conjuntamente significativas?

C11.8 Para este ejercicio use los datos de PHILLIPS.RAW.

i) Estime un modelo AR(1) para la tasa de desempleo. Mediante esta ecuación haga una predicción de la tasa de desempleo para 2004. Compárela con la tasa de desempleo real de ese año. (Esta información se encuentra en cualquier edición reciente del informe *Economic Report of the President*.)

ii) Añada un rezago de la inflación al modelo AR(1) del inciso i). ¿Es inf_{t-1} estadísticamente significativa?

iii) Con la ecuación del inciso ii) haga una predicción de la tasa de desempleo para 2004. ¿El resultado es mejor o peor que el del modelo del inciso i)?

iv) Utilice el método de la sección 6.4 para elaborar un intervalo de predicción al 95% para la tasa de desempleo de 2004. ¿Se encuentra la tasa de desempleo de 2004 en el intervalo?

C11.9 Use los datos de TRAFFIC2.RAW para este ejercicio. En el ejercicio para computadora C10.11 se pidió hacer un análisis de estos datos.

i) Calcule el coeficiente de correlación de primer orden para la variable *prcfat*. ¿Le preocupa que *prcfat* contenga una raíz unitaria? Haga lo mismo para la tasa de desempleo (*unem*).

ii) Estime un modelo de regresión múltiple que relacione la primera diferencia de *prcfat*, Δ*prcfat*, con las mismas variables del inciso vi) del ejercicio para computadora C10.11, con la excepción de que también debe obtener la primera diferencia de la tasa de desempleo. Luego incluya una tendencia lineal en el tiempo, variables binarias mensuales (feb, ..., dec), la variable de fin de semana (*wkends*) y las dos variables de políticas (*spdlaw* y *beltlaw*), sin diferenciarlas. ¿Los resultados le parecen interesantes?

iii) Comente el enunciado siguiente: "Siempre se debe obtener la primera diferencia de cualquier serie de tiempo para la que se sospeche que tiene una raíz unitaria antes de hacer la regresión simple, puesto que es la estrategia segura y debe dar resultados parecidos al uso de niveles". [Para resolver este ejercicio, tal vez quiera hacer la regresión del inciso vi) del ejercicio para computadora C10.11, si es que aún no la ha hecho.]

C11.10 Utilice la base de datos PHILLIPS.RAW para responder esta pregunta. Ahora debe usar los datos de los 56 años.

i) Vuelva a estimar la ecuación (11.19) y reporte los resultados de la manera acostumbrada. ¿Las estimaciones del intercepto y la pendiente cambian de manera notable cuando añade los datos de los años recientes?

ii) Obtenga una nueva estimación de la tasa natural de desempleo. Compare esta nueva estimación con aquélla reportada en el ejemplo 11.5.

iii) Calcule la autocorrelación de primer orden para *unem*. En su opinión, ¿la raíz está cerca de uno?

iv) Use $\Delta unem$ como la variable explicativa en lugar de *unem*. ¿Cuál variable explicativa da una R-cuadrada mayor?

C11.11 La ley de Okun —vea, por ejemplo, Mankiw (1994, capítulo 2)—, supone la siguiente relación entre el cambio porcentual anual del PIB real, *pcrgdp*, y el cambio en la tasa de desempleo anual, $\Delta unem$:

$$pcrgdp = 3 - 2 \cdot \Delta unem.$$

Si la tasa de desempleo es estable, el PIB real crece a 3% anual. Por cada incremento de un punto porcentual en la tasa de desempleo, el PIB real decrece dos puntos porcentuales. (Esto no debe interpretarse en ningún sentido causal; es más parecido a una descripción estadística.)

Para ver si los datos sobre la economía estadounidense respaldan la ley de Okun, se especifica un modelo que permite desviaciones a través de un término de error, $pcrgdp_t = \beta_0 + \beta_1 \Delta unem_t + u_t$.

i) Use los datos de OKUN.RAW para estimar la ecuación. ¿Obtiene exactamente 3 para el intercepto y −2 para la pendiente? ¿Esperaba que así fuera?

ii) Encuentre el estadístico t para probar H_0: $\beta_1 = -2$. ¿Rechaza H_0 contra la alternativa de dos colas a cualquier nivel de significancia razonable?

iii) Calcule el estadístico t para probar H_0: $\beta_0 = 3$. ¿Rechaza H_0 al nivel de 5% contra la alternativa de dos colas? ¿El rechazo es "sólido"?

iv) Obtenga el estadístico F y el valor-p para probar H_0: $\beta_0 = 3$, $\beta_1 = -2$ contra la alternativa de que H_0 es falsa. En general, ¿diría usted que los datos rechazan o tienden a respaldar la ley de Okun?

C11.12 Para este ejercicio, utilice la base de datos MINWAGE.RAW, concentrándose en las series de salario y empleo para el sector 232 (accesorios de vestir de hombres y niños). La variable *gwage*232 es el crecimiento mensual (el cambio en los logaritmos) del salario promedio en el sector 232; *gemp*232 es el crecimiento en el empleo en el sector 232; *gmwage* es el crecimiento en el salario mínimo federal y *gcpi* es el crecimiento en el Índice de Precios al Consumidor (urbano).

i) Calcule la autocorrelación de primer orden en *gwage*232. ¿Esta serie parece ser débilmente dependiente?

ii) Estime el modelo dinámico

$$gwage232_t = \beta_0 + \beta_1 gwage232_{t-1} + \beta_2 gmwage_t + \beta_3 gcpi_t + u_t$$

por MCO. Manteniendo fijo el crecimiento en el salario del último mes y el crecimiento en el IPC, ¿un incremento en el salario mínimo federal da como resultado un incremento contemporáneo en $gwage232_t$? Explique.

iii) Ahora añada el crecimiento rezagado en el empleo, $gemp232_{t-1}$, a la ecuación del inciso ii). ¿Es en términos estadísticos significativo?

iv) Comparado con el modelo sin $gwage232_{t-1}$ y $gemp232_{t-1}$, ¿la adición de dos variables rezagadas afecta mucho al coeficiente de *gmwage*?

v) Ejecute la regresión de $gmwage_t$ sobre $gwage232_{t-1}$ y $gemp232_{t-1}$, y reporte la R-cuadrada. Comente cómo el valor de la R-cuadrada le ayuda a explicar su respuesta del inciso iv).

CAPÍTULO 12

Correlación serial y heterocedasticidad en regresiones de series de tiempo

En este capítulo se analizará el problema crítico de la correlación serial en los términos de error de un modelo de regresión múltiple. En el capítulo 11 se vio que cuando la dinámica de un modelo se especifica de forma completa, de manera apropiada, los errores no están serialmente correlacionados. Por consiguiente, se pueden usar las pruebas de correlación serial para detectar especificaciones dinámicas incorrectas. Además, los modelos de rezagos distribuidos finitos y los modelos estáticos con frecuencia tienen errores serialmente correlacionados incluso si no hay una especificación incorrecta subyacente al modelo. Por tanto, es importante saber las consecuencias y soluciones para la correlación serial en estas útiles clases de modelos.

En la sección 12.1 se presentan las propiedades de MCO cuando los errores contienen correlación serial. En la sección 12.2 se demuestra cómo probar si hay correlación serial. Se estudian las pruebas que se aplican a los modelos con regresores estrictamente exógenos y las pruebas que son asintóticamente válidas con regresores generales, incluidas las variables dependientes rezagadas. La sección 12.3 explica cómo corregir la correlación serial bajo el supuesto de las variables explicativas estrictamente exógenas, mientras que la sección 12.4 muestra cómo el uso de datos diferenciados a menudo elimina la correlación serial en los errores. La sección 12.5 cubre los avances más recientes sobre la manera de ajustar los errores estándar usuales de MCO y los estadísticos de prueba, en presencia de una correlación serial muy general.

En el capítulo 8 se vio la prueba y corrección de la heterocedasticidad en aplicaciones de corte transversal. En la sección 12.6 se muestra cómo los métodos utilizados en el caso del corte transversal pueden extenderse al caso de las series de tiempo. La mecánica es en esencia la misma, pero existen algunas sutilezas asociadas con la correlación temporal en las observaciones de las series de tiempo que deben abordarse. Asimismo, se tocan de manera breve las consecuencias de las formas dinámicas de la heterocedasticidad.

12.1 Propiedades de MCO con errores correlacionados serialmente

Insesgamiento y consistencia

En el capítulo 10, se demostró el insesgamiento del estimador de MCO bajo los primeros tres supuestos de Gauss-Markov para las regresiones de series de tiempo (ST.1 a ST.3). En particular, el teorema 10.1 no da por sentado nada respecto a la correlación serial en los errores. Se deduce que, siempre y cuando las variables explicativas sean estrictamente exógenas, las $\hat{\beta}_j$ son insesgadas, sin importar el grado de correlación serial en los errores. Esto es análogo a la observación de que la heterocedasticidad en los errores no causa un sesgo en las $\hat{\beta}_j$.

En el capítulo 11 se relajó el supuesto de exogeneidad estricta a $\mathrm{E}(u_t|\mathbf{x}_t) = 0$ y se mostró que, cuando los datos son débilmente dependientes, las $\hat{\beta}_j$ siguen siendo consistentes (aunque no necesariamente insesgadas). Este resultado no dependía de ningún supuesto sobre la correlación serial en los errores.

Eficiencia e inferencia

Dado que el teorema de Gauss-Markov (teorema 10.4) requiere tanto homocedasticidad como errores no correlacionados serialmente, los estimadores de MCO ya no son MELI en presencia de la correlación serial. Aún más importante resulta que los errores estándar usuales de MCO y los estadísticos de prueba no son válidos, incluso asintóticamente. Se puede ver esto al calcular la varianza del estimador de MCO bajo los primeros cuatro supuestos de Gauss-Markov y el modelo de **correlación serial AR(1)** para los términos de error. Dicho con más precisión, se supone que

$$u_t = \rho u_{t-1} + e_t,\ t = 1, 2, \ldots, n \tag{12.1}$$

$$|\rho| < 1, \tag{12.2}$$

donde e_t son variables aleatorias no correlacionadas con media cero y varianza σ_e^2; recuerde del capítulo 11 que el supuesto 12.2 es la condición de estabilidad.

Considere la varianza del estimador de pendiente de MCO en el modelo de regresión simple

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

y, sólo para simplificar la fórmula, se supone que el promedio muestral de x_t es cero ($\bar{x} = 0$). Por tanto, el estimador $\hat{\beta}_1$ de MCO de β_1 puede escribirse como

$$\hat{\beta}_1 = \beta_1 + \mathrm{STC}_x^{-1} \sum_{t=1}^{n} x_t u_t, \tag{12.3}$$

donde $\mathrm{STC}_x = \sum_{t=1}^{n} x_t^2$. Ahora bien, al calcular la varianza de $\hat{\beta}_1$ (condicional en $\mathbf{X}$), se debe representar la correlación serial en las u_t:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{STC}_x^{-2}\mathrm{Var}\left(\sum_{t=1}^{n} x_t u_t\right) \\
&= \mathrm{STC}_x^{-2}\left(\sum_{t=1}^{n} x_t^2 \mathrm{Var}(u_t) + 2\sum_{t=1}^{n-1}\sum_{j=1}^{n-t} x_t x_{t+j}\mathrm{E}(u_t u_{t+j})\right) \\
&= \sigma^2/\mathrm{STC}_x + 2(\sigma^2/\mathrm{STC}_x^2)\sum_{t=1}^{n-1}\sum_{j=1}^{n-t} \rho^j x_t x_{t+j},
\end{aligned} \tag{12.4}$$

donde $\sigma^2 = \mathrm{Var}(u_t)$ y se ha usado el hecho de que $\mathrm{E}(u_t u_{t+j}) = \mathrm{Cov}(u_t, u_{t+j}) = \rho^j \sigma^2$ [véase la ecuación (11.4)]. El primer término de la ecuación (12.4), σ^2/STC_x, es la varianza de $\hat{\beta}_1$ cuando $\rho = 0$, que es la varianza usual de MCO bajo los supuestos de Gauss-Markov. Si se ignora la correlación serial y se estima la varianza en la forma usual, el estimador de la varianza, por lo general, será sesgado cuando $\rho \neq 0$ ya que ignora el segundo término en la ecuación (12.4). Como se verá en ejemplos posteriores, $\rho > 0$ es más común, en cuyo caso, $\rho^j > 0$ para toda j. Además, las variables independientes en los modelos de regresión a menudo se correlacionan de forma positiva en el tiempo, de modo que $x_t x_{t+j}$ es positivo para casi todos los pares t y $t + j$. Por ende, en la mayor parte de las aplicaciones económicas, el término $\sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}$ es positivo, y por ello la fórmula usual de la varianza de MCO σ^2/STC_x *subestima* la varianza verdadera del estimador de MCO. Si ρ es grande o x_t tiene un grado alto de correlación serial positiva —caso común—, el sesgo del estimador usual de la varianza de MCO puede ser sustancial. En tal caso, se considerará que el estimador MCO de la pendiente es más preciso de lo que en realidad es.

Cuando $\rho < 0$, ρ^j es negativa cuando j es impar y positiva cuando es par, y por ello resulta difícil determinar el signo de $\sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}$. De hecho, es posible que la fórmula usual de MCO para la varianza, en realidad *sobrestime* la verdadera varianza de $\hat{\beta}_1$. En cualquier caso, el estimador usual de la varianza es sesgado para $\text{Var}(\hat{\beta}_1)$ en presencia de correlación serial.

Dado que el error estándar de $\hat{\beta}_1$ es una estimación de la desviación estándar de $\hat{\beta}_1$, no es válido utilizar el error estándar usual de MCO en presencia de correlación serial. Por tanto, los estadísticos t ya no pueden utilizarse para probar hipótesis simples. Dado que un error estándar más pequeño significa un estadístico t mayor, los estadísticos t usuales a menudo serán muy grandes cuando $\rho > 0$. Tampoco son válidos los estadísticos F y ML para probar hipótesis múltiples.

Pregunta 12.1

Suponga que, más que el modelo AR(1), u_t sigue el modelo MA(1) $u_t = e_t + \alpha e_{t-1}$. Determine $\text{Var}(\hat{\beta}_1)$ y demuestre que es distinto de la fórmula usual si $\alpha \neq 0$.

Bondad de ajuste

En ocasiones se destaca la afirmación de que la correlación serial en los errores de un modelo de regresión de series de tiempo invalida nuestras medidas usuales de bondad de ajuste, R-cuadrada y R-cuadrada ajustada. Por fortuna, éste no es el caso, siempre y cuando los datos sean estacionarios y débilmente dependientes. Para ver por qué estas medidas siguen siendo válidas, recuerde que se definió que la R-cuadrada poblacional en un contexto de corte transversal era $1 - \sigma_u^2/\sigma_y^2$ (vea la sección 6.3). Esta definición sigue siendo apropiada en el contexto de la regresión de series de tiempo con datos estacionarios débilmente dependientes: las varianzas tanto de los errores como de la variable dependiente no cambian con el tiempo. Por la ley de los números grandes, tanto, R^2 como $\bar{R}^2$ estiman de manera consistente la R-cuadrada poblacional. El argumento es en esencia el mismo que en el caso del corte transversal en presencia de heterocedasticidad (vea la sección 8.1). Debido a que nunca hay un estimador insesgado de la R-cuadrada poblacional, no tiene sentido hablar sobre sesgos en R^2 provocados por la correlación serial. En realidad, todo lo que se puede decir es que nuestras medidas de bondad de ajuste siguen siendo estimadores consistentes del parámetro poblacional. Este argumento no es válido si $\{y_t\}$ es un proceso I(1) ya que $\text{Var}(y_t)$ crece con t; la bondad de ajuste no tiene mucho sentido en este caso. Como se vio en la sección 10.5, las tendencias en la media de y_t, o la estacionalidad, pueden y deben tomarse en cuenta al calcular la R-cuadrada. Otras desviaciones de la estacionariedad no generan dificultad en la interpretación de R^2 y $\bar{R}^2$ en las formas acostumbradas.

Correlación serial en presencia de variables dependientes rezagadas

A menudo, a los principiantes en econometría se les advierte de los peligros de los errores correlacionados serialmente en presencia de variables dependientes rezagadas. Casi todo libro sobre la materia contiene en cierta forma el planteamiento siguiente: "los estimadores de MCO son inconsistentes en presencia de variables dependientes rezagadas y de errores correlacionados serialmente". Por desgracia, como aseveración general, este enunciado es falso. Hay una versión del planteamiento que es correcta, pero es necesario ser muy precisos.

Para ilustrar esto, suponga que el valor esperado de y_t dado y_{t-1} es lineal:

$$E(y_t|y_{t-1}) = \beta_0 + \beta_1 y_{t-1}, \tag{12.5}$$

en donde se supone estabilidad, $|\beta_1| < 1$. Se sabe que siempre se puede escribir esto con un término de error como

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \tag{12.6}$$

$$E(u_t|y_{t-1}) = 0. \tag{12.7}$$

Por construcción, este modelo satisface el supuesto clave de la media condicional cero ST.3′ para la consistencia de MCO y, por ende, los estimadores de MCO $\hat{\beta}_0$ y $\hat{\beta}_1$ son consistentes. Es importante notar que, sin supuestos adicionales, los errores $\{u_t\}$ *pueden* estar correlacionados serialmente. La condición (12.7) asegura que u_t no se correlaciona con y_{t-1}, pero podría hacerlo con y_{t-2}. Entonces, ya que $u_{t-1} = y_{t-1} - \beta_0 - \beta_1 y_{t-2}$, la covarianza entre u_t y u_{t-1} es $-\beta_1 \text{Cov}(u_t, y_{t-2})$, que no necesariamente es cero. Así, los errores muestran correlación serial y el modelo contiene una variable dependiente rezagada, pero MCO estima consistentemente β_0 y β_1 ya que éstos son los parámetros de la esperanza condicional (12.5). La correlación serial en los errores hará que los estadísticos usuales de MCO no sean válidos para fines de pruebas, pero no afectará la consistencia.

Ahora bien, ¿cuándo MCO son inconsistentes si los errores se correlacionan serialmente y los regresores contienen una variable dependiente rezagada? Esto sucede cuando el modelo se escribe en forma de error, exactamente como en la ecuación (12.6), pero entonces se *supone* que $\{u_t\}$ sigue un modelo AR(1) estable como en las ecuaciones (12.1) y (12.2), en donde

$$E(e_t|u_{t-1}, u_{t-2}, \ldots) = E(e_t|y_{t-1}, y_{t-2}, \ldots) = 0. \tag{12.8}$$

Por la suposición de que e_t no está correlacionada con y_{t-1} sucede que $\text{Cov}(y_{t-1}, u_t) = \rho\text{Cov}(y_{t-1}, u_{t-1})$, lo cual no es cero a menos que $\rho = 0$. Esto hace que los estimadores de MCO de β_0 y β_1 de la regresión de y_t sobre y_{t-1} sean inconsistentes.

Ahora se observa que la estimación por MCO de la ecuación (12.6), cuando los errores u_t siguen también un modelo AR(1), conduce a estimadores inconsistentes. Sin embargo, la exactitud de esta afirmación no la vuelve menos desatinada. Cabe preguntar: ¿qué sentido tendría estimar los parámetros de la ecuación (12.6) cuando los errores siguen un modelo AR(1)? Resulta difícil pensar en casos en los que esto fuera interesante. Cuando menos en la ecuación (12.5) los parámetros indican el valor esperado de y_t, dada y_{t-1}. Al combinar las ecuaciones (12.6) y (12.1), se ve que y_t en realidad sigue un modelo autoregresivo de segundo orden, o modelo AR(2). Para verlo, escriba $u_{t-1} = y_{t-1} - \beta_0 - \beta_1 y_{t-2}$ y sustituya esto en $u_t = \rho u_{t-1} + e_t$. Por tanto, la ecuación (12.6) puede rescribirse como

$$\begin{aligned} y_t &= \beta_0 + \beta_1 y_{t-1} + \rho(y_{t-1} - \beta_0 - \beta_1 y_{t-2}) + e_t \\ &= \beta_0(1 - \rho) + (\beta_1 + \rho)y_{t-1} - \rho\beta_1 y_{t-2} + e_t \\ &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t, \end{aligned}$$

donde $\alpha_0 = \beta_0(1 - \rho)$, $\alpha_1 = \beta_1 + \rho$, y $\alpha_2 = -\rho\beta_1$. Dada la ecuación (12.8), se deduce que

$$\mathrm{E}(y_t|y_{t-1}, y_{t-2}, \ldots) = \mathrm{E}(y_t|y_{t-1}, y_{t-2}) = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2}. \tag{12.9}$$

Esto significa que el valor esperado de y_t, dadas todas las y, anteriores, depende de *dos* rezagos de y. Es la ecuación (12.9) la que nos interesa utilizar para cualquier propósito práctico, incluido el pronóstico, como se verá en el capítulo 18. De especial interés son los parámetros α_j. Bajo condiciones de estabilidad adecuadas para un modelo AR(2), lo cual se cubre en la sección 12.3, la estimación por MCO de la ecuación (12.9) genera estimadores consistentes y asintóticamente normales de α_j.

En resumen, se necesita una buena razón para tener en un modelo, tanto una variable dependiente rezagada como una forma particular de correlación serial en los errores. A menudo, la correlación serial en los errores de un modelo dinámico señala sencillamente que la función dinámica de la regresión no se ha especificado por completo: en el ejemplo anterior se debió agregar y_{t-2} a la ecuación.

En el capítulo 18 se verán algunos ejemplos de modelos con variables dependientes rezagadas en los que los errores están correlacionados serialmente y también con y_{t-1}. Pero incluso en estos casos, los errores no siguen un proceso autoregresivo.

12.2 Métodos de prueba de la correlación serial

En esta sección, se analizan diversos métodos de prueba de correlación serial en los términos de error del modelo de regresión lineal múltiple

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t.$$

Considere primero el caso en que los regresores son estrictamente exógenos. Recuerde que esto exige que el error u_t no se correlacione con los regresores en todos los periodos (vea la sección 10.3), y de esta manera se descartan, entre otras cosas, los modelos con variables dependientes rezagadas.

Prueba *t* de correlación serial AR(1) con regresores estrictamente exógenos

Si bien existen numerosas maneras en que los términos de error en un modelo de regresión múltiple pueden estar serialmente correlacionados, el modelo más popular —y el más sencillo para trabajar— es el modelo AR(1) de las ecuaciones (12.1) y (12.2). En la sección anterior se explicaron las consecuencias de realizar MCO cuando los errores están en general correlacionados serialmente, y se obtuvo la varianza del estimador de pendiente de MCO en un modelo de regresión simple con errores AR(1). Ahora se mostrará cómo probar la presencia de la correlación serial AR(1). La hipótesis nula es que *no* hay correlación serial. Por tanto, de la misma manera que con las pruebas de heterocedasticidad, se supone lo mejor y se requiere proporcionar a los datos evidencia razonablemente sólida de que se viola el supuesto ideal de no correlación serial.

Primero se deriva una prueba para muestras grandes de acuerdo con el supuesto de que las variables explicativas son estrictamente exógenas: el valor esperado de u_t, dada toda la historia de las variables independientes, es cero. Además, en la ecuación (12.1) se debe suponer que

$$\mathrm{E}(e_t|u_{t-1}, u_{t-2}, \ldots) = 0 \tag{12.10}$$

y

$$\mathrm{Var}(e_t|u_{t-1}) = \mathrm{Var}(e_t) = \sigma_e^2. \tag{12.11}$$

Estos son los supuestos estándar del modelo AR(1) (que se siguen cuando $\{e_t\}$ es una secuencia i.i.d.) y nos permiten aplicar los resultados de muestras grandes del capítulo 11 a la regresión dinámica.

Al igual que en la prueba de heterocedasticidad, la hipótesis nula es que el supuesto de Gauss-Markov adecuado es verdadero. En el modelo AR(1), la hipótesis nula de que los errores no se correlacionan de forma serial es

$$\mathrm{H}_0\colon \rho = 0. \tag{12.12}$$

¿Cómo se puede probar esta hipótesis? Si se observara la u_t entonces, bajo los supuestos (12.10) y (12.11), se aplicarían de inmediato los resultados de normalidad asintótica del teorema 11.2 al modelo de regresión dinámica

$$u_t = \rho u_{t-1} + e_t, t = 2, \ldots, n. \tag{12.13}$$

(De acuerdo con la hipótesis nula $\rho = 0$, $\{u_t\}$ a todas luces es débilmente dependiente.) En otras palabras, se puede estimar ρ de la regresión de u_t sobre u_{t-1}, para toda $t = 2, \ldots, n$, sin un intercepto y utilizar el estadístico t usual para $\hat{\rho}$. Sin embargo, esto no funciona ya que no se observan los errores u_t. No obstante, igual que con la prueba de heterocedasticidad, se puede reemplazar u_t con el residual de MCO correspondiente, $\hat{u}_t$. En virtud de que $\hat{u}_t$ depende de los estimadores MCO $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, no es obvio que al utilizar $\hat{u}_t$ en lugar de u_t en la regresión no tenga un efecto en la distribución del estadístico t. Por fortuna, resulta que, por el supuesto de exogeneidad estricta, la distribución de muestra grande del estadístico t no se ve afectada por el uso de los residuales de MCO en lugar de los errores. La prueba rebasa el alcance de este libro, pero se deduce de la obra de Wooldridge (1991b).

Se puede resumir la prueba asintótica para correlación serial AR(1) de manera muy sencilla.

Prueba de correlación serial AR(1) con regresores estrictamente exógenos:

i) Efectúe la regresión por MCO de y_t sobre $x_{t1}, \ldots, x_{tk}$ y obtenga los residuales de MCO, $\hat{u}_t$, para toda $t = 1, 2, \ldots, n$.

ii) Realice la regresión de

$$\hat{u}_t \text{ sobre } \hat{u}_{t-1}, \text{ para toda } t = 2, \ldots, n, \tag{12.14}$$

para obtener el coeficiente $\hat{\rho}$ de $\hat{u}_{t-1}$ y su estadístico t, $t_{\hat{\rho}}$. (Esta regresión puede contener o no un intercepto; el estadístico t para $\hat{\rho}$ se verá un poco afectado, pero de cualquier modo es asintóticamente válido.)

iii) Utilice $t_{\hat{\rho}}$ para probar H_0: $\rho = 0$ contra H_1: $\rho \neq 0$ en la forma común. (En realidad, ya que con frecuencia se espera que $\rho > 0$ la alternativa puede ser H_1: $\rho > 0$.) Por lo general, se llega a la conclusión de que la correlación serial es un problema que se debe enfrentar sólo si H_0 se rechaza al nivel de 5%. Como siempre, es mejor reportar el valor-p de la prueba.

Al decidir si la correlación serial necesita corregirse, se debe recordar la diferencia entre la significancia práctica y la estadística. Con un tamaño de muestra grande es posible encontrar correlación serial aun cuando $\hat{\rho}$ sea pequeño en términos prácticos; cuando $\hat{\rho}$ está cercano a cero, los procedimientos de inferencia usuales de MCO no están muy desviados [vea la ecuación (12.4)]. Tales resultados son poco frecuentes en las aplicaciones de series de tiempo ya que, por lo común, las bases de datos de las series de tiempo son muy cortas.

Ejemplo 12.1

[Prueba de correlación serial AR(1) en la curva de Phillips]

En el capítulo 10 se estimó una curva estática de Phillips que explicó el intercambio entre la inflación y el desempleo en Estados Unidos (vea el ejemplo 10.1). En el capítulo 11 se estudió una determinada curva de Phillips aumentada por las expectativas, en la que se supusieron expectativas adaptativas (vea el ejemplo 11.5). Ahora se prueba la correlación serial en el término de error de cada ecuación. Dado que la curva aumentada por las expectativas utiliza $\Delta inf_t = inf_t - inf_{t-1}$ como variable dependiente, se tiene una observación menos.

Para la curva estática de Phillips, la regresión en la ecuación (12.14) produce $\hat{\rho} = .573$, $t = 4.93$ y el valor-$p = .000$ (con 48 observaciones hasta 1996). Esta es una evidencia contundente de corrección serial positiva de primer orden. Una consecuencia de ello es que los errores estándar y los estadísticos t del capítulo 10 no son válidos. En comparación, la prueba de correlación serial AR(1) en la curva aumentada por las expectativas obtiene $\hat{\rho} = -.036$, $t = -.287$ y un valor-$p = .775$ (con 47 observaciones): no hay evidencia de correlación serial AR(1) en la curva de de Phillips aumentada por las expectativas.

Aunque la prueba en la ecuación (12.14) se deriva del modelo AR(1), también puede detectar otras clases de correlación serial. Recuerde que $\hat{\rho}$ es un estimador consistente de la correlación entre u_t y u_{t-1}. Cualquier correlación serial que provoque que los errores adyacentes se correlacionen puede detectarse mediante esta prueba. Por otra parte, no detecta correlación serial donde los errores adyacentes no están correlacionados, $\mathrm{Corr}(u_t, u_{t-1}) = 0$. (Por ejemplo, u_t y u_{t-2} podrían estar correlacionados.)

Pregunta 12.2

¿Cómo utilizaría la regresión (12.14) para construir un intervalo de confianza aproximado de 95% para ρ?

Al usar el estadístico t usual en la ecuación (12.14), se debe suponer que los errores en (12.13) satisfacen el supuesto de homocedasticidad adecuado, (12.11). De hecho, resulta sencillo hacer la prueba robusta a la heterocedasticidad en e_t: para ello se usa simplemente el estadístico t robusto a la heterocedasticidad usual del capítulo 8. Para la curva estática de Phillips del ejemplo 12.1, el estadístico t robusto a la heterocedasticidad es 4.03, que es menor al estadístico t que no es robusto, pero sigue siendo muy significativo. En la sección 12.6, se estudia más a fondo la heterocedasticidad en las regresiones de series de tiempo, incluidas sus formas dinámicas.

Prueba de Durbin-Watson bajo los supuestos clásicos

Otra prueba para la correlación serial AR(1) es la de Durbin-Watson. El **estadístico de Durbin-Watson (*DW*)** se basa también en los residuales de MCO:

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}. \qquad \boxed{12.15}$$

Algo de álgebra simple demuestra que el DW y $\hat{\rho}$ de (12.14) están estrechamente relacionados:

$$DW \approx 2(1 - \hat{\rho}). \qquad \boxed{12.16}$$

Una razón por la que esta relación no es exacta es que $\hat{\rho}$ tiene $\sum_{t=2}^{n} \hat{u}_{t-1}^2$ en su denominador, mientras que el estadístico DW tiene la suma de cuadrados de todos los residuales de MCO en su denominador. Incluso con tamaños de muestra moderados, la aproximación en la ecuación (12.16) con frecuencia es muy cercana. Por tanto, las pruebas basadas en DW y la prueba t basada en $\hat{\rho}$ son conceptualmente iguales.

Durbin y Watson (1950) obtienen la distribución de DW (condicional en $\mathbf{X}$), lo que requiere todo el conjunto de supuestos del modelo lineal clásico, incluida la normalidad de los términos de error. Por desgracia, esta distribución depende de los valores de las variables independientes. (También depende del tamaño de la muestra, del número de regresores y de si la regresión contiene un intercepto.) Aun cuando algunos paquetes de econometría tabulan los valores críticos y los valores-*p* para DW, algunos no lo hacen. De cualquier forma, dependen de todo el conjunto de supuestos del MLC.

Varios textos sobre econometría dan a conocer límites superiores e inferiores para los valores críticos que dependen del nivel de significancia deseado, la hipótesis alternativa, el número de observaciones y el número de regresores. (Se da por sentado que un intercepto se incluye en el modelo.) Por lo general, la prueba de DW se calcula para la alternativa

$$\mathrm{H}_1: \rho > 0. \qquad \boxed{12.17}$$

A partir de la aproximación en la ecuación (12.16), $\hat{\rho} \approx 0$ implica que $DW \approx 2$, y $\hat{\rho} > 0$ implica que $DW < 2$. Por consiguiente, para rechazar la hipótesis nula (12.12) a favor de (12.17), se busca un valor de DW que sea considerablemente menor que dos. Por desgracia, debido a los problemas en la obtención de la distribución nula de DW, se debe comparar DW con dos conjuntos de valores críticos. Por lo general, estos se etiquetan como d_U (para el *superior*) y d_L (para el *inferior*). Si $DW < d_L$, entonces se rechaza H_0 en favor de (12.17); si $DW > d_U$, no se rechaza H_0. Cuando $d_L \leq DW \leq d_U$, la prueba no es concluyente.

Como ejemplo, si se elige un nivel de significancia de 5% con $n = 45$ y $k = 4$, $d_U = 1.720$ entonces $d_L = 1.336$ [vea Savin y White (1977)]. Si $DW < 1.336$, se rechaza la hipótesis nula de que no hay correlación serial al nivel de 5%; si $DW > 1.72$, no se rechaza H_0; si $1.336 \leq DW \leq 1.72$, la prueba no es concluyente.

En el ejemplo 12.1, para la curva estática de Phillips, se calcula que $DW = .80$. Se puede obtener el valor crítico de 1% inferior de Savin y White (1977) para $k = 1$ y $n = 50$: $d_L = 1.32$. Por tanto, se rechaza la hipótesis nula de que no hay correlación serial contra la alternativa de una correlación serial positiva al nivel de 1%. (Usando la prueba t anterior, se puede concluir que el valor-p es igual a cero hasta tres posiciones decimales). Para la curva de Phillips aumentada por las expectativas, $DW = 1.77$, lo cual está bien dentro de la región de no rechazo incluso al nivel de 5% por ciento ($d_U = 1.59$).

El hecho de que una distribución de muestreo exacta para DW pueda tabularse es la única ventaja que DW tiene sobre la prueba t de la ecuación (12.14). Dado que los valores críticos tabulados son exactamente válidos sólo bajo el conjunto completo de supuestos del MLC y que éstos pueden conducir a una amplia región no concluyente, las desventajas prácticas del estadístico DW son sustanciales. El estadístico t de la ecuación (12.14) es fácil de calcular y asintóticamente válido sin errores normalmente distribuidos. El estadístico t también es válido en presencia de heterocedasticidad que depende de las x_{tj}. Además, es fácil hacerlo robusto a cualquier forma de heterocedasticidad.

Prueba de correlación serial AR(1) sin regresores estrictamente exógenos

Cuando las variables explicativas no son estrictamente exógenas, de manera que una o más x_{tj} están correlacionadas con u_{t-1}, ni la prueba t de la regresión de la ecuación (12.14) ni el estadístico de Durbin-Watson son válidos, incluso en muestras grandes. El caso principal de regresores sin exogeneidad estricta ocurre cuando el modelo contiene una variable dependiente rezagada: y_{t-1} y u_{t-1} evidentemente están correlacionadas. Durbin (1970) sugirió dos alternativas para el estadístico DW cuando el modelo contiene una variable dependiente rezagada y los otros regresores son no aleatorios (o, en términos más generales, estrictamente exógenos). La primera alternativa es el *estadístico* h *de Durbin*. Este estadístico tiene el inconveniente práctico de que no siempre se puede calcular, por lo que no se trata aquí.

El estadístico alternativo al de Durbin es simple de calcular y es válido cuando hay cualquier número de variables explicativas sin exogeneidad estricta.

Prueba de correlación serial con regresores generales:

i) Realice la regresión por MCO de y_t sobre $x_{t1}, \dots, x_{tk}$ y obtenga los residuales de MCO, $\hat{u}_t$, para toda $t = 1, 2, \dots, n$.

ii) Efectúe la regresión de

$$\hat{u}_t \text{ sobre } x_{t1}, x_{t2}, \dots, x_{tk}, \hat{u}_{t-1}, \text{ para toda } t = 2, \dots, n \tag{12.18}$$

para obtener el coeficiente $\hat{\rho}$ de $\hat{u}_{t-1}$ y su estadístico t, $t_{\hat{\rho}}$.

iii) Emplee $t_{\hat{\rho}}$ para probar H_0: $\rho = 0$ contra H_1: $\rho \neq 0$ en la forma usual (o contra una alternativa de una cola).

En la ecuación (12.18), se realiza la regresión de los residuales de MCO sobre *todas* las variables independientes, incluidos un intercepto y el residual rezagado. El estadístico t del residual rezagado es una prueba válida de (12.12) para el modelo AR(1) de la ecuación (12.13) [cuando se añade $\text{Var}(u_t|\mathbf{x}_t, u_{t-1}) = \sigma^2$ bajo H_0]. Cualquier número de variables dependientes rezagadas podrían aparecer entre las x_{tj}, y también se pueden permitir otras variables explicativas sin exogeneidad estricta.

La inclusión de $x_{t1}, \dots, x_{tk}$ explícitamente permite que cada x_{tj} esté correlacionada con u_{t-1}, y esto asegura que $t_{\hat{\rho}}$ tenga una distribución t aproximada para muestras grandes. El estadístico t de (12.14) ignora la posible correlación entre x_{tj} y u_{t-1}, de manera que no es válido sin regresores estrictamente exógenos. Por cierto, debido a que $\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \dots - \hat{\beta}_k x_{tk}$, se puede mostrar que el estadístico t de $\hat{u}_{t-1}$ es el mismo si y_t se usa en lugar de $\hat{u}_t$ como variable dependiente en (12.18).

El estadístico t de la ecuación (12.18) se vuelve fácilmente robusto a la heterocedasticidad de forma desconocida [en particular, cuando $\text{Var}(u_t|\mathbf{x}_t, u_{t-1})$ no es constante]: simplemente use el estadístico t robusto a la heterocedasticidad para $\hat{u}_{t-1}$.

Ejemplo 12.2

[Prueba de correlación serial AR(1) en la ecuación de salario mínimo]

En el capítulo 10 (vea el ejemplo 10.9), se estimó el efecto del salario mínimo sobre la tasa de empleo en Puerto Rico. Ahora se revisa si los errores parecen contener correlación serial, usando la prueba que no supone exogeneidad estricta del salario mínimo o del producto nacional bruto. [Se añade la variable log del PNB real de Puerto Rico a la ecuación (10.38), como en el ejercicio para computadora C10.3.] Se está suponiendo que los procesos estocásticos subyacentes son débilmente dependientes, pero se les permite contener una tendencia lineal en el tiempo al incluir t en la regresión.

Si $\hat{u}_t$ denota los residuales de MCO, se hace la regresión de

$$\hat{u}_t \text{ sobre } \log(mincov_t), \log(prgnp_t), \log(usgnp_t), t \text{ y } \hat{u}_{t-1},$$

usando las 37 observaciones disponibles. El coeficiente estimado de $\hat{u}_{t-1}$ es $\hat{\rho} = .481$ con $t = 2.89$ (valor-p de dos colas $= .007$). Por consiguiente, existe evidencia contundente de correlación serial AR(1) en los errores, lo que significa que los estadísticos t para las $\hat{\beta}_j$ que antes se obtuvieron no son válidos para hacer inferencias. Recuerde, sin embargo, que las $\hat{\beta}_j$ siguen siendo consistentes si u_t no se correlaciona contemporáneamente con cada variable explicativa. A propósito, si en su lugar usa la regresión (12.14), se obtiene $\hat{\rho} = .417$ y $t = 2.63$, por lo que el resultado de la prueba es parecido en este caso.

Prueba de correlación serial de orden superior

La prueba de la ecuación (12.18) se extiende con facilidad a órdenes superiores de correlación serial. Por ejemplo, suponga que se desea probar

$$H_0: \rho_1 = 0, \rho_2 = 0 \tag{12.19}$$

en el modelo AR(2),

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t.$$

Este modelo alterno de correlación serial nos permite probar la *correlación serial de segundo orden*. Como siempre, se estima el modelo por medio de MCO y se obtienen los residuales de MCO, $\hat{u}_t$. Luego, se lleva a cabo la regresión de

$$\hat{u}_t \text{ sobre } x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1} \text{ y } \hat{u}_{t-2}, \text{ para toda } t = 3, \ldots, n,$$

con el propósito de obtener la prueba F de significancia conjunta para $\hat{u}_{t-1}$ y $\hat{u}_{t-2}$. Si estos dos rezagos son conjuntamente significativos a un nivel muy pequeño, por ejemplo, de 5%, entonces se rechaza (12.19) y se llega a la conclusión de que los errores se correlacionan serialmente.

De manera más general, se puede probar la correlación serial del modelo autoregresivo de orden q:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_q u_{t-q} + e_t. \tag{12.20}$$

La hipótesis nula es

$$\mathrm{H}_0: \rho_1 = 0, \rho_2 = 0, \ldots, \rho_q = 0. \tag{12.21}$$

Prueba de correlación serial AR(q):

i) Realice la regresión por MCO de y_t sobre $x_{t1}, \ldots, x_{tk}$ y obtenga los residuales de MCO, $\hat{u}_t$, para toda $t = 1, 2, \ldots, n$.

ii) Efectúe la regresión de

$$\hat{u}_t \text{ sobre } x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}, \text{ para toda } t = (q + 1), \ldots, n. \tag{12.22}$$

iii) Calcule la prueba F de significancia conjunta de $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}$ en (12.22). [También se puede utilizar el estadístico F con y_t como variable dependiente en (12.22), ya que se obtiene una respuesta idéntica.]

Si se supone que las x_{tj} son estrictamente exógenas, de modo que cada x_{tj} no se correlaciona con $u_{t-1}, u_{t-2}, \ldots, u_{t-q}$, entonces las x_{tj} pueden omitirse de (12.22). La inclusión de las x_{tj} en la regresión hace que la prueba sea válida con o sin el supuesto de exogeneidad estricta. La prueba requiere el supuesto de homocedasticidad

$$\mathrm{Var}(u_t|\mathbf{x}_t, u_{t-1}, \ldots, u_{t-q}) = \sigma^2. \tag{12.23}$$

Una versión robusta a la heterocedasticidad se calcula tal como se explicó en el capítulo 8.

Una alternativa para calcular la prueba F es usar la forma del multiplicador de Lagrange (*ML*) del estadístico. (En el capítulo 5 se planteó el estadístico *ML* para probar las restricciones de exclusión en el análisis de corte transversal.) El estadístico *ML* para probar (12.21) es sencillamente

$$LM = (n - q)R^2_{\hat{u}}, \tag{12.24}$$

donde $R^2_{\hat{u}}$ es sólo la R-cuadrada usual de la regresión (12.22). Bajo la hipótesis nula, $ML \overset{a}{\sim} \chi^2_q$. Ésta a menudo se denomina **prueba de Breusch-Godfrey** para correlación serial AR(q). El estadístico *ML* exige también (12.23), pero puede volverse robusto a la heterocedasticidad. [Para detalles, vea Wooldridge (1991b).]

Ejemplo 12.3

[Prueba de correlación serial AR(3)]

En el estudio de evento de la industria de cloruro de bario (vea el ejemplo 10.5), se usaron datos mensuales, de modo que se podrían probar órdenes superiores de correlación serial. Para fines ilustrativos, se prueba la correlación serial AR(3) de los errores subyacentes en la ecuación (10.22). Usando la regresión (12.22), el estadístico F de significancia conjunta para $\hat{u}_{t-1}$, $\hat{u}_{t-2}$ y $\hat{u}_{t-3}$ es $F = 5.12$. En un principio, se tenía $n = 131$ y se pierden tres observaciones en la regresión auxiliar (12.22). Como para este ejemplo se estiman 10 parámetros en (12.22), los *gl* del estadístico F son 3 y 118. El valor-p del estadístico F es .0023 de modo que hay evidencia contundente de correlación serial AR(3).

Con datos trimestrales o mensuales que no se hayan ajustado estacionalmente, a veces se desea probar formas estacionales de correlación serial. Por ejemplo, con datos trimestrales se podría postular el modelo autorregresivo

$$u_t = \rho_4 u_{t-4} + e_t. \qquad \textbf{12.25}$$

De las pruebas de correlación serial AR(1), queda bastante claro cómo proceder. Cuando los regresores son estrictamente exógenos, se puede usar una prueba t para $\hat{u}_{t-4}$ en la regresión de

$$\hat{u}_t \text{ sobre } \hat{u}_{t-4}, \text{ para toda } t = 5, \ldots, n.$$

También se dispone de una modificación del estadístico de Durbin-Watson [vea Wallis (1972)]. Cuando las x_{tj} no son estrictamente exógenas, se usa la regresión en (12.18), con $\hat{u}_{t-4}$ que reemplaza a $\hat{u}_{t-1}$.

En el ejemplo 12.3, los datos son mensuales y no están ajustados estacionalmente. Por tanto, tiene sentido probar la correlación entre u_t y u_{t-12}. Una regresión de $\hat{u}_t$ sobre $\hat{u}_{t-12}$ proporciona $\hat{\rho}_{12} = -.187$ y un valor-p = .028, de modo que hay evidencia de autocorrelación estacional *negativa*. (La inclusión de los regresores cambia las cosas sólo moderadamente: $\hat{\rho}_{12} = -.170$ y un valor-p = .052.) Esto es inusual y no tiene una explicación obvia.

Pregunta 12.3

Suponga que cuenta con datos trimestrales y que desea probar la presencia de correlación serial de primer o de cuarto orden. ¿Cómo procedería con regresores estrictamente exógenos?

12.3 Corrección de correlación serial con regresores estrictamente exógenos

Si se detecta correlación serial luego de aplicar una de las pruebas de la sección 12.2, se tiene que hacer algo al respecto. Si nuestro objetivo es estimar un modelo con la dinámica completa, es necesario especificarlo de nuevo. En aplicaciones en las que nuestro objetivo no es estimar un modelo dinámico completo es necesario que se determine una forma de hacer inferencia estadística: como se vio en la sección 12.1, los estadísticos de prueba de MCO usuales ya no son válidos. En esta sección se comienza con el caso importante de correlación serial AR(1). El método tradicional para este problema supone regresores fijos. Lo que se necesita en realidad son regresores estrictamente exógenos. Por consiguiente, no se debe utilizar estas correcciones cuando las variables explicativas comprendan variables dependientes rezagadas.

Obtención del mejor estimador lineal insesgado en el modelo AR(1)

En lo que sigue, se tienen como ciertos los supuestos de Gauss-Markov ST.1 a ST.4, pero se relaja el supuesto ST.5. En concreto, se supone que los errores siguen el modelo AR(1)

$$u_t = \rho u_{t-1} + e_t, \text{ para toda } t = 1, 2, \ldots. \qquad \textbf{12.26}$$

Recuerde que el supuesto ST.3 implica que u_t tiene media cero condicional en **X**. En el siguiente análisis se da por sentado el condicionamiento sobre **X** con el fin de simplificar la notación. Así, se escribe la varianza de u_t como

$$\mathrm{Var}(u_t) = \sigma_e^2/(1 - \rho^2). \tag{12.27}$$

Por simplicidad, considere el caso con sólo una variable explicativa:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \text{ para toda } t = 1, 2, \ldots, n.$$

Dado que el problema de esta ecuación es la correlación serial en u_t, tiene sentido transformarla para eliminar la correlación serial. Para $t \geq 2$, se escribe

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + u_{t-1}$$
$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

Ahora bien, si se multiplica la primera ecuación por ρ y se la resta de la segunda ecuación, se obtiene

$$y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + e_t, t \geq 2,$$

donde se ha usado el hecho de que $e_t = u_t - \rho u_{t-1}$. Se puede escribir esto como

$$\tilde{y}_t = (1 - \rho)\beta_0 + \beta_1 \tilde{x}_t + e_t, t \geq 2, \tag{12.28}$$

donde

$$\tilde{y}_t = y_t - \rho y_{t-1}, \tilde{x}_t = x_t - \rho x_{t-1} \tag{12.29}$$

son llamados los **datos cuasi diferenciados**. (Si $\rho = 1$, estos son los datos diferenciados, pero recuerde que se está suponiendo que $|\rho| < 1$.) Los términos de error en la ecuación (12.28) no están correlacionados serialmente; de hecho, esta ecuación satisface todos los supuestos de Gauss-Markov. Esto significa que si se conoce ρ, se puede estimar β_0 y β_1 al hacer la regresión de $\tilde{y}_t$ en $\tilde{x}_t$, siempre y cuando se divida el intercepto estimado entre $(1 - \rho)$.

Los estimadores de MCO de la ecuación (12.28) no son MELI porque no utilizan el primer periodo. Esto puede arreglarse con facilidad al escribir la ecuación para $t = 1$ como

$$y_1 = \beta_0 + \beta_1 x_1 + u_1. \tag{12.30}$$

Como cada e_t no está correlacionado con u_1, se puede añadir la ecuación (12.30) a la ecuación (12.28) y aún tener errores no correlacionados serialmente. Sin embargo, al usar la ecuación (12.27), se tiene que $\mathrm{Var}(u_1) = \sigma_e^2/(1 - \rho^2) > \sigma_e^2 = \mathrm{Var}(e_t)$. [La ecuación (12.27) desde luego no es válida cuando $|\rho| \geq 1$, que es la razón por la cual se supuso la condición de estabilidad.] Por ende, se debe multiplicar la ecuación (12.30) por $(1 - \rho^2)^{1/2}$ para obtener errores con la misma varianza:

$$(1 - \rho^2)^{1/2} y_1 = (1 - \rho^2)^{1/2}\beta_0 + \beta_1(1 - \rho^2)^{1/2} x_1 + (1 - \rho^2)^{1/2} u_1$$

o

$$\tilde{y}_1 = (1 - \rho^2)^{1/2}\beta_0 + \beta_1 \tilde{x}_1 + \tilde{u}_1, \tag{12.31}$$

donde $\tilde{u}_1 = (1 - \rho^2)^{1/2}u_1$, $\tilde{y}_1 = (1 - \rho^2)^{1/2}y_1$, y así sucesivamente. El error en la ecuación (12.31) tiene varianza $\mathrm{Var}(\tilde{u}_1) = (1 - \rho^2)\mathrm{Var}(u_1) = \sigma_e^2$, de manera que se puede usar la ecuación (12.31) junto con la ecuación (12.28) en una regresión de MCO. Esto da los estimadores MELI de β_0 y β_1 bajo los supuestos ST.1 a ST.4 y el modelo AR(1) para u_t. Este es otro ejemplo de un estimador de *mínimos cuadrados generalizados* (o MCG). En el capítulo 8 se vieron otros estimadores de MCG en el contexto de la heterocedasticidad.

La adición de más regresores genera muy pocos cambios. Para $t \geq 2$, se usa la ecuación

$$\tilde{y}_t = (1 - \rho)\beta_0 + \beta_1\tilde{x}_{t1} + \dots + \beta_k\tilde{x}_{tk} + e_t, \tag{12.32}$$

donde $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,\,j}$. Para $t = 1$, se tiene $\tilde{y}_1 = (1 - \rho^2)^{1/2}y_1$, $\tilde{x}_{1j} = (1 - \rho^2)^{1/2}x_{1j}$, y el intercepto es $(1 - \rho^2)^{1/2}\beta_0$. Para un ρ, dado, es muy fácil transformar los datos y llevar a cabo MCO. A menos que $\rho = 0$, el estimador MCG, es decir MCO sobre los datos transformados, por lo general será diferente del estimador MCO original. El estimador MCG resulta ser MELI, y dado que los errores en la ecuación transformada no están correlacionados serialmente y son homocedásticos, los estadísticos t y F de la ecuación transformada son válidos (cuando menos asintóticamente, y exactamente si los errores e_t están normalmente distribuidos).

Estimación por MCG factibles con errores AR(1)

El problema con el estimador de MCG es que ρ rara vez se conoce en la práctica. Sin embargo, se sabe cómo obtener un estimador consistente de ρ: tan solo se hace la regresión de los residuales de MCO sobre sus contrapartes rezagadas, precisamente como en la ecuación (12.14). A continuación se usa esta estimación, $\hat{\rho}$, en lugar de ρ para obtener las variables cuasi diferenciadas. Luego, se estima por MCO la ecuación

$$\tilde{y}_t = \beta_0\tilde{x}_{t0} + \beta_1\tilde{x}_{t1} + \dots + \beta_k\tilde{x}_{tk} + error_t, \tag{12.33}$$

donde $\tilde{x}_{t0} = (1 - \hat{\rho})$ para $t \geq 2$, y $\tilde{x}_{10} = (1 - \hat{\rho}^2)^{1/2}$. Esto da como resultado el estimador **de MCG factibles (MCGF)** de las β_j. El término de error en la ecuación (12.33) contiene e_t y también los términos que involucran el error de estimación en $\hat{\rho}$. Por fortuna, el error de estimación en $\hat{\rho}$ no afecta la distribución asintótica de los estimadores de MCGF.

Estimación por MCG factibles del modelo AR(1):

i) Realice la regresión por MCO de y_t sobre $x_{t1}, \dots, x_{tk}$ y obtenga los residuales de MCO, $\hat{u}_t$, $t = 1, 2, \dots, n$.

ii) Efectúe la regresión de la ecuación (12.14) y obtenga $\hat{\rho}$.

iii) Aplique MCO a la ecuación (12.33) para estimar $\beta_0, \beta_1, \dots, \beta_k$. Los errores estándar usuales, los estadísticos t y los estadísticos F son asintóticamente válidos.

El costo de usar $\hat{\rho}$ en lugar de ρ es que el estimador de MCG factibles no tiene propiedades de muestra finita manejables. En particular, no es insesgado, aunque es consistente cuando los datos son débilmente dependientes. Además, incluso si en la ecuación (12.32) e_t está distribuido normalmente, los estadísticos t y F sólo tienen una distribución t y F aproximada, debido al error de estimación en $\hat{\rho}$. Esto es adecuado para la mayoría de los propósitos, aunque se debe tener cuidado con los tamaños de muestra pequeños.

Como el estimador de MCGF no es insesgado, desde luego que no se puede decir que es MELI. Sin embargo, es asintóticamente más eficiente que el estimador de MCO cuando el modelo AR(1) para la correlación serial es válido (y las variables explicativas son estrictamente exógenas). De nuevo, este enunciado da por sentado que las series de tiempo son débilmente dependientes.

Existen varios nombres para la estimación por MCGF del modelo AR(1) que provienen de distintos métodos de estimación de ρ y de un tratamiento diferente de la primera observación. La **estimación Cochrane-Orcutt (CO)** omite la primera observación y utiliza $\hat{\rho}$ de (12.14), mientras que la **estimación Prais-Winsten (PW)** utiliza la primera observación de la manera sugerida antes. De forma asintótica, no existe diferencia si se usa la primera observación o no, pero muchas muestras de series de tiempo son pequeñas, así que las diferencias en las aplicaciones pueden ser notorias.

En la práctica, tanto el método de Cochrane-Orcutt como el de Prais-Winsten se utilizan en un esquema iterativo. Una vez que se halla el estimador de MCGF usando $\hat{\rho}$ de (12.14), se puede calcular un nuevo conjunto de residuales, obtener un nuevo estimador de ρ de (12.14), transformar los datos usando la nueva estimación de ρ y estimar la ecuación (12.33) por MCO. Se puede repetir el proceso completo muchas veces, hasta que la estimación de ρ cambie muy poco respecto a la iteración anterior. Muchos paquetes de regresión implementan un procedimiento iterativo de forma automática, de manera que no se tiene que realizar ningún trabajo adicional. Resulta difícil decir si sirve hacer más de una iteración. Parece de utilidad en algunos casos pero, en teoría, las propiedades de muestra grande del estimador iterado son las mismas que las del estimador que aplica sólo la primera iteración. Para obtener detalles sobre éstos y otros métodos, vea Davidson y MacKinnon (1993, capítulo 10).

Ejemplo 12.4

[Estimación de Prais-Winsten en el estudio de evento]

Se estima la ecuación del ejemplo 10.5, aplicando la estimación de Prais-Winsten. Para comparar, también se presentan los resultados de MCO en la tabla 12.1.

Los coeficientes que son estadísticamente significativos en la estimación de Prais-Winsten no difieren mucho de los estimadores de MCO [en particular, los coeficientes de log(*chempi*), log(*rtwex*) y *afdec6*]. No debe sorprendernos que los coeficientes que son estadísticamente insignificantes cambien, quizá de manera marcada, de un método de estimación a otro.

Advierta cómo los errores estándar de la segunda columna son más grandes de manera uniforme que los de la primera. Esto es común. Los errores estándar de Prais-Winsten consideran la correlación serial; los errores estándar de MCO no. Como se vio en la sección 12.1, por lo común estos últimos subestiman la variación de muestreo real de las estimaciones por MCO y no debe confiarse en ellos cuando una correlación serial significativa está presente. Por tanto, el efecto sobre las importaciones chinas después de la decisión de la Comisión de Comercio Internacional ahora es menos significativo estadísticamente de lo que se pensaba ($t_{afdec6} = -1.69$).

Por último, se reporta una R-cuadrada para la estimación de PW, lo que en este caso está muy por debajo de la R-cuadrada para la estimación de MCO. Sin embargo, estas R-cuadradas no deben compararse. Para MCO, la R-cuadrada, como de costumbre, se basa en la regresión con las variables dependientes e independientes no transformadas. Para PW, la R-cuadrada proviene de la regresión final de la variable

TABLA 12.1

Variable dependiente: log(*chnimp*)

Coeficiente	MCO	Prais-Winsten
log(*chempi*)	3.12 (0.48)	2.94 (0.63)
log(*gas*)	.196 (.907)	1.05 (0.98)
log(*rtwex*)	.983 (.400)	1.13 (0.51)
befile6	.060 (.261)	−.016 (.322)
affile6	−.032 (.264)	−.033 (.322)
afdec6	−.565 (.286)	−.577 (.342)
intercepto	−17.80 (21.05)	−37.08 (22.78)
$\hat{\rho}$	———	.293
Observaciones *R*-cuadrada	131 .305	131 .202

dependiente *transformada* sobre las variables independientes transformadas. No está claro lo que mide esta R^2 en realidad, aunque se reporta de la manera usual.

Comparación de MCO y MCGF

En algunas aplicaciones de los métodos Cochrane-Orcutt o Prais-Winsten, las estimaciones por MCGF difieren, en la práctica, de manera importante de las estimaciones por MCO (éste no es el caso del ejemplo 12.4). Por lo general, esto se ha interpretado como una prueba de superioridad de los estimadores de MCG factibles sobre los estimadores de MCO. Por desgracia, las cosas no son tan simples. Para ver por qué, considere el modelo de regresión

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

donde los procesos de series de tiempo son estacionarios. Ahora, suponiendo que la ley de los grandes números es válida, la consistencia de MCO para β_1 se mantiene si

$$\mathrm{Cov}(x_t, u_t) = 0. \tag{12.34}$$

Con anterioridad, se afirmó que los estimadores de MCGF eran consistentes de acuerdo con el supuesto de exogeneidad estricta, que es más restrictivo que la ecuación (12.34). De hecho, puede demostrarse que el supuesto más débil que debe mantenerse para que los estimadores de MCGF sean consistentes, *además de* (12.34), es que la suma de x_{t-1} y x_{t+1} no se correlacione con u_t:

$$\mathrm{Cov}[(x_{t-1} + x_{t+1}), u_t] = 0. \tag{12.35}$$

En términos prácticos, la consistencia de MCGF exige que u_t no se correlacione con x_{t-1}, x_t, ni x_{t+1}.

¿Cómo se puede mostrar que la condición (12.35) es necesaria junto con la condición (12.34)? El argumento es simple si se supone que se conoce ρ y se omite el primer periodo, como en Cochrane-Orcutt. El argumento cuando se usa $\hat{\rho}$ es más difícil y no produce un mejor discernimiento. Dado que una observación no puede afectar las propiedades asintóticas de un estimador, al omitirla no se ve afectado el argumento. Ahora bien, con ρ conocido, el estimador de MCG utiliza $x_t - \rho x_{t-1}$ como el regresor de una ecuación donde $u_t - \rho u_{t-1}$ es el error. Por el teorema 11.1 se sabe que la condición fundamental para la consistencia de MCO es que el error y el regresor no estén correlacionados. En este caso, se necesita que $\mathrm{E}[(x_t - \rho x_{t-1})(u_t - \rho u_{t-1})] = 0$. Si se expande la esperanza, se obtiene

$$\begin{aligned}\mathrm{E}[(x_t - \rho x_{t-1})(u_t - \rho u_{t-1})] &= \mathrm{E}(x_t u_t) - \rho \mathrm{E}(x_{t-1}u_t) - \rho \mathrm{E}(x_t u_{t-1}) + \rho^2 \mathrm{E}(x_{t-1}u_{t-1}) \\ &= -\rho[\mathrm{E}(x_{t-1}u_t) + \mathrm{E}(x_t u_{t-1})]\end{aligned}$$

debido a que $\mathrm{E}(x_t u_t) = \mathrm{E}(x_{t-1}u_{t-1}) = 0$ por el supuesto (12.34). Ahora, bajo la estacionariedad, $\mathrm{E}(x_t u_{t-1}) = \mathrm{E}(x_{t+1}u_t)$ debido a que sólo se está desplazando el índice de tiempo un periodo adelante. Por consiguiente,

$$\mathrm{E}(x_{t-1}u_t) + \mathrm{E}(x_t u_{t-1}) = \mathrm{E}[(x_{t-1} + x_{t+1})u_t],$$

y la última esperanza es la covarianza de la ecuación (12.35) debido a que $\mathrm{E}(u_t) = 0$. Se ha mostrado que la ecuación (12.35) es necesaria junto con la ecuación (12.34) para que MCG sean consistentes para β_1. (Desde luego, si $\rho = 0$, no se necesita la ecuación (12.35) debido a que se está de nuevo en MCO.)

Estos cálculos muestran que MCO y MCGF podrían dar estimaciones significativamente distintas si (12.35) no se satisface. En este caso, se prefieren MCO —que aún son consistentes bajo (12.34)— a MCGF (que son inconsistentes). Si x tiene un efecto rezagado sobre y, o x_{t+1} reacciona a los cambios en u_t, MCGF pueden generar resultados engañosos.

Como MCO y MCGF son procedimientos de estimación diferentes, no es de esperar que den los mismos resultados. Si proporcionan estimaciones similares de las β_j, entonces son preferibles los MCGF si hay evidencia de correlación serial, ya que el estimador es más eficiente y los estadísticos de prueba de MCGF son al menos válidos asintóticamente. Surge un problema más complejo cuando hay diferencias prácticas en las estimaciones por MCO y MCGF: resulta difícil determinar si tales diferencias son estadísticamente significativas. Puede emplearse el método general propuesto por Hausman (1978), pero rebasa el alcance de este libro.

El siguiente ejemplo proporciona un caso donde los MCO y MCGF son, en la práctica, diferentes de manera importante.

Ejemplo 12.5

[Curva estática de Phillips]

La tabla 12.2 presenta las estimaciones por MCO e iteradas de Prais-Winsten para la curva estática de Phillips del ejemplo 10.1, usando las observaciones hasta 1996.

TABLA 12.2

Variable dependiente: *inf*

Coeficiente	MCO	Prais-Winsten
unem	.468 (.289)	−.716 (.313)
intercepto	1.424 (1.719)	8.296 (2.231)
$\hat{\rho}$	———	.781
Observaciones *R*-cuadrada	49 .053	49 .136

El coeficiente de interés es el de *unem* y difiere de forma marcada entre PW y MCO. Como la estimación de PW es consistente para el intercambio entre inflación y desempleo, nuestra tendencia es concentrarnos en las estimaciones de PW. De hecho, estas estimaciones son bastante cercanas a lo que se obtiene al diferenciar primero tanto *inf* como *unem* (vea el ejercicio para computadora C11.4), lo que tiene sentido porque la cuasi diferenciación utilizada en PW con $\hat{\rho} = .781$ es similar a la primera diferencia. Puede ser que *inf* y *unem* no se relacionen en niveles, pero tienen una relación negativa en las primeras diferencias.

Ejemplos como la curva estática de Phillips plantean problemas difíciles para los investigadores empíricos. Por una parte, si se está realmente interesado en una relación estática, y si el desempleo y la inflación son procesos I(0), entonces MCO producen estimadores consistentes sin supuestos adicionales. Pero podría ocurrir que el desempleo, la inflación o ambos tengan raíces unitarias, en cuyo caso MCO no necesariamente tienen sus propiedades deseables habituales; en el capítulo 18 se analizará esto más a fondo. En el ejemplo 12.5, MCGF proporciona estimaciones económicamente más sensibles; debido a que es parecido a la regresión en primera diferencia, MCGF tiene la ventaja de eliminar (de manera aproximada) las raíces unitarias.

Corrección de la correlación serial de orden superior

También es posible corregir órdenes superiores de correlación serial. En Harvey (1990) se da un tratamiento general. Aquí, se ilustra el enfoque para la correlación serial AR(2):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t,$$

donde $\{e_t\}$ satisface los supuestos establecidos para el modelo AR(1). Ahora resultan más complejas las condiciones de estabilidad. Puede demostrarse que ellas son [vea Harvey (1990)]

$$\rho_2 > -1, \rho_2 - \rho_1 < 1 \text{ y } \rho_1 + \rho_2 < 1.$$

Por ejemplo, el modelo es estable si $\rho_1 = .8$ y $\rho_2 = -.3$; el modelo es inestable si $\rho_1 = .7$ y $\rho_2 = .4$.

Suponiendo que las condiciones de estabilidad son válidas, se puede obtener la transformación que elimina la correlación serial. En el modelo de regresión simple, esto es sencillo cuando $t > 2$:

$$y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} = \beta_0(1 - \rho_1 - \rho_2) + \beta_1(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + e_t$$

o

$$\tilde{y}_t = \beta_0(1 - \rho_1 - \rho_2) + \beta_1\tilde{x}_t + e_t, t = 3, 4, \ldots, n. \quad \boxed{12.36}$$

Si se conocen ρ_1 y ρ_2, se puede estimar con facilidad esta ecuación por MCO después de obtener las variables transformadas. Como pocas veces se conocen ρ_1 y ρ_2, hay que estimarlas. Como es costumbre, se utilizan los residuales de MCO, $\hat{u}_t$: se obtienen $\hat{\rho}_1$ y $\hat{\rho}_2$ de la regresión de

$$\hat{u}_t \text{ sobre } \hat{u}_{t-1}, \hat{u}_{t-2}, t = 3, \ldots, n.$$

[Esta es la misma regresión utilizada para probar la correlación serial AR(2) con regresores estrictamente exógenos.] Luego, se emplean $\hat{\rho}_1$ y $\hat{\rho}_2$ en lugar de ρ_1 y ρ_2 para obtener las variables transformadas. Esto da una versión del estimador de MCG factibles. Si se tienen múltiples variables explicativas, entonces cada una se transforma por $\tilde{x}_{tj} = x_{tj} - \hat{\rho}_1 x_{t-1,j} - \hat{\rho}_2 x_{t-2,j}$, cuando $t > 2$.

El tratamiento de las primeras dos observaciones es un poco difícil. Puede mostrarse que la variable dependiente y cada variable independiente (incluido el intercepto) deberían transformarse por

$$\tilde{z}_1 = \{(1 + \rho_2)[(1 - \rho_2)^2 - \rho_1^2]/(1 - \rho_2)\}^{1/2} z_1$$
$$\tilde{z}_2 = (1 - \rho_2^2)^{1/2} z_2 - [\rho_1(1 - \rho_1^2)^{1/2}/(1 - \rho_2)]z_1,$$

donde z_1 y z_2 denotan, ya sea una variable dependiente o una independiente en $t = 1$ y $t = 2$, respectivamente. No se deducirán estas transformaciones. En breve, éstas eliminan la correlación serial entre las primeras dos observaciones y hacen que las varianzas de su error sean iguales a σ_e^2.

Por fortuna, los paquetes econométricos orientados al análisis de series de tiempo estiman con facilidad modelos de análisis de series de tiempo con errores generales AR(q) pocas veces es necesario que uno mismo calcule de forma directa las variables transformadas.

12.4 Diferenciación y correlación serial

En el capítulo 11 se presentó la diferenciación como una transformación para hacer que un proceso integrado sea débilmente dependiente. Hay otra manera de ver los méritos de la diferenciación cuando se manejan datos altamente persistentes. Suponga que se empieza con el modelo de regresión simple:

$$y_t = \beta_0 + \beta_1 x_t + u_t, t = 1, 2, \ldots, \quad \boxed{12.37}$$

donde u_t sigue el proceso AR(1) de (12.26). Como se mencionó en la sección 11.3 y como se analizará en forma más detallada en el capítulo 18, los procedimientos usuales de inferencia de

MCO pueden ser muy engañosos cuando las variables y_t y x_t son integradas de primer orden, o I(1). En el caso extremo en que los errores $\{u_t\}$ en (12.37) siguen una caminata aleatoria, la ecuación no tiene sentido porque, entre otras cosas, la varianza de u_t crece con t. Es más lógico diferenciar la ecuación:

$$\Delta y_t = \beta_1 \Delta x_t + \Delta u_t,\ t = 2, \ldots, n. \tag{12.38}$$

Si u_t sigue una caminata aleatoria, entonces $e_t \equiv \Delta u_t$ tiene media cero, varianza constante y no se correlaciona serialmente. Así, suponiendo que e_t y Δx_t no se correlacionan, se puede estimar la ecuación (12.38) por MCO, donde se pierde la primera observación.

Incluso si u_t no sigue una caminata aleatoria pero ρ es positiva y grande, a menudo es buena idea tomar la primera diferencia, ya que esto eliminará la mayor parte de la correlación serial. Desde luego, la ecuación (12.38) es diferente de la (12.37), pero al menos se puede tener más confianza en los errores estándar de MCO y en los estadísticos t de (12.38). La consideración de múltiples variables explicativas no modifica nada.

Ejemplo 12.6

[Diferenciación de la ecuación de la tasa de interés]

En el ejemplo 10.2, se estimó una ecuación que relacionaba la tasa de las letras del Tesoro a tres meses con la inflación y el déficit federal [vea la ecuación (10.15)]. Si se estima la regresión de los residuales de esta ecuación sobre un solo rezago, se obtiene $\hat{\rho} = .623$ (.110), que es un valor grande y estadísticamente significativo. Por tanto, como mínimo, la correlación serial es un problema en esta ecuación.

Si se diferencian los datos y se hace la regresión, se obtiene

$$\begin{aligned} \Delta i3_t &= \underset{(.171)}{.042} + \underset{(.092)}{.149}\ \Delta inf_t - \underset{(.148)}{.181}\ \Delta def_t + \hat{e}_t \\ n &= 55,\ R^2 = .176,\ \bar{R}^2 = .145 \end{aligned} \tag{12.39}$$

Los coeficientes de esta regresión son muy diferentes de aquellos de la ecuación en niveles, lo que sugiere ya sea que las variables explicativas no son estrictamente exógenas o que una o más variables tienen una raíz unitaria. De hecho, la correlación entre $i3_t$ e $i3_{t-1}$ es cercana a .885, lo cual puede indicar un problema al tratar de considerar a (10.15) como una regresión con sentido. Además, la regresión en diferencias en esencia no tiene correlación serial: una regresión de $\hat{e}_t$ sobre $\hat{e}_{t-1}$ da $\hat{\rho} = .072$ (.134). Dado que obtener la primera diferencia elimina posibles raíces unitarias así como correlación serial, probablemente se tiene más confianza en las estimaciones y los errores estándar de (12.39) que en los de (10.15). La ecuación en diferencias muestra que los cambios anuales en las tasas de interés sólo tienen una relación positiva y débil con los cambios anuales en la inflación, y el coeficiente de Δdef_t es de hecho negativo (aun cuando no sea estadísticamente significativo, incluso en el nivel de significancia de 20% contra una alternativa de dos colas).

Como se explicó en el capítulo 11, la decisión de hacer o no una diferenciación es difícil. Pero este análisis señala otro beneficio de la diferenciación, que es la eliminación de la correlación serial. Se revisará de nuevo este tema en el capítulo 18.

Pregunta 12.4

Suponga que después de estimar un modelo por MCO, usted estima ρ de la regresión (12.14) y obtiene $\hat{\rho} = .92$. ¿Qué haría al respecto?

12.5 Inferencia robusta a la correlación serial después de MCO

En años recientes, se ha vuelto muy popular la estimación de modelos por MCO, pero sin corregir los errores estándar ante formas muy arbitrarias de correlación serial (y heterocedasticidad). Aun cuando se sabe que MCO será ineficiente, existen algunas buenas razones para optar por este método. Primero, las variables explicativas tal vez no sean estrictamente exógenas. En este caso, MCGF ni siquiera es consistente, mucho menos eficiente. Segundo, en la mayoría de las aplicaciones de MCGF, se supone que los errores siguen un modelo AR(1). Quizá sea mejor calcular los errores estándar para las estimaciones de MCO que sean robustos ante formas más generales de correlación serial.

Para entender esta idea, considere la ecuación (12.4), que es la varianza del estimador de pendiente de MCO en un modelo de regresión simple con errores AR(1). Se puede estimar esta varianza de manera muy sencilla al sustituir los estimadores estándar de ρ y σ^2. Los únicos problemas con esto son que supone que el modelo AR(1) es válido y también da por sentada la homocedasticidad. Es posible relajar ambas suposiciones.

Un tratamiento general de los errores estándar que son robustos a la heterocedasticidad y a la correlación serial se proporciona en Davidson y MacKinnon (1993). Aquí se expone un método sencillo para calcular el error estándar robusto de cualquier coeficiente de MCO.

Nuestro tratamiento aquí sigue a Wooldridge (1989). Considere el modelo estándar de regresión lineal múltiple

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \; t = 1, 2, \ldots, n, \tag{12.40}$$

que se ha estimado por MCO. En concreto, se está interesado en obtener un error estándar robusto a la correlación serial para $\hat{\beta}_1$. Esto resulta ser muy fácil. Escriba x_{t1} como una función lineal de las variables independientes restantes y un término de error

$$x_{t1} = \delta_0 + \delta_2 x_{t2} + \ldots + \delta_k x_{tk} + r_t,$$

donde el error r_t tenga media cero y no esté serialmente correlacionado con $x_{t2}, x_{t3}, \ldots, x_{tk}$.

Por tanto, se puede mostrar que la varianza asintótica del estimador $\hat{\beta}_1$ de MCO es

$$\text{Avar}(\hat{\beta}_1) = \left(\sum_{t=1}^{n} \text{E}(r_t^2)\right)^{-2} \text{Var}\left(\sum_{t=1}^{n} r_t u_t\right).$$

Bajo el supuesto de no correlación serial ST.5′, $\{a_t \equiv r_t u_t\}$ no está serialmente correlacionado, así que los errores estándar de MCO (bajo la homocedasticidad) o los errores estándar robustos a la heterocedasticidad serán válidos. Pero si el supuesto ST.5′ no se cumple, nuestra expresión para $\text{Avar}(\hat{\beta}_1)$ debe tomar en cuenta la correlación entre a_t y a_s, cuando $t \neq s$. En la práctica, es común suponer que, una vez que los términos estén alejados por más de unos cuantos periodos, la correlación es fundamentalmente cero. Recuerde que bajo la dependencia débil, la correlación debe aproximarse a cero, de modo que este es un tratamiento razonable.

Siguiendo el esquema de trabajo general de Newey y West (1987), Wooldridge (1989) muestra que $\text{Avar}(\hat{\beta}_1)$ puede estimarse como sigue: "ee($\hat{\beta}_1$)" denota el error estándar de MCO usual (pero incorrecto) y $\hat{\sigma}$ indica el error estándar usual de la regresión (o la raíz del error cuadrático medio) obtenidos de la estimación de la ecuación (12.40) por MCO. Sean $\hat{r}_t$ los residuales de la regresión auxiliar de

$$x_{t1} \text{ sobre } x_{t2}, x_{t3}, \ldots, x_{tk} \tag{12.41}$$

(incluyendo una constante, como de costumbre). Para un entero a elegir $g > 0$, defina

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + 2\sum_{h=1}^{g} [1 - h/(g+1)] \left(\sum_{t=h+1}^{n} \hat{a}_t \hat{a}_{t-h} \right), \tag{12.42}$$

donde

$$\hat{a}_t = \hat{r}_t \hat{u}_t, \; t = 1, 2, \ldots, n.$$

Esto parece un poco complicado, pero en la práctica es fácil de obtener. El entero g en la ecuación (12.42) controla cuánta correlación serial se está permitiendo en el cálculo del error estándar. Una vez que se tiene $\hat{v}$, el **error estándar robusto a la correlación serial** de $\hat{\beta}_1$ es simplemente

$$\text{ee}(\hat{\beta}_1) = [\text{“ee}(\hat{\beta}_1)\text{”}/\hat{\sigma}]^2 \sqrt{\hat{v}}. \tag{12.43}$$

En otras palabras, se toma el error estándar usual de MCO de $\hat{\beta}_1$ y se divide entre $\hat{\sigma}$, se eleva al cuadrado el resultado, y luego se multiplica por la raíz cuadrada de $\hat{v}$. Esto puede usarse para construir intervalos de confianza y estadísticos t para $\hat{\beta}_1$.

Es conveniente ver cómo es $\hat{v}$ en algunos casos simples. Cuando $g = 1$,

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + \sum_{t=2}^{n} \hat{a}_t \hat{a}_{t-1}, \tag{12.44}$$

y cuando $g = 2$,

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + (4/3)\left(\sum_{t=2}^{n} \hat{a}_t \hat{a}_{t-1}\right) + (2/3)\left(\sum_{t=3}^{n} \hat{a}_t \hat{a}_{t-2}\right). \tag{12.45}$$

Cuanto más grande sea g, más términos se incluyen para corregir la correlación serial. El propósito del factor $[1 - h/(g + 1)]$ en la ecuación (12.42) es asegurar que $\hat{v}$ de hecho sea no negativo [Newey y West (1987) comprueban esto]. Desde luego, es necesario que $\hat{v} \geq 0$, ya que es la estimación de una varianza y la raíz cuadrada de $\hat{v}$ aparece en la ecuación (12.43).

El error estándar en (12.43) también es robusto a la heterocedasticidad arbitraria. [En el material publicado sobre series de tiempo, a los errores estándar robustos a la correlación serial en ocasiones se les llama errores estándar *consistentes a la heterocedasticidad y a la autocorrelación*, o HAC (por sus siglas en inglés).] De hecho, si se omite el segundo término en la ecuación (12.42), entonces la ecuación (12.43) se vuelve el error estándar usual robusto a la heterocedasticidad que se estudió en el capítulo 8 (sin el ajuste de grados de libertad).

La teoría que subyace en el error estándar de la ecuación (12.43) es técnica y algo sutil. Recuerde que se empezó por afirmar que no se conoce la forma de la correlación serial. De ser éste el caso, ¿cómo se puede elegir el entero g? La teoría sostiene que la ecuación (12.43) funciona en el caso de formas de correlación serial bastante arbitrarias, siempre que g crezca con el tamaño de muestra n. La idea es que, con tamaños de muestra mayores se puede ser más flexible en cuanto a la cantidad de correlación en (12.42). Recientemente se han publicado muchos trabajos más sobre la relación entre g y n, pero no se abundará en ellos aquí. Para datos anuales, la elección de una g, pequeña, como $g = 1$ o $g = 2$, probablemente represente la mayor parte de la correlación serial. Para datos trimestrales o mensuales, probablemente g tenga que ser mayor (como $g = 4$ u 8 para los datos trimestrales y $g = 12$ o 24 para los mensuales), suponiendo que se cuenta con datos suficientes. Newey y West (1987) recomiendan tomar g como la parte entera de $4(n/100)^{2/9}$; otros han propuesto la parte entera de $n^{1/4}$. El programa econométrico Eviews® implementa la propuesta de Newey-West. Por ejemplo, para $n = 50$ (lo cual es razonable para datos anuales posteriores a la Segunda Guerra Mundial), $g = 3$. (La parte entera de $n^{1/4}$ resulta $g = 2$.)

Se resume la manera de obtener un error estándar robusto a la correlación serial para $\hat{\beta}_1$. Desde luego, dado que se puede listar cualquier variable independiente primero, el procedimiento siguiente funciona al calcular un error estándar para cualquier coeficiente de pendiente.

Error estándar robusto a la correlación serial para $\hat{\beta}_1$:

i) Estime la ecuación (12.40) por MCO, lo que genera "ee($\hat{\beta}_1$)", $\hat{\sigma}$, y los residuales de MCO $\{\hat{u}_t: t = 1, \ldots, n\}$.

ii) Calcule los residuales $\{\hat{r}_t: t = 1, \ldots, n\}$ de la regresión auxiliar (12.41). Luego, calcule $\hat{a}_t = \hat{r}_t\hat{u}_t$ (para cada t).

iii) Para su g elegida, calcule $\hat{v}$ como en la ecuación (12.42).

iv) Determine ee($\hat{\beta}_1$) usando la ecuación (12.43).

De forma empírica, los errores estándar robustos a la correlación serial por lo general son mayores que los errores estándar usuales de MCO cuando hay correlación serial. Esto se debe a que en la mayoría de los casos, los errores se correlacionan serialmente de forma positiva. Sin embargo, es posible contar con una alta correlación serial en $\{u_t\}$ y tener al mismo tiempo semejanzas entre los errores estándar usuales y los errores estándar robustos a la correlación serial (CS) de algunos coeficientes: son las autocorrelaciones muestrales de $\hat{a}_t = \hat{r}_t\hat{u}_t$ las que determinan que el error estándar sea robusto para $\hat{\beta}_1$.

El uso de errores estándar robustos a la CS ha quedado rezagado respecto al uso de los errores estándar robustos sólo a la heterocedasticidad, por varias razones. En primer lugar, los cortes transversales grandes, donde los errores estándar robustos a la heterocedasticidad tienen buenas propiedades, son más comunes que las series de tiempo grandes. Los errores estándar robustos a la CS no se desempeñan muy bien cuando hay una alta correlación serial y el tamaño muestral es pequeño (por pequeño se entiende un tamaño incluso de 100). En segundo lugar, como se debe elegir el entero g en la ecuación (12.42), el cálculo de los errores estándar robustos a la CS no es automático. Como ya se mencionó, algunos paquetes de econometría han automatizado la elección, pero aún así uno tiene que aceptar la elección.

Otra razón importante de que los errores estándar robustos a la CS no se calculen en forma rutinaria es que, en presencia de correlación serial importante, MCO pueden ser muy ineficientes, en especial en tamaños de muestra pequeños. Luego de aplicar MCO y corregir los errores estándar de la correlación serial, los coeficientes a menudo son insignificantes, o menos significativos de lo que fueron con los errores estándar de MCO usuales.

Si se tiene confianza en que las variables explicativas son estrictamente exógenas, pero no se cree que los errores sigan un proceso AR(1), aún se pueden obtener estimadores más eficientes que los de MCO al usar un estimador estándar de MCG factibles, tal como Prais-Winsten o Cochrane-Orcutt. Con una correlación serial alta, es probable que sea mejor la transformación cuasi diferenciada utilizada por PW o CO que no hacer nada y sólo usar MCO. Pero si los errores no siguen un modelo AR(1), entonces los errores estándar reportados por la estimación de PW o CO serán incorrectos. No obstante, se puede cuasi diferenciar manualmente los datos después de estimar ρ, usar MCO en los datos transformados y luego usar los errores estándar robustos a la CS en la ecuación transformada. El cálculo de un error estándar robusto a la CS después de la cuasi diferenciación asegurará que cualquier correlación serial adicional sea considerada en la inferencia estadística. En realidad, es probable que los errores estándar robustos a la CS funcionen mejor después de que gran parte de la correlación serial se ha eliminado usando la cuasi diferenciación [o cualquier otra transformación, como la usada en la correlación serial AR(2)]. Este método es análogo al uso de mínimos cuadrados ponderados en presencia de heterocedasticidad, donde luego se calculan los errores estándar que son robustos para tener la función de varianza especificada incorrectamente (vea la sección 8.4).

Los errores estándar robustos a la CS, luego de realizar la estimación por MCO, son más útiles cuando se tienen dudas acerca de la exogeneidad estricta de algunas de las variables explicativas, de modo que métodos como los de Prais-Winsten y Cochrane-Orcutt ni siquiera son consistentes. También es válido utilizar los errores estándar robustos a la CS en modelos con variables dependientes rezagadas, suponiendo, desde luego, que hay una buena razón para permitir la correlación serial en tales modelos.

Ejemplo 12.7

[El salario mínimo en Puerto Rico]

Se obtiene un error estándar robusto a la CS para el efecto del salario mínimo en la ecuación de empleo de Puerto Rico. En el ejemplo 12.2 se encontró una fuerte evidencia de correlación serial AR(1). Al igual que en ese ejemplo, se utilizan como controles adicionales log(*usgnp*), log(*prgnp*) y una tendencia lineal en el tiempo.

El estimador de MCO de la elasticidad de la tasa de empleo respecto al salario mínimo es $\hat{\beta}_1 = -.2123$, y el error estándar usual de MCO es "ee($\hat{\beta}_1$)" $= .0402$. El error estándar de la regresión es $\hat{\sigma} = .0328$. Además, al utilizar el procedimiento anterior con $g = 2$ [vea la ecuación (12.45)], se obtiene $\hat{v} = .000805$. Esto da el error estándar robusto a la CS/heterocedasticidad de $\text{ee}(\hat{\beta}_1) = [(.0402/.0328)^2]\sqrt{.000805} \approx .0426$. Curiosamente, el error estándar robusto es apenas ligeramente mayor que el usual de MCO. El estadístico t robusto es cercano a -4.98, y por ende la elasticidad estimada aún es muy significativa en términos estadísticos.

En comparación, el estimador iterado de PW de β_1 es $-.1477$, con un error estándar de .0458. Así, la estimación por MCGF está mucho más próxima a cero que la estimación por MCO, y se podría sospechar que hay una violación al supuesto de exogeneidad estricta. O bien, la diferencia entre las estimaciones por MCO y por MCGF se podría explicar por error de muestreo. Resulta muy difícil determinarlo.

Kiefer y Vogelsang (2005) proporcionan una manera diferente de hacer inferencia válida en presencia de correlación serial arbitraria. En vez de preocuparse por la tasa a la cual se permite crecer a g (como función de n) con el fin de que el estadístico t tenga distribuciones normales estándar asintóticas, Kiefer y Vogelsang derivan la distribución en muestras grandes del estadístico t cuando se permite que $b = (g + 1)/n$ se establezca como una fracción diferente de cero. [En el escenario de Newey-West, $(g + 1)/n$ siempre converge a cero.] Por ejemplo, cuando $b = 1$, $g = n - 1$, lo cual significa que se incluye *cada* término de covarianza en la ecuación (12.42). El estadístico t resultante no tiene una distribución normal estándar en muestras grandes, pero Kiefer y Vogelsang muestran que sí tiene una distribución asintótica y tabulan los valores críticos apropiados. Para una prueba de dos colas de 5%, el valor crítico es 4.771, y para una prueba de dos colas a un nivel de 10%, el valor crítico es 3.764. En comparación con los valores críticos de la distribución normal estándar, se necesita un estadístico t considerablemente más grande. Pero no hay que preocuparse por elegir el número de covarianzas en la ecuación (12.42).

Se concluye esta sección con la observación de que es posible construir estadísticos tipo F robustos a la correlación serial para probar hipótesis múltiples, pero es un tema demasiado avanzado para cubrirlo aquí [vea Wooldridge (1991b, 1995) y Davidson y MacKinnon (1993) para los tratamientos].

12.6 Heterocedasticidad en regresiones de series de tiempo

En el capítulo 8 se analizaron las pruebas y la corrección de la heterocedasticidad para aplicaciones de corte transversal. La heterocedasticidad se presenta también en los modelos de regresión de series de tiempo, y su presencia, mientras no ocasiona sesgo ni inconsistencia en las $\hat{\beta}_j$, invalida los errores estándar usuales y los estadísticos t y F, al igual que en el caso de corte transversal.

En las aplicaciones de regresión de series de tiempo, la heterocedasticidad a menudo recibe poca atención, si es que la recibe: por lo regular es más persistente el problema de los errores serialmente correlacionados. No obstante, resultará de ayuda cubrir de manera breve algunos problemas que surgen al aplicar las pruebas y las formas de corregir la heterocedasticidad en las regresiones de series de tiempo.

Como los estadísticos usuales de MCO son válidos asintóticamente de acuerdo con los supuestos ST.1′ a ST.5′, es interesante ver lo que sucede cuando el supuesto de homocedasticidad ST.4′ no se cumple. El supuesto ST.3′ descarta especificaciones incorrectas, tales como variables omitidas y ciertos tipos de error de medición, mientras que ST.5′ excluye la correlación serial en los errores. Resulta importante recordar que los errores serialmente correlacionados generan problemas que no son capaces de resolver las pruebas y los ajustes de la heterocedasticidad.

Estadísticos robustos a la heterocedasticidad

Al estudiar la heterocedasticidad para las regresiones de corte transversal, se observó que no guarda ninguna relación con la ausencia de insesgamiento o consistencia de los estimadores de MCO. Se sostienen exactamente las mismas conclusiones en el caso de las series de tiempo, como se ve al revisar los supuestos necesarios para el insesgamiento (teorema 10.1) y la consistencia (teorema 11.1).

En la sección 8.2 se analizó cómo se ajustan los errores estándar usuales de MCO y los estadísticos t y F al considerar la presencia de heterocedasticidad de forma desconocida. Estos mismos ajustes se aplican a las regresiones de series de tiempo bajo los supuestos ST.1′, ST.2′, ST.3′ y ST.5′. Así pues, considerando que el único supuesto violado es el de homocedasticidad, la inferencia válida se obtiene con facilidad en la mayor parte de los paquetes de econometría.

Pruebas de heterocedasticidad

A veces, se desea probar la heterocedasticidad en las regresiones de series de tiempo, en especial si nos preocupa el desempeño de los estadísticos robustos a la heterocedasticidad en tamaños de muestra relativamente pequeños. Las pruebas que se revisaron en el capítulo 8 se aplican de forma directa, pero con unas cuantas salvedades. En primer lugar, los errores u_t no deben correlacionarse serialmente; cualquier correlación serial por lo general invalida las pruebas de heterocedasticidad. De ahí que tenga sentido probar primero la correlación serial, con ayuda de una prueba robusta a la heterocedasticidad, si se sospecha de la existencia de ésta. Una vez que se han realizado las acciones pertinentes para corregir la correlación serial, se pone a prueba la heterocedasticidad.

En segundo lugar, considere la ecuación utilizada para motivar la prueba de heterocedasticidad de Breusch-Pagan:

$$u_t^2 = \delta_0 + \delta_1 x_{t1} + \dots + \delta_k x_{tk} + v_t, \tag{12.46}$$

donde la hipótesis nula es H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$. Para que sea válido el estadístico F, con $\hat{u}_t^2$ reemplazando a u_t^2 como variable dependiente, se debe suponer que los errores $\{v_t\}$ son en sí homocedásticos (como en el caso de corte transversal) y que no están serialmente correlacionados. Esto se da por sentado de manera implícita al calcular todas las pruebas estándar de heterocedasticidad, incluida la versión de la prueba de White que se estudió en la sección 8.3. La suposición de que las $\{v_t\}$ no están serialmente correlacionadas excluye ciertas formas de heterocedasticidad dinámica, tema que se tratará en la siguiente subsección.

Pregunta 12.5

¿Cómo calcularía la prueba de heterocedasticidad de White en la ecuación (12.47)?

Si se encuentra heterocedasticidad en u_t (y las u_t no se correlacionan serialmente), entonces se pueden utilizar los estadísticos de prueba robustos a la heterocedasticidad. Una alternativa es emplear **mínimos cuadrados ponderados**, como en la sección 8.4. La mecánica de los mínimos cuadrados ponderados para el caso de series de tiempo es idéntica a aquélla para el caso de corte transversal.

Ejemplo 12.8

[Heterocedasticidad y la hipótesis de los mercados eficientes]

En el ejemplo 11.4 se estimó el modelo simple

$$return_t = \beta_0 + \beta_1 return_{t-1} + u_t. \qquad \textbf{12.47}$$

La HME establece que $\beta_1 = 0$. Cuando se probó esta hipótesis con los datos de NYSE.RAW, se obtuvo $t_{\beta_1} = 1.55$ con $n = 689$. Con una muestra tan grande, ésta no es una evidencia muy contundente en contra de la HME. Aun cuando la HME plantea que el rendimiento esperado, dada la información observada en el pasado, debe ser constante, no indica nada sobre la varianza condicional. De hecho, la prueba de heterocedasticidad de Breusch-Pagan implica hacer la regresión de los residuales cuadrados de MCO $\hat{u}_t^2$ sobre $return_{t-1}$:

$$\begin{aligned} \hat{u}_t^2 &= 4.66 - 1.104\, return_{t-1} + residual_t \\ &\quad\ (0.43)\quad (0.201) \\ n &= 689,\ R^2 = .042. \end{aligned} \qquad \textbf{12.48}$$

El estadístico t de $return_{t-1}$ es cercano a -5.5, lo que señala una fuerte evidencia de heterocedasticidad. Como el coeficiente de $return_{t-1}$ es negativo, se tiene el interesante hallazgo de que la volatilidad del rendimiento de las acciones es menor cuando el rendimiento anterior fue alto y viceversa. Por tanto, se ha descubierto una característica común en muchos estudios financieros: el valor esperado de los rendimientos de las acciones no depende de los rendimientos anteriores, pero la varianza de éstos sí.

Heterocedasticidad condicional autorregresiva

En años recientes, los economistas han volcado su interés en las formas dinámicas de heterocedasticidad. Desde luego, si $\mathbf{x}_t$ contiene una variable dependiente rezagada, entonces la heterocedasticidad dada en la ecuación (12.46) es dinámica. Pero las formas dinámicas de heterocedasticidad aparecen incluso en modelos que no son dinámicos en la ecuación de regresión.

Para verlo, considere un modelo de regresión estática simple:

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

y suponga que los supuestos de Gauss-Markov son válidos. Esto quiere decir que los estimadores de MCO son MELI. El supuesto de homocedasticidad establece que $\text{Var}(u_t|\mathbf{Z})$ es constante, donde $\mathbf{Z}$ denota los n resultados de z_t. Aun cuando la varianza de u_t dada $\mathbf{Z}$ es constante, la heterocedasticidad puede surgir de otras formas. Engle (1982) propuso que se considerara la varianza condicional de u_t, dados los errores pasados (en los que el condicionamiento sobre $\mathbf{Z}$ se deja implícito). Engle sugirió lo que se conoce como modelo de **heterocedasticidad condicional autorregresiva (ARCH**, por sus siglas en inglés). El modelo ARCH de primer orden es

$$\text{E}(u_t^2|u_{t-1}, u_{t-2}, \ldots) = \text{E}(u_t^2|u_{t-1}) = \alpha_0 + \alpha_1 u_{t-1}^2, \quad \boxed{\mathbf{12.49}}$$

donde queda implícito el condicionamiento sobre $\mathbf{Z}$. Esta ecuación representa la varianza condicional de u_t dada la u_t anterior sólo si $\text{E}(u_t|u_{t-1}, u_{t-2}, \ldots) = 0$, lo cual significa que los errores no están correlacionados serialmente. Como las varianzas condicionales deben ser positivas, este modelo sólo tiene sentido si $\alpha_0 > 0$ y $\alpha_1 \geq 0$; si $\alpha_1 = 0$, no hay dinámica en la ecuación de la varianza.

Es instructivo escribir la ecuación (12.49) como sigue

$$u_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + v_t, \quad \boxed{\mathbf{12.50}}$$

donde el valor esperado de v_t (dadas $u_{t-1}, u_{t-2}, \ldots$) es cero por definición. (Sin embargo, las v_t no son independientes de las u_t pasadas debido a la restricción $v_t \geq -\alpha_0 - \alpha_1 u_{t-1}^2$.) La ecuación (12.50) parece un modelo autorregresivo en u_t^2 (de ahí el nombre ARCH). La condición de estabilidad para esta ecuación es $\alpha_1 < 1$, como en el modelo AR(1) usual. Cuando $\alpha_1 > 0$, los errores cuadrados contienen correlación serial (positiva) aun cuando las mismas u_t no la tengan.

¿Qué implicaciones tiene la ecuación (12.50) para MCO? Ya que se comenzó por suponer que los supuestos de Gauss-Markov son válidos, los estimadores de MCO son MELI. Además, aun si u_t no estuviera distribuida normalmente, se sabe que los estadísticos de prueba usuales de MCO son válidos asintóticamente bajo los supuestos ST.1′ a ST.5′, los que se satisfacen en los modelos estáticos y de rezagos distribuidos con errores ARCH.

Si MCO aún tienen propiedades deseables bajo ARCH, ¿por qué preocuparse por las formas ARCH de heterocedasticidad en los modelos estáticos y de rezagos distribuidos? Es de interés por dos razones. En primer lugar, es posible obtener estimadores consistentes (aunque no insesgados) de las β_j que son *asintóticamente* más eficientes que los estimadores de MCO. Un procedimiento de mínimos cuadrados ponderados, basado en la estimación de la ecuación (12.50), resolverá el problema. Un procedimiento de máxima verisimilitud bajo el supuesto de que los errores u_t tienen una distribución normal condicional también sirve. En segundo lugar, los economistas de diversos campos se han interesado en la dinámica de la varianza condicional. La aplicación original de Engle fue a la varianza de la inflación del Reino Unido, en la que descubrió que una mayor magnitud del error en el periodo anterior (u_{t-1}^2 mayor) se asociaba con una varianza del error mayor en el periodo en curso. Puesto que la varianza a menudo se utiliza para medir la volatilidad y esta es un elemento clave en las teorías de fijación de precios de los activos, los modelos ARCH se han vuelto importantes en las finanzas empíricas.

Los modelos ARCH se aplican también cuando hay dinámica en la media condicional. Suponga que se tiene la variable dependiente y_t, una variable exógena contemporánea z_t y

$$\text{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1},$$

de manera que a lo sumo un rezago de y y de z aparecen en la regresión dinámica. El enfoque tradicional es suponer que $\mathrm{Var}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)$ es constante, como se vio en el capítulo 11. Pero esta varianza podría seguir un modelo ARCH:

$$\begin{aligned}\mathrm{Var}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) &= \mathrm{Var}(u_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)\\ &= \alpha_0 + \alpha_1 u_{t-1}^2,\end{aligned}$$

donde $u_t = y_t - \mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)$. Como se sabe a partir del capítulo 11, la presencia de ARCH no afecta la consistencia de MCO y los errores estándar robustos a la heterocedasticidad y los estadísticos de prueba son válidos. (Recuerde, son válidos para cualquier forma de heterocedasticidad y ARCH es sólo una forma particular de heterocedasticidad.)

Si usted está interesado en el modelo ARCH y sus extensiones, consulte Bollerslev, Chou y Kroner (1992) y Bollerslev, Engle y Nelson (1994) para revisiones recientes de la literatura.

Ejemplo 12.9

[ARCH en rendimiento accionario]

En el ejemplo 12.8 se vio que había heterocedasticidad en los rendimientos accionarios semanales. Esta heterocedasticidad en realidad está mejor representada por el modelo ARCH (12.50). Si se calculan los residuales de MCO de la ecuación (12.47), se elevan al cuadrado y se hace la regresión de los mismos sobre los residuales cuadrados rezagados, se obtiene

$$\begin{aligned}\hat{u}_t^2 &= \underset{(.44)}{2.95} + \underset{(.036)}{.337}\,\hat{u}_{t-1}^2 + residual_t\\ n &= 688,\ R^2 = .114.\end{aligned} \tag{12.51}$$

El estadístico t de $\hat{u}_{t-1}^2$ es mayor a 9, lo que indica fuerte ARCH. Como se estudió antes, un error grande en el tiempo $t - 1$ implica una varianza mayor en los rendimientos accionarios actuales.

Es importante ver que, aun cuando los residuales *cuadrados* de MCO están autocorrelacionados, los residuales mismos de MCO no lo están (ya que son consistentes con la HME). Al hacer la regresión de $\hat{u}_t$ sobre $\hat{u}_{t-1}$ se tiene $\hat{\rho} = .0014$ con $t_{\hat{\rho}} = .038$.

Heterocedasticidad y correlación serial en modelos de regresión

Nada anula la posibilidad de que tanto la heterocedasticidad como la correlación serial estén presentes en un modelo de regresión. Si no se está seguro, siempre se puede usar MCO y calcular los errores estándar completamente robustos, como se describe en la sección 12.5.

La mayoría de las ocasiones, la correlación de series de tiempo se considera el problema más importante, debido a que por lo general tiene un mayor impacto en los errores estándar y en la eficiencia de los estimadores del que tiene la heterocedasticidad. Como se concluyó en la sección 12.2, la obtención de pruebas para la correlación serial que son robustas a la heterocedasticidad arbitraria es muy sencilla. Si en una de estas pruebas se detecta correlación serial, se puede emplear la transformación de Cochrane-Orcutt (o Prais-Winsten) [vea la ecuación (12.32)] y, en la ecuación transformada, usar errores estándar y estadísticos de prueba robustos a la heterocedasticidad. O, incluso se puede probar la heterocedasticidad en la ecuación (12.32) usando las pruebas de Breusch-Pagan o de White.

Otra opción es que se pueden modelar la heterocedasticidad y la correlación serial y corregir ambas mediante un procedimiento combinado de mínimos cuadrados ponderados y AR(1). En concreto, considere el modelo

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t \\ u_t &= \sqrt{h_t}v_t \\ v_t &= \rho v_{t-1} + e_t, \ |\rho| < 1, \end{aligned} \tag{12.52}$$

donde las variables explicativas $\mathbf{X}$ son independientes de e_t para toda t y h_t es una función de x_{tj}. El proceso $\{e_t\}$ tiene media cero, varianza constante σ_e^2 y no está correlacionado serialmente. Por tanto, $\{v_t\}$ satisface un proceso AR(1) estable. El error u_t es heterocedástico, además de contener correlación serial:

$$\mathrm{Var}(u_t|\mathbf{x}_t) = \sigma_v^2 h_t,$$

donde $\sigma_v^2 = \sigma_e^2/(1 - \rho^2)$. Pero $v_t = u_t/\sqrt{h_t}$ es homocedástica y sigue un modelo AR(1) estable. De modo que la ecuación transformada

$$y_t/\sqrt{h_t} = \beta_0(1/\sqrt{h_t}) + \beta_1(x_{t1}/\sqrt{h_t}) + \dots + \beta_k(x_{tk}/\sqrt{h_t}) + v_t \tag{12.53}$$

tiene errores AR(1). Ahora bien, si se tiene un tipo particular de heterocedasticidad en mente, es decir, se conoce h_t, se puede estimar la ecuación (12.53) usando los métodos estándar de CO o PW.

En la mayoría de los casos, primero se tiene que estimar h_t. El método siguiente combina el método de mínimos cuadrados ponderados de la sección 8.4 con la corrección de la correlación serial de la sección 12.3.

MCG factibles con heterocedasticidad y correlación serial AR(1):

i) Estime la ecuación (12.52) por MCO y guarde los residuales, $\hat{u}_t$.

ii) Efectúe la regresión de $\log(\hat{u}_t^2)$ sobre $x_{t1}, \dots, x_{tk}$ (o sobre $\hat{y}_t, \hat{y}_t^2$) y obtenga los valores ajustados, por ejemplo, $\hat{g}_t$.

iii) Obtenga las estimaciones de h_t: $\hat{h}_t = \exp(\hat{g}_t)$.

iv) Estime la ecuación transformada

$$\hat{h}_t^{-1/2} y_t = \hat{h}_t^{-1/2}\beta_0 + \beta_1 \hat{h}_t^{-1/2} x_{t1} + \dots + \beta_k \hat{h}_t^{-1/2} x_{tk} + error_t \tag{12.54}$$

por los métodos estándar de Cochrane-Orcutt o Prais-Winsten.

Los estimadores de MCG factibles obtenidos del procedimiento son asintóticamente eficientes, siempre y cuando los supuestos del modelo (12.52) se cumplan. Resulta más importante que todos los errores estándar y los estadísticos de prueba de la estimación CO o PW son asintóticamente válidos. Si se permite que la función de varianza se especifique incorrectamente, o la posibilidad de que alguna correlación serial no siga un modelo AR(1), entonces se puede aplicar la cuasi diferenciación a la ecuación (12.54), estimando la ecuación resultante por MCO, y luego obtener los errores estándar de Newey-West. Al hacer esto, se estaría utilizando un procedimiento que podría ser asintóticamente eficiente al mismo tiempo que garantizara que nuestra inferencia sea válida (asintóticamente) si se ha especificado incorrectamente nuestro modelo, ya sea de heterocedasticidad o de correlación serial.

RESUMEN

Se ha cubierto el importante problema de la correlación serial de los errores en modelos de regresión múltiple. La correlación positiva entre errores adyacentes es común, en especial en modelos estáticos y con rezagos distribuidos finitos. Esto provoca que los errores estándar de MCO y los estadísticos usuales induzcan a errores (aunque las $\hat{\beta}_j$ pueden seguir siendo insesgadas, o al menos consistentes). Por lo general, los errores estándar de MCO subestiman la verdadera incertidumbre en la estimación de parámetros.

El modelo más popular de correlación serial es el AR(1). Si se toma éste como punto de partida, es fácil probar la presencia de correlación serial AR(1) usando los residuales de MCO. Un estadístico t asintóticamente válido se obtiene al efectuar la regresión de los residuales de MCO sobre los residuales rezagados, suponiendo que los regresores son estrictamente exógenos y que se cumple el supuesto de la homocedasticidad. Hacer la prueba robusta a la heterocedasticidad es sencillo. El estadístico de Durbin-Watson está disponible bajo los supuestos del modelo lineal clásico, pero puede conducir a un resultado no concluyente, y tiene poco que ofrecer por encima de la prueba t.

Para modelos con una variable dependiente rezagada u otros regresores no estrictamente exógenos, la prueba t estándar sobre el coeficiente de $\hat{u}_{t-1}$ sigue siendo válida, siempre y cuando todas las variables independientes se incluyan como regresores junto con $\hat{u}_{t-1}$. Se puede usar un estadístico F o ML para probar la correlación serial de orden superior.

En modelos con regresores estrictamente exógenos es posible usar un procedimiento de MCG factibles —Cochrane-Orcutt o Prais-Winsten— para corregir la correlación serial AR(1). Esto da estimaciones diferentes de las estimaciones por MCO: las estimaciones por MCGF se obtienen de aplicar MCO sobre las variables *cuasi diferenciadas*. Todos los estadísticos de prueba usuales de la ecuación transformada son asintóticamente válidos. La mayoría de los paquetes de regresión tienen funciones integradas para estimar modelos con errores AR(1).

Otra manera de tratar la correlación serial, en especial cuando el supuesto de exogeneidad estricta no se cumple, es usar MCO pero calculando los errores estándar robustos a la correlación serial (que también son robustos a la heterocedasticidad). Muchos paquetes de regresión siguen un método sugerido por Newey y West (1987); también se pueden usar paquetes de regresión estándar para obtener un error estándar a la vez.

Por último, se analizaron algunas características especiales de la heterocedasticidad en modelos de series de tiempo. Como en el caso de corte transversal, el tipo más importante de heterocedasticidad es aquel que depende de las variables explicativas; esto es lo que determina si los estadísticos usuales de MCO son válidos. Las pruebas de Breusch-Pagan y White cubiertas en el capítulo 8 pueden aplicarse de manera directa, con la advertencia de que los errores no deben correlacionarse serialmente. En años recientes, los economistas, en especial aquellos que estudian los mercados financieros, han mostrado interés en las formas dinámicas de heterocedasticidad. El modelo ARCH es el principal ejemplo.

TÉRMINOS CLAVE

Correlación serial AR(1)
Datos cuasi diferenciados
Error estándar robusto a la correlación serial
Estadístico de Durbin-Watson (*DW*)
Estimación de Cochrane-Orcutt (CO)
Estimación de Prais-Winsten (PW)
Heterocedasticidad condicional autorregresiva (ARCH)
MCG factibles (MCGF)
Mínimos cuadrados ponderados
Prueba de Breusch-Godfrey

PROBLEMAS

12.1 Cuando en un modelo de regresión los errores tienen correlación serial AR(1), ¿por qué los errores estándar de MCO tienden a subestimar la variación de muestreo en las $\hat{\beta}_j$? ¿Siempre es cierto que los errores estándar de MCO son muy pequeños?

12.2 Explique por qué es incorrecta la siguiente afirmación: "Los métodos de Cochrane-Orcutt y de Prais-Winsten se usan ambos para obtener errores estándar válidos para las estimaciones por MCO cuando hay una correlación serial".

12.3 En el ejemplo 10.6, se estimó una variante del modelo de Fair para pronosticar los resultados de la elección presidencial en Estados Unidos.

i) ¿Qué argumento puede formularse para justificar que el término de error en la ecuación no esté correlacionado serialmente? (*Sugerencia:* ¿Con qué frecuencia se llevan a cabo las elecciones presidenciales?)

ii) Cuando se hace la regresión de los residuales de MCO de la ecuación (10.23) sobre los residuales rezagados, se obtiene $\hat{\rho} = -.068$ y ee$(\hat{\rho}) = .240$. ¿A qué conclusión llega acerca de la correlación serial de u_t?

iii) ¿Le preocupa el tamaño de muestra pequeño de esta aplicación al probar la correlación serial?

12.4 Cierto o falso: "Si en un modelo de regresión los errores contienen ARCH deben correlacionarse serialmente".

12.5 i) En el estudio de evento referente a las zonas empresariales del ejercicio para computadora C10.5, una regresión de los residuales de MCO sobre los residuales rezagados produce $\hat{\rho} = .841$ y ee$(\hat{\rho}) = .053$. ¿Qué consecuencias tiene esto para MCO?

ii) Si usted quisiera utilizar MCO, pero también deseara obtener un error estándar válido para el coeficiente de la zona empresarial (*ze*), ¿qué haría?

12.6 En el ejemplo 12.8, se encontró evidencia de heterocedasticidad en u_t en la ecuación (12.47). Así, se calculan los errores estándar robustos a la heterocedasticidad (presentados entre [·]) junto con los errores estándar usuales:

$$\begin{aligned}\widehat{return}_t = {} & \underset{\substack{(.081)\\ [.085]}}{.180} + \underset{\substack{(.038)\\ [.069]}}{.059}\, return_{t-1}\\ n = 689,\ & R^2 = .0035,\ \bar{R}^2 = .0020.\end{aligned}$$

¿Qué efecto tiene haber usado el estadístico t robusto a la heterocedasticidad en la significancia de $return_{t-1}$?

EJERCICIOS EN COMPUTADORA

C12.1 En el ejemplo 11.6 se estimó un modelo de rezagos distribuidos finitos en primeras diferencias:

$$\Delta gfr_t = \gamma_0 + \delta_0 \Delta pe_t + \delta_1 \Delta pe_{t-1} + \delta_2 \Delta pe_{t-2} + u_t.$$

Utilice la base de datos FERTIL3.RAW para probar si hay correlación serial AR(1) en los errores.

C12.2 i) Con los datos de WAGEPRC.RAW, estime el modelo de rezagos distribuidos del problema 11.5. Emplee la regresión (12.14) para probar la correlación serial AR(1).

ii) Vuelva a estimar el modelo usando la estimación iterada de Cochrane-Orcutt. ¿Cuál es su nueva estimación de la propensión de largo plazo?

iii) Con la estimación iterada de CO, determine el error estándar de la PLP (esto exige que estime una ecuación modificada). Determine si el estimador de la PLP es estadísticamente diferente de uno, a un nivel de 5%.

C12.3 i) En el inciso (i) del ejercicio para computadora C11.6, se le pidió que estimara el modelo del acelerador de la inversión en inventarios. Pruebe esta ecuación para la correlación serial AR(1).

ii) Si encuentra evidencia de correlación serial, vuelva a estimar la ecuación mediante el método de Cochrane-Orcutt y compare los resultados.

C12.4 i) Utilice la base de datos NYSE.RAW para estimar la ecuación (12.48). Sea $\hat{h}_t$ los valores ajustados de esta ecuación (las estimaciones de la varianza condicional). ¿Cuántas $\hat{h}_t$ son negativas?

ii) Agregue $return^2_{t-1}$ a la ecuación (12.48) y calcule de nuevo los valores ajustados, $\hat{h}_t$. ¿Algunas $\hat{h}_t$ son negativas?

iii) Utilice las $\hat{h}_t$ del inciso ii) para estimar la ecuación (12.47) por medio de mínimos cuadrados ponderados (como en la sección 8.4). Compare su estimación de β_1 con la de la ecuación (11.16). Pruebe H_0: $\beta_1 = 0$ y compare los resultados cuando se utilizan MCO.

iv) Ahora, estime la ecuación (12.47) por MCP, usando el modelo ARCH estimado en la ecuación (12.51) para obtener las $\hat{h}_t$. ¿Cambia esto sus resultados del inciso iii)?

C12.5 Considere la versión del modelo de Fair del ejemplo 10.6. Ahora, en lugar de predecir la proporción del voto bipartidista que corresponde a los demócratas, estime un modelo de probabilidad lineal que determine si los demócratas ganan o no.

i) Utilice la variable binaria *demwins* en lugar de *demvote* en la ecuación (10.23) e informe los resultados como acostumbra. ¿Qué factores influyen en la probabilidad de ganar? Utilice los datos sólo hasta 1992.

ii) ¿Cuántos valores ajustados son menores que cero? ¿Cuántos son mayores que uno?

iii) Aplique la siguiente regla de predicción: si $\widehat{demwins} > .5$, la predicción es que los demócratas ganan; de lo contrario, quienes triunfan son los republicanos. Con esta regla, determine cuántas de las veinte elecciones predice el modelo de forma correcta.

iv) Sustituya los valores de las variables explicativas para 1996. ¿Cuál es la probabilidad predicha de que Clinton ganara la elección? Clinton ganó; ¿obtuvo usted la predicción correcta?

v) Utilice la prueba *t* robusta a la heterocedasticidad para saber si hay correlación serial AR(1)en los errores. ¿Qué encontró?

vi) Obtenga los errores estándar robustos a la heterocedasticidad para las estimaciones del inciso i). ¿Hay cambios notables en cualesquiera de los estadísticos *t*?

C12.6 i) En el ejercicio para computadora C10.7, usted estimó una relación simple entre el crecimiento del consumo y el crecimiento del ingreso disponible. Pruebe en la ecuación la correlación serial AR(1) (con ayuda de CONSUMP.RAW).

ii) En el ejercicio para computadora C11.7, usted probó la hipótesis del ingreso permanente efectuando la regresión del crecimiento en el consumo sobre un rezago. Luego de llevar a cabo esta regresión, pruebe la heterocedasticidad realizando la regresión de los residuales cuadrados sobre gc_{t-1} y gc^2_{t-1}. ¿A qué conclusión llega?

C12.7 i) Para el ejemplo 12.4, usando los datos de BARIUM.RAW, obtenga las estimaciones iterativas de Cochrane-Orcutt.

ii) ¿Las estimaciones de Prais-Winsten y Cochrane-Orcutt son parecidas? ¿Esperaba usted que lo fueran?

C12.8 Utilice la base de datos TRAFFIC2.RAW para este ejercicio.

i) Realice la regresión por MCO de *prcfat* sobre una tendencia lineal en el tiempo, variables binarias mensuales y las variables *wkends*, *unem*, *spdlaw* y *beltlaw*. Pruebe si los errores tienen correlación serial AR(1) usando la regresión de la ecuación (12.14). ¿Tiene sentido utilizar la prueba que supone exogeneidad estricta de los regresores?

ii) Obtenga los errores estándar robustos a la correlación serial y la heterocedasticidad para los coeficientes de *spdlaw* y *beltlaw*, usando cuatro rezagos en el estimador de Newey-West. ¿Cómo afecta esto a la significancia estadística de las dos variables de políticas?

iii) Ahora estime el modelo usando la estimación iterativa de Prais-Winsten y compare su resultado con las estimaciones por MCO. ¿Hay cambios importantes en los coeficientes de las variables de políticas o en su significado estadístico?

C12.9 El archivo FISH.RAW contiene 97 observaciones de precios y cantidades diarias sobre los precios del pescado en el mercado Fulton Fish Market de la ciudad de Nueva York. Utilice la variable log(*avgprc*) el logaritmo del precio promedio del pescado, como la variable dependiente.

i) Realice la regresión de log(*avgprc*) sobre cuatro variables binarias diarias (*mon*, *tues*, *wed*, *thurs*), con el día viernes como base. Incluya una tendencia lineal en el tiempo. ¿Existe evidencia de que los precios varían de manera sistemática a lo largo de la semana?

ii) Ahora, añada las variables *wave2* y *wave3*, las cuales son mediciones de la altura de las olas durante varios días pasados. ¿Estas variables son individualmente significativas? Describa un mecanismo mediante el cual el mar tempestuoso aumente el precio del pescado.

iii) ¿Qué ocurrió con la tendencia en el tiempo cuando *wave2* y *wave3* se añadieron a la regresión? ¿Qué debe estar pasando?

iv) Explique por qué se supone que todas las variables explicativas en la regresión son estrictamente exógenas.

v) Pruebe los errores para la correlación serial AR(1).

vi) Obtenga los errores estándar de Newey-West usando cuatro rezagos. ¿Qué ocurre con los estadísticos t de *wave2* y *wave3*? ¿Esperaba un cambio más grande o más pequeño en comparación con los estadísticos t de MCO?

vii) Ahora obtenga las estimaciones de Prais-Winsten para el modelo estimado en el inciso ii). ¿Son *wave2* y *wave3* conjuntamente significativas en términos estadísticos?

C12.10 Emplee los datos de PHILLIPS.RAW para responder estas preguntas.

i) Usando la base de datos completa, estime la ecuación de la curva estática de Phillips $inf_t = \beta_0 + \beta_1 unem_t + u_t$ e informe los resultados de la manera usual.

ii) Obtenga los residuales de MCO del inciso i), $\hat{u}_t$, y obtenga ρ de la regresión de $\hat{u}_t$ sobre $\hat{u}_{t-1}$. (Está bien incluir un intercepto en esta regresión.) ¿Existe evidencia contundente de correlación serial?

iii) Ahora estime el modelo de la curva estática de Phillips mediante la estimación iterativa de Prais-Winsten. Compare la estimación de β_1 con aquélla obtenida en la tabla 12.2. ¿Hay mucha diferencia en la estimación cuando se añaden los años posteriores?

iv) En vez de usar Prais-Winsten, utilice la estimación iterativa de Cochrane-Orcutt. ¿Qué tan parecidas son las estimaciones finales de ρ? ¿Qué tan parecidas son las estimaciones de β_1 de PW y CO?

C12.11 Use los datos de NYSE.RAW para responder a estas preguntas.

i) Estime el modelo de la ecuación (12.47) y obtenga los residuales cuadrados de MCO. Calcule los valores promedio, mínimo y máximo de $\hat{u}_t^2$ en la muestra.

ii) Utilice los residuales cuadrados de MCO para estimar el modelo de heterocedasticidad siguiente:

$$\mathrm{Var}(u_t|return_{t-1}, return_{t-2}, \dots) = \mathrm{Var}(u_t|return_{t-1}) = \delta_0 + \delta_1 return_{t-1} + \delta_2 return_{t-1}^2.$$

Informe los coeficientes estimados, los errores estándar reportados, la R-cuadrada y la R-cuadrada ajustada.

iii) Considere la varianza condicional como una función del rezago $return_{-1}$. ¿Para qué valor de $return_{-1}$ la varianza es la menor y cuál es la varianza?

iv) Al predecir la varianza dinámica, ¿el modelo del inciso ii) produce alguna estimación negativa de la varianza?

v) ¿El modelo del inciso ii) parece ajustarse mejor o peor que el modelo ARCH(1) del ejemplo 12.9? Explique por qué.

vi) Para la regresión ARCH(1) de la ecuación (12.51), añada el segundo rezago, $\hat{u}_{t-2}^2$. ¿Este rezago le parece importante? ¿El modelo ARCH(2) se ajusta mejor que el modelo del inciso ii)?

C12.12 Use los datos de INVEN.RAW para este ejercicio; vea también el ejercicio para computadora C11.6.

i) Obtenga los residuales de MCO del modelo del acelerador $\Delta inven_t = \beta_0 + \beta_1 \Delta GDP_t + u_t$ y utilice la regresión de $\hat{u}_t$ sobre $\hat{u}_{t-1}$ para probar la correlación serial. ¿Cuál es la estimación de ρ? ¿Qué tan grave parece ser el problema de la correlación serial?

ii) Estime el modelo del acelerador por PW y compare la estimación de β_1 con la estimación por MCO. ¿Por qué esperaría que se parecieran?

C12.13 Utilice la base de datos OKUN.RAW para responder esta pregunta; vea también el ejercicio para computadora C11.11.

i) Estime la ecuación $pcrgdp_t = \beta_0 + \beta_1 \Delta unem_t + u_t$ y pruebe si los errores tienen correlación serial AR(1), sin dar por sentado que $\{\Delta unem_t: t = 1, 2, \dots\}$ es estrictamente exógena. ¿A qué conclusión llega?

ii) Haga la regresión de los residuales cuadrados, $\hat{u}_t^2$, sobre $\Delta unem_t$ (esta es la prueba de Breusch-Pagan para la heterocedasticidad en el caso de la regresión simple). ¿A qué conclusión llega?

iii) Obtenga el error estándar robusto a la heterocedasticidad para la estimación por MCO de $\hat{\beta}_1$. ¿Es considerablemente diferente del error estándar usual de MCO?

C12.14 Use los datos de MINWAGE.RAW para este ejercicio, concentrándose en el sector 232.

i) Estime la ecuación

$$gwage232_t = \beta_0 + \beta_1 gmwage_t + \beta_2 gcpi_i + u_t,$$

y pruebe si los errores tienen correlación serial AR(1). ¿Importa si usted supone que $gmwage_t$ y $gcpi_t$ son estrictamente exógenas? ¿Cuál es su conclusión general?

ii) Obtenga el error estándar de Newey-West para las estimaciones por MCO del inciso i), considerando 12 rezagos. ¿Cómo se comparan los errores estándar de Newey-West con los errores estándar usuales de MCO?

iii) Ahora obtenga los errores estándar robustos a la heterocedasticidad para MCO, y compárelos con los errores estándar usuales y los errores estándar de Newey-West. ¿Le parece que en esta aplicación la correlación serial o la heterocedasticidad constituyen un problema mayor?

iv) Utilice la prueba de Breusch-Pagan en la ecuación original para verificar si los errores exhiben una heterocedasticidad fuerte.

v) Añada los rezagos de 1 a 12 de *gmwage* a la ecuación del inciso i). Obtenga el valor-p de la prueba F conjunta para los rezagos 1 a 12, y compárelo con el valor-p para la prueba robusta a la heterocedasticidad. ¿Cómo afecta el ajuste de la heterocedastidad la significancia de los rezagos?

vi) Obtenga el valor-p para la prueba de significancia conjunta del inciso v) usando el método de Newey-West. ¿Cuál es su conclusión ahora?

vii) Si usted omite los rezagos de *gmwage*, ¿es muy diferente la estimación de la propensión de largo plazo?

PARTE 3

Temas avanzados

Ahora se estudiarán temas más especializados que, por lo general, no se cubren en los cursos de introducción de un semestre. Algunos de estos temas requieren algunas habilidades matemáticas más que el análisis de regresión múltiple que se ha visto en la primera y segunda partes. En el capítulo 13 se muestra cómo aplicar la regresión múltiple a combinaciones independientes de cortes transversales. Los subtemas expuestos son muy parecidos al análisis de corte transversal estándar, con la salvedad de que es posible estudiar cómo cambian las relaciones a lo largo del tiempo al incluir variables binarias temporales. También se muestra la manera de analizar los conjuntos de datos de panel en un esquema de regresión. El capítulo 14 cubre otros métodos de datos de panel más avanzados que se utilizan de manera rutinaria en el trabajo aplicado.

En los capítulos 15 y 16 se investiga el problema de las variables explicativas endógenas. El capítulo 15 presenta el método de variables instrumentales como un recurso para resolver el problema de las variables omitidas así como el error de medición. En la economía empírica se aplica con mucha frecuencia el método de mínimos cuadrados en dos etapas y resulta indispensable para estimar modelos de ecuaciones simultáneas, tema que se tratará en el capítulo 16.

El capítulo 17 cubre algunos temas muy avanzados que son de uso común en el análisis de corte transversal, incluyendo modelos para variables dependientes limitadas y métodos para corregir el sesgo en la selección de muestreo. El capítulo 18 toma un rumbo distinto, ya que cubre algunos avances recientes en la econometría de series de tiempo, los cuales han demostrado ser útiles en la estimación de las relaciones dinámicas.

El capítulo 19 es de mucha ayuda para aquellos estudiantes que deban redactar un trabajo final del curso o algún otro trabajo de ciencias sociales aplicadas. El capítulo ofrece sugerencias sobre cómo elegir un tema, recolectar y analizar datos, y redactar el trabajo.

CAPÍTULO 13

Combinación de cortes transversales en el tiempo: métodos simples para datos de panel

Hasta ahora, se ha cubierto el análisis de regresión múltiple usando sólo datos de corte transversal o de series de tiempo. Si bien en las aplicaciones a menudo surgen estos dos casos, los conjuntos de datos que cuentan con dimensiones tanto de cortes transversales como de series de tiempo se utilizan cada vez con más frecuencia en la investigación empírica. En estos conjuntos aún se emplean métodos de regresión múltiple. De hecho, los datos con aspectos de corte transversal y de series de tiempo por lo común esclarecen importantes cuestiones de política. En este capítulo se verán varios ejemplos.

Se analizarán dos tipos de conjuntos de datos en este capítulo. Una **combinación independiente de cortes transversales** se obtiene mediante un muestreo aleatorio de una población grande en distintos puntos del tiempo (por lo general, aunque no necesariamente, en años diferentes). Por ejemplo, en cada año, se puede extraer una muestra aleatoria de los salarios por hora, la educación, la experiencia, etc., de la fuerza laboral activa de Estados Unidos. O bien, en cualquier otro año, es posible extraer una muestra aleatoria de los precios de venta, la superficie, el número de baños, etc. de las casas vendidas en determinada zona metropolitana. Desde un punto de vista estadístico, estos conjuntos de datos cuentan con una importante característica: constan de observaciones *independientemente* muestreadas. Este es un aspecto clave también en el análisis de datos de corte transversal: entre otras cosas, excluye la correlación en los términos de error para distintas observaciones.

Una combinación independiente de cortes transversales difiere de una sola muestra aleatoria en cuanto a que es probable que el muestreo de la población en distintos puntos del tiempo conduzca a observaciones que no se distribuyen de modo idéntico. Por ejemplo, en la mayoría de los países las distribuciones de salarios y educación han cambiado con el tiempo. Como se verá más adelante, esto es sencillo de abordar en la práctica si se permite que el intercepto de un modelo de regresión múltiple, y en algunos casos las pendientes, cambie con el tiempo. En la sección 13.1 se cubrirán estos modelos y en la sección 13.2 se estudiará la manera de emplear la combinación de cortes transversales en el tiempo para evaluar los cambios de políticas.

Un conjunto de **datos de panel**, incluso si tiene tanto una dimensión de corte transversal como una de serie de tiempo, difiere en algunos importantes aspectos de una combinación independiente de cortes transversales. Para recolectar datos de panel, a los cuales en ocasiones se les llama datos longitudinales, se da seguimiento (o se intenta) a los *mismos* individuos, familias, empresas, ciudades, estados o cualquier otra cosa a lo largo del tiempo. Por ejemplo, para un conjunto de datos de panel sobre salarios individuales, horas de trabajo, educación y otros factores se hace una recolección aleatoria, eligiendo a personas de una población en un momento determinado. Luego, se vuelve a entrevistar a esos *mismos* sujetos en diversos momentos posteriores; lo que proporciona datos sobre el salario, las horas de trabajo, la educación, etc., del mismo grupo de personas en años distintos.

Es muy fácil reunir conjuntos de datos de panel de distritos escolares, ciudades, municipios, estados y países, y el análisis de políticas mejora enormemente con el uso de dichos conjuntos;

más adelante se verán algunos ejemplos. Para el análisis econométrico de datos de panel, no se puede dar por sentado que las observaciones se distribuyan de forma independiente en el tiempo. Por ejemplo, los factores inobservables (como la capacidad) que influyen en el salario de una persona en 1990 también influirán en el salario de esa persona en 1991; los factores inobservables que afectan la tasa de delitos de una ciudad en 1985 también la afectarán en 1990. Por este motivo, se han ideado modelos y métodos especiales para analizar datos de panel. En las secciones 13.3, 13.4 y 13.5 se describirá el sencillo método de la diferenciación para eliminar atributos inobservables constantes en el tiempo de las unidades bajo estudio. Como los métodos para datos de panel son un poco más avanzados, se usará principalmente la intuición cuando se describan las propiedades estadísticas de los procedimientos de estimación, y se dejarán los pormenores para el apéndice del capítulo. Se continúa la misma estrategia en el capítulo 14, donde se cubren métodos más complejos para datos de panel.

13.1 Combinación independiente de cortes transversales en el tiempo

Muchos estudios de personas, familias y empresas se repiten a intervalos regulares, a menudo cada año. Un ejemplo es la encuesta *Current Population Survey* (o CPS) que cada año hace un muestreo aleatorio de los hogares estadounidenses (vea, por ejemplo, la base de datos CPS78_85.RAW, que contiene datos de la CPS de 1978 a 1985). Si se extrae una muestra aleatoria en cada periodo, al combinar los resultados de cada muestra se obtiene una combinación independiente de cortes transversales.

Una razón para utilizar una combinación independiente de cortes transversales es que el tamaño de muestra se incrementa. Al combinar muestras aleatorias extraídas de la misma población, pero en distintos puntos del tiempo, se obtienen estimadores más precisos y estadísticos con mayor potencia de prueba. Este tipo de combinación es útil a este respecto sólo en la medida en que la relación entre la variable dependiente y al menos algunas de las variables independientes permanece constante con el paso del tiempo.

Como se mencionó en la introducción, la utilización de una combinación de cortes transversales sólo plantea complicaciones estadísticas menores. Por lo común, para reflejar el hecho de que es posible que la población tenga distintas distribuciones en diferentes periodos, se permite que el intercepto difiera a través de los periodos, que por lo general son años. Esto se consigue fácilmente si se incluyen variables binarias para todos los años excepto uno, el primero, que se elige como año base de la muestra. También es posible que la varianza del error cambie con el tiempo, algo que se analizará más adelante.

A veces resulta de interés el patrón de coeficientes de las variables binarias anuales. Por ejemplo, a un demógrafo puede interesarle la siguiente pregunta: *Después* de controlar la educación, ¿se ha modificado el patrón de fertilidad entre las mujeres mayores de 35 años entre 1972 y 1984? El siguiente ejemplo ilustra cómo es fácil responder a esta pregunta con sólo aplicar el análisis de regresión múltiple con **variables binarias anuales**.

Ejemplo 13.1

[Fertilidad de las mujeres en el tiempo]

La base de datos FERTIL1.RAW, que es similar a la base de datos utilizada por Sander (1992), proviene de la encuesta *General Social Survey* del Centro Nacional de Investigación de Opinión de Estados Unidos, para los años pares del periodo de 1972 a 1984, inclusive. Se utilizan estos datos para estimar un modelo que explique el número total de hijos que tiene una mujer (*kids*).

Una pregunta interesante es: después de controlar otros factores observables, ¿qué ha ocurrido con las tasas de fertilidad respecto al tiempo? Los factores que se controlan son: los años de educación, la edad, la raza, la religión, la región del país donde vivían a la edad de 16 años y las condiciones de vida a esa edad. Las estimaciones se muestran en la tabla 13.1.

TABLA 13.1

Determinantes de la fertilidad de las mujeres

Variable dependiente: *kids*		
Variables independientes	**Coeficientes**	**Errores estándar**
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constante	−7.742	3.052
$n = 1{,}129$ $R^2 = .1295$ $\bar{R}^2 = .1162$		

El año base es 1972. Los coeficientes sobre las variables binarias anuales muestran un marcado descenso en la fertilidad a principios de la década de 1980. Por ejemplo, el coeficiente de *y82* implica que, manteniendo fijos la educación, la edad y otros factores, una mujer tuvo en promedio .52 hijos menos o aproximadamente medio hijo menos en 1982 que en 1972. Este descenso es muy grande: si se mantienen fijos *educ*, *age* y los otros factores, se predice que 100 mujeres tendrán alrededor de 52 hijos menos en 1982 que 100 mujeres comparables en 1972. Dado que se está controlando la educación, esta disminución es distinta de la que ocurre en la fertilidad debida al incremento en el nivel educativo. (Los años medios de educación son 12.2 para 1972 y 13.3 para 1984). Los coeficientes sobre *y82* y *y84* representan los descensos en la fertilidad por razones que no se capturan en las variables explicativas.

Dado que las binarias de los años 1982 y 1984 son muy significativas de manera individual, no es de extrañarse que como un grupo las binarias de los años sean muy significativas conjuntamente: la R-cuadrada para la regresión sin las binarias de los años es .1019 y esto conduce a $F_{6,1111} = 5.87$ y el valor-$p \approx 0$.

Las mujeres con más educación tienen menos hijos y la estimación es muy significativa estadísticamente. Si todo lo demás se mantiene igual, 100 mujeres con educación universitaria tendrán en promedio alrededor de 51 hijos menos que 100 mujeres que sólo tienen bachillerato: .128(4) = .512. La edad tiene un efecto cada vez menor sobre la fertilidad. (El cambio de pendiente en la función cuadrática es aproximadamente en $age = 46$, edad en la cual la mayoría de las mujeres han dejado de tener hijos.)

El modelo estimado en la tabla 13.1 supone que el efecto de cada variable explicativa, en particular la educación, ha permanecido constante. Esto puede ser cierto o no; en el ejercicio para computadora C13.1 se le pedirá que explore este problema.

Por último, puede haber heterocedasticidad en el término de error que subyace a la ecuación estimada. Esto se resuelve mediante los métodos estudiados en el capítulo 8. Hay una diferencia interesante aquí: ahora la varianza del error puede cambiar con el tiempo incluso si no cambia con los valores de *educ*, *age*, *black*, etc. No obstante, los errores estándar robustos a la heterocedasticidad y los estadísticos de prueba son válidos. La prueba de Breusch-Pagan se obtendría al hacer la regresión de los residuales cuadrados de MCO sobre *todas* las variables independientes de la tabla 13.1, incluyendo variables binarias de los años. (Para el caso especial del estadístico de White, los valores ajustados $\widehat{kids}$ y los valores ajustados cuadrados se usan como las variables independientes, como siempre.) Un procedimiento de mínimos cuadrados ponderados debe considerar las varianzas que posiblemente cambien con el tiempo. En el procedimiento analizado en la sección 8.4 las variables binarias de los años se incluyeron en la ecuación (8.32).

También se puede interactuar una variable binaria anual con las variables explicativas clave para ver si el efecto de esa variable ha cambiado a lo largo de un cierto periodo. El siguiente ejemplo examina cómo la rentabilidad de la educación y las diferencias de género han cambiado de 1978 a 1985.

Pregunta 13.1

Cuando estudia los datos de la tabla 13.1, una persona afirma que, si todo lo demás permanece igual en la tabla, se espera que una mujer negra tenga un hijo más que una mujer que no es negra. ¿Está de acuerdo con esta afirmación?

Ejemplo 13.2

[Cambios en la rentabilidad de la educación y en la diferencia de salario por género]

Una ecuación log(*wage*) (donde *wage* es el salario por hora) que combina datos de los años 1978 (año base) y 1985 es

$$\begin{aligned}\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85{\cdot}educ + \beta_2 exper \\ + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85{\cdot}female + u,\end{aligned} \tag{13.1}$$

donde la mayoría de las variables explicativas ya deben resultarle familiares. La variable *union* es una variable binaria igual a uno si la persona está afiliada a un sindicato e igual a cero si no lo está. La variable *y85* es una variable binaria igual a uno si la observación es de 1985 e igual a cero si es de 1978. Hay 550 personas en la muestra de 1978 y un grupo distinto de 534 personas en la muestra de 1985.

El intercepto de 1978 es β_0 y el de 1985 es $\beta_0 + \delta_0$. El impacto de la educación en 1978 es β_1 y en 1985 es $\beta_1 + \delta_1$. Por tanto, δ_1 mide cómo el impacto de la educación se ha modificado en un periodo de siete años. Por último, en 1978 la diferencia en log(*wage*) entre las mujeres y los hombres es β_5, y en 1985 es $\beta_5 + \delta_5$. De esta manera, se puede comprobar la hipótesis nula de que esa diferencia no ha cambiado durante el periodo de siete años, al probar H_0: $\delta_5 = 0$. La alternativa de que la diferencia de género se ha *reducido* es H_1: $\delta_5 > 0$. Por simplicidad, se ha supuesto que la experiencia y la afiliación a un sindicato ejercen el mismo efecto sobre los salarios en ambos periodos.

Antes de presentar las estimaciones, hay otro aspecto que se debe tomar en cuenta, a saber, el salario por hora aquí está en dólares nominales (o corrientes). Como los salarios nominales aumentan simplemente debido a la inflación, en realidad se está interesado en el efecto de cada variable explicativa sobre los salarios reales. Suponga que optamos por medir los salarios en dólares de 1978. Esto exige deflactar los salarios de 1985 a dólares de 1978. (Si se usa el índice de precios al consumidor del informe *Economic Report of the President*, de 1997 el factor de deflación es 107.6/65.2 ≈ 1.65.) Aunque basta con dividir cada salario de 1985 entre 1.65, esto no es necesario *siempre y cuando* se incluya una variable binaria anual de 1985 en la regresión *y* log(*wage*) se emplee como variable dependiente (en oposición a *wage*). El uso del salario real o nominal en una forma funcional logarítmica influye sólo en el coeficiente de la variable binaria anual, *y85*. Para ver esto, sea *P85* el factor de deflación de los salarios de 1985 (1.65 si se usa el IPC). Por tanto, el logaritmo del salario real para cada persona *i* en la muestra de 1985 es

$$\log(wage_i/P85) = \log(wage_i) - \log(P85).$$

Ahora, si bien $wage_i$ difiere de una persona a otra, *P85* no lo hace. Por consiguiente, el intercepto de 1985 absorberá log(*P85*). (Esta conclusión se modificaría si, por ejemplo, se utilizara un índice de precios distinto para personas que viven en diversas partes del país.) En resumen, para estudiar el cambio que ha sufrido el impacto de la educación o la diferencia de género, no es necesario convertir los salarios nominales en salarios reales en la ecuación (13.1). El ejercicio para computadora C13.2 le pide que verifique esto para el actual ejemplo.

Si no se toman en cuenta los diferentes interceptos de 1978 y 1985, el uso de los salarios nominales puede producir resultados que inducen a graves errores. Si se usa *wage* en lugar de log(*wage*) como variable dependiente, es importante considerar el salario real e incluir una variable binaria anual.

El análisis anterior por lo general se aplica cuando se usan valores de dólar, ya sea para la variable dependiente o para las variables independientes. Siempre y cuando los montos en dólares aparezcan en forma logarítmica y se utilicen variables binarias para todos los periodos (excepto, desde luego, el periodo base), el uso de deflactores de precios agregados sólo afectará a los interceptos; ninguna de las estimaciones de la pendiente cambiará.

Ahora se usa la base de datos CPS78_85.RAW para estimar la ecuación:

$$\begin{aligned}
\widehat{\log(wage)} = {} & \underset{(.093)}{.459} + \underset{(.124)}{.118}\, y85 + \underset{(.0067)}{.0747}\, educ + \underset{(.0094)}{.0185}\, y85{\cdot}educ \\
& + \underset{(.0036)}{.0296}\, exper - \underset{(.00008)}{.00040}\, exper^2 + \underset{(.030)}{.202}\, union \\
& - \underset{(.037)}{.317}\, female + \underset{(.051)}{.085}\, y85{\cdot}female \\
& n = 1{,}084,\ R^2 = .426,\ \bar{R}^2 = .422.
\end{aligned} \tag{13.2}$$

Se estima que el impacto de la educación en 1978 es de alrededor de 7.5%; el impacto de la educación en 1985 es cercano a 1.85 puntos porcentuales *mayor*, es decir alrededor de 9.35%. Dado que el estadístico t sobre el término de interacción es $.0185/.0094 \approx 1.97$, la diferencia en el impacto de la educación es estadísticamente significativa al nivel de 5% contra una alternativa de dos colas.

¿Qué pasa con la diferencia de género? En 1978, si todos los demás factores permanecen iguales, una mujer ganaría alrededor de 31.7% menos que un hombre (27.2% es una estimación más precisa). En 1985 la diferencia en log(*wage*) es $-.317 + .085 = -.232$. Por consiguiente, la diferencia de género parece haber disminuido de 1978 a 1985 cerca de 8.5 puntos porcentuales. El estadístico t sobre el término de interacción es de alrededor de 1.67, lo cual indica que es significativo al nivel de 5% contra una alternativa de una cola positiva.

¿Qué pasa si se interactúan *todas* las variables independientes con *y85* en la ecuación (13.2)? Esto es idéntico a estimar dos ecuaciones separadas, una para 1978 y otra para 1985. A veces, esto es recomendable. Por ejemplo, en el capítulo 7 se trató un estudio de Krueger (1993) en el cual estimó el rendimiento del uso de computadoras en el trabajo. Krueger estima dos ecuaciones separadas, una que utiliza la encuesta CPS de 1984 y otra que utiliza la de 1989. Al comparar cómo cambia el impacto de la educación en el tiempo y si el uso de computadoras está controlado o no, estima que de una tercera parte a un medio del incremento observado en el impacto de la educación durante el periodo de cinco años puede atribuirse a un mayor uso de las computadoras. [Vea las tablas VIII y IX en Krueger (1993).]

Prueba de Chow para el cambio estructural en el tiempo

En el capítulo 7 se estudió cómo la prueba de Chow, que es sencillamente una prueba F, se usa para determinar si una función de regresión múltiple difiere en dos grupos. Esta prueba también se puede aplicar a dos periodos diferentes. Una forma de la prueba toma la suma de residuales cuadrados de la estimación combinada como la SRC restringida. La SRC no restringida es la suma de las SRC de los dos periodos estimados por separado. La mecánica del cálculo del estadístico es exactamente la misma que aquélla vista en la sección 7.4. También está disponible una versión robusta a la heterocedasticidad (vea la sección 8.2).

El ejemplo 13.2 propone otra forma de calcular la prueba de Chow para dos periodos: se interactúa cada variable con una variable binaria anual para uno de los dos años y se prueba la significancia conjunta de la variable binaria anual y de todos los términos de interacción. Dado que en un modelo de regresión el intercepto cambia a menudo con el tiempo (debido, digamos, a la inflación en el ejemplo del costo de la vivienda), esta prueba de Chow completa puede detectar estas modificaciones. A menudo resulta más interesante permitir que el intercepto difiera y luego probar si los coeficientes de la pendiente cambian con el tiempo (como se hizo en el ejemplo 13.2).

Una prueba de Chow también puede calcularse para más de dos periodos. Al igual que en el caso de dos periodos, por lo general es más interesante permitir que los interceptos cambien con el tiempo y luego probar si los coeficientes de la pendiente han sufrido cambios en el tiempo. Comúnmente podemos probar la constancia de los coeficientes de pendiente al interactuar todas las variables binarias de los periodos (salvo aquellas que definen el grupo base) con una, varias o todas las variables explicativas y probar la significancia conjunta de los términos de interacción. Los ejercicios para computadora C13.1 y C13.2 son ejemplos de ello. Para varios periodos y variables explicativas, la construcción de un conjunto completo de interacciones llega a ser tediosa. Otra posibilidad es adaptar el método descrito en la parte vi) del ejercicio para computadora C7.11. Primero se estima el modelo restringido al hacer una regresión combinada que tome en cuenta los diferentes interceptos en el tiempo; esto da como resultado la SRC_r. Luego se hace una regresión para cada uno de los, digamos, T periodos y se obtiene la suma de los residuales cuadrados para cada periodo. La suma no restringida de los residuales cuadrados se obtiene como $\text{SRC}_{nr} = \text{SRC}_1 + \text{SRC}_2 + \ldots + \text{SRC}_T$. Si hay k variables explicativas (sin incluir el intercepto o las variables binarias de tiempo) con k periodos, entonces se está probando $(T - 1)k$ restricciones y hay $T + Tk$ parámetros estimados en el modelo no restringido. Por tanto, si $n = n_1 + n_2 + \ldots + n_T$ es el número total de observaciones, entonces los gl de la prueba F son $(T - 1)k$ y $n - T - Tk$. Se calcula el estadístico F como siempre: $[(\text{SRC}_r - \text{SRC}_{nr})/\text{SRC}_{nr}][(n - T - Tk)/(T - 1)k]$. Lamentablemente, al igual que con cualquier prueba F basada en las sumas de residuales cuadrados o R-cuadradas, esta prueba no es robusta a la heterocedasticidad (incluidas las varianzas que cambian con el tiempo). Para obtener una prueba robusta a la heterocedasticidad, se deben construir los términos de interacción y hacer una regresión combinada.

13.2 Análisis de políticas con combinación de cortes transversales

La combinación de cortes transversales es muy útil para evaluar el impacto de ciertos eventos o políticas. El ejemplo que sigue es un estudio de evento que muestra cómo se utilizan dos conjuntos de datos de corte transversal, recolectados antes y después de la ocurrencia de un evento, para determinar el efecto sobre variables económicas.

Ejemplo 13.3

[Efecto de la ubicación de un incinerador de basura sobre los precios de las viviendas]

Kiel y McClain (1995) estudiaron el efecto que tuvo un nuevo incinerador de basura en el valor de las viviendas en North Andover, Massachusetts. Utilizaron datos de varios años y un análisis econométrico muy complejo. Aunque se emplearán datos de dos años y algunos modelos simplificados, el análisis que sigue es similar.

El rumor de que se construiría un nuevo incinerador de basura en North Andover comenzó después de 1978 y la construcción inició en 1981. Se esperaba que el incinerador entrara en operación poco después del comienzo de su construcción; en realidad comenzó a operar en 1985. Se utilizarán datos sobre los precios de las casas que se vendieron en 1978 y otra muestra de aquellas que se vendieron en 1981. La hipótesis es que el costo de las casas ubicadas cerca del incinerador estaría por debajo del precio de las viviendas más lejanas.

A manera de ejemplo, se considera que una casa está cerca del incinerador si se halla en un radio de 3 millas. [En el ejercicio para computadora C13.3, se le pide que utilice la distancia real desde la casa al incinerador, como en el caso de Kiel y McClain (1995).] Se comenzará con el análisis del efecto sobre el precio de las viviendas en términos de dólares, lo que requiere que el precio se mida en dólares constantes. Se miden todos los precios en dólares de 1978, con la ayuda del índice de precios de la vivienda en Boston. Sea *rprice* el precio de la casa en términos reales.

Un analista novato utilizaría solamente los datos de 1981 y estimaría un modelo muy sencillo:

$$rprice = \gamma_0 + \gamma_1 nearinc + u, \tag{13.3}$$

donde *nearinc* es una variable binaria igual a uno si la casa está cerca del incinerador y a cero en caso contrario. La estimación de esta ecuación usando los datos de KIELMC.RAW es

$$\begin{aligned} \widehat{rprice} &= 101{,}307.5 - 30{,}688.27\, nearinc \\ &\qquad (3{,}093.0) \qquad (5{,}827.71) \\ n &= 142,\ R^2 = .165. \end{aligned} \tag{13.4}$$

Dado que se trata de una regresión simple sobre una sola variable binaria, el intercepto es el precio de venta promedio de las casas que no están cerca del incinerador y el coeficiente de *nearinc* es la diferencia entre el precio promedio de las casas cercanas y aquél de las casas lejanas. La estimación muestra que el precio de venta promedio para el primer grupo fue de \$30,688.27 menor que para el segundo grupo. El estadístico t es mayor que cinco en valor absoluto, de manera que se rechaza enfáticamente la hipótesis de que el precio medio de las casas cercanas y el de las casas lejanas es el mismo.

Por desgracia, la ecuación (13.4) *no* implica que la ubicación del incinerador esté reduciendo el valor de las casas. De hecho, si se lleva a cabo la misma regresión para 1978 (aun antes de que se corrieran los rumores de la construcción del incinerador), se obtiene

$$\begin{aligned} \widehat{rprice} &= \underset{(2,653.79)}{82,517.23} - \underset{(4,744.59)}{18,824.37}\, nearinc \\ n &= 179,\ R^2 = .082. \end{aligned} \tag{13.5}$$

Por tanto, incluso *antes* de que se hiciera algún comentario de un incinerador, el valor medio de una casa cercana al lugar era de \$18,824.37 menor que el de una casa lejana (\$82,517.23); también la diferencia es estadísticamente significativa. Esto es congruente con el punto de vista de que el incinerador se construyó en una zona donde el valor de la vivienda es bajo.

¿Cómo distinguir entonces si la construcción de un nuevo incinerador reduce el valor de la vivienda? La clave es considerar la forma en que el coeficiente de *nearinc* cambió entre 1978 y 1981. La diferencia en el valor medio de las casas fue mucho mayor en 1981 que en 1978 (\$30,688.27 en comparación con \$18,824.37), incluso como porcentaje del valor medio de las casas no cercanas al sitio del incinerador. La diferencia entre los dos coeficientes de *nearinc* es

$$\hat{\delta}_1 = -30,688.27 - (-18,824.37) = -11,863.9.$$

Esta es nuestra estimación del efecto del incinerador en los valores de las casas cercanas al mismo. En la economía empírica, $\hat{\delta}_1$ se ha llamado el **estimador de diferencia en diferencias** debido a que se expresa como

$$\hat{\delta}_1 = (\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}) - (\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}), \tag{13.6}$$

donde *nr* significa “cerca del sitio del incinerador” y *fr* quiere decir “lejos del incinerador”. En otras palabras, $\hat{\delta}_1$ es la diferencia respecto al tiempo de la diferencia promedio de los precios de la vivienda entre las dos ubicaciones.

Para probar si $\hat{\delta}_1$ es estadísticamente diferente de cero, se debe encontrar su error estándar usando un análisis de regresión. De hecho, $\hat{\delta}_1$ se obtiene mediante la estimación

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u, \tag{13.7}$$

usando los datos combinados de ambos años. El intercepto, β_0, es el precio promedio de una casa no cercana al incinerador en 1978. El parámetro δ_0 captura los cambios de valor de *todas* las casas en North Andover de 1978 a 1981. [Una comparación de las ecuaciones (13.4) y (13.5) mostró que los valores de las casas en North Andover, respecto al índice de costos de la vivienda en Boston, se incrementó marcadamente en este periodo.] El coeficiente de *nearinc*, β_1, mide el efecto de la ubicación que no se debe a la presencia del incinerador: como se vio en la ecuación (13.5), incluso en 1978, las casas cercanas al incinerador se vendieron a menor precio que las más alejadas del mismo.

TABLA 13.2

Efectos de la ubicación del incinerador en los precios de la vivienda

Variable dependiente: *rprice*			
Variable independiente	**(1)**	**(2)**	**(3)**
constante	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Otros controles	No	*age*, age^2	Conjunto completo
Observaciones *R*-cuadrada	321 .174	321 .414	321 .660

El parámetro de interés se halla en el término de interacción *y81·nearinc*: δ_1 mide la disminución en el valor de las casas debida al nuevo incinerador, siempre y cuando se suponga que las casas, tanto las cercanas como las lejanas al incinerador, no se revalorizaron a tasas distintas por otras razones.

Las estimaciones de la ecuación (13.7) se proporcionan en la columna 1 de la tabla 13.2. El único número que no es posible obtener de las ecuaciones (13.4) y (13.5) es el error estándar de $\hat{\delta}_1$. El estadístico t de $\hat{\delta}_1$ es cercano a -1.59, lo cual es marginalmente significativo contra una alternativa de una cola (valor-$p \approx .057$).

Kiel y McClain (1995) incluyeron varias características de la vivienda en sus análisis de la ubicación del incinerador. Existen dos buenas razones para hacer esto. En primer lugar, el tipo de las casas vendidas en 1981 podría haber sido sistemáticamente distinto a aquél de las casas vendidas en 1978; si éste fuera el caso, es importante controlar las características que pudieran haber sido diferentes. Resulta igualmente importante, incluso si las características de la vivienda promedio son las mismas para ambos años, que el hecho de incluirlas reduce enormemente la varianza del error, lo que entonces puede reducir el error estándar de $\hat{\delta}_1$. (Vea la sección 6.3 para un análisis de esto.) En la columna 2 se ha controlado la edad de las casas con ayuda de una función cuadrática. Esto aumenta considerablemente la *R*-cuadrada (al reducir la varianza residual). El coeficiente de *y81·nearinc* ahora es mucho mayor en magnitud y su error estándar es menor.

Además de las variables de edad de la columna 2, la columna 3 controla la distancia a la carretera interestatal en pies (*intst*), el área del terreno en pies (*land*), el área de la casa en pies (*area*), el número de habitaciones (*rooms*) y el número de baños (*baths*). Esto genera una estimación de *y81·nearinc* más cercana a aquella obtenida sin ningún control, pero también produce un error estándar menor: el estadístico t para $\hat{\delta}_1$ es cercano a -2.84. Por consiguiente, se encuentra un efecto mucho más significativo en la columna 3 que en la 1. Las estimaciones de la columna 3 son preferibles porque controlan la mayor parte de los factores y cuentan con los errores estándar más pequeños (salvo en la constante, que no es importante aquí). El hecho de que *nearinc* tenga un coeficiente mucho menor y no significativo en la columna 3 indica que las características incluidas en dicha columna capturan en buena medida las características de las casas que resultan más importantes para determinar el precio de las viviendas.

Con el fin de introducir el método, se utilizó el nivel de los precios reales de la vivienda en la tabla 13.2. Tiene más sentido usar log(*price*) [o log(*rprice*)] en el análisis para obtener un efecto porcentual aproximado. El modelo básico se vuelve

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u. \qquad \mathbf{13.8}$$

Ahora bien, $100 \cdot \delta_1$ es la reducción porcentual aproximada en el valor de las casas debida al incinerador. [Como en el ejemplo 13.2, el uso de log(*price*) en lugar de log(*rprice*) sólo influye en el coeficiente de *y81*.] Si se usan las mismas 321 observaciones combinadas se tiene

$$\begin{aligned} \widehat{\log(price)} &= \underset{(.31)}{11.29} + \underset{(.045)}{.457}\, y81 - \underset{(.055)}{.340}\, nearinc - \underset{(.083)}{.063}\, y81 \cdot nearinc \\ n &= 321,\ R^2 = .409. \end{aligned} \qquad \mathbf{13.9}$$

El coeficiente del término de interacción implica que, debido al nuevo incinerador, las casas cercanas al mismo pierden cerca de 6.3% de su valor. Sin embargo, esta estimación no es estadísticamente distinta de cero. Pero cuando se utiliza un conjunto completo de controles, como en la columna 3 de la tabla 13.2 (pero con *intst*, *land* y *area* en forma logarítmica), el coeficiente de *y81·nearinc* se vuelve $-.132$ con un estadístico t cercano a -2.53. De nuevo, es importante controlar otros factores. Al utilizar la forma logarítmica, se estima que las casas cercanas al incinerador se devaluaron alrededor de 13.2%.

La metodología empleada en el ejemplo anterior cuenta con numerosas aplicaciones, en especial cuando los datos surgen de un **experimento natural** (o de un **cuasiexperimento**). Un experimento natural tiene lugar cuando algún evento exógeno, a menudo un cambio en las políticas gubernamentales, modifica el ambiente en que se desenvuelven las personas, las familias, las empresas o las ciudades. Un experimento de esta índole siempre cuenta con un grupo de control, que no se ve afectado por el cambio de políticas, y un grupo de tratamiento, que sí se ve afectado por este cambio. A diferencia de un experimento verdadero, en el cual los grupos de control y de tratamiento se eligen de manera aleatoria y explícita, en los experimentos naturales los grupos surgen a partir del cambio en una política particular. Para controlar las diferencias sistemáticas entre estos dos grupos, se necesitan datos de dos años, un año antes y otro después del cambio de la política. Así, nuestra muestra por lo común se descompone en cuatro grupos: el grupo de control antes del cambio, el grupo de control después del cambio, el grupo de tratamiento antes del cambio y el grupo de tratamiento después del cambio.

Sea C el grupo de control y T el grupo de tratamiento, donde dT es igual a uno para los miembros del grupo de tratamiento T e igual a cero si no lo son. De esta manera, suponiendo que $d2$ indica una variable binaria para el segundo periodo (después del cambio de la política), la ecuación que nos interesa es

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + otros\ factores, \qquad \mathbf{13.10}$$

donde y es la variable resultante que nos interesa. Al igual que en el ejemplo 13.3, δ_1 mide el efecto de la política. Sin otros factores en la regresión, $\hat{\delta}_1$ será el estimador de diferencia en diferencias:

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}), \qquad \mathbf{13.11}$$

TABLA 13.3

Ejemplo del estimador de diferencia en diferencias

	Antes	Después	Antes - Después
Control	β_0	$\beta_0 + \delta_0$	δ_0
Tratamiento	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Tratamiento − Control	β_1	$\beta_1 + \delta_1$	δ_1

donde la barra indica el promedio, el primer subíndice denota el año y el segundo subíndice, el grupo.

La configuración general de la diferencia en diferencias se muestra en la tabla 13.3. Esta sugiere que el parámetro δ_1, algunas veces llamado **efecto promedio del tratamiento** (debido a que mide el efecto del "tratamiento" o la política sobre el resultado promedio de y), puede estimarse de dos maneras: 1) Calcular las diferencias en los promedios entre los grupos de tratamiento y de control para cada periodo, y luego obtener la diferencia de los resultados respecto al tiempo; esto se hace sencillamente como en la ecuación (13.11); 2) calcular el cambio en los promedios respecto al tiempo para cada uno de los grupos de tratamiento y de control, y después determinar la diferencia entre estos cambios, lo cual simplemente se escribe $\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{1,T}) - (\bar{y}_{2,C} - \bar{y}_{1,C})$. Como es natural, la estimación de $\hat{\delta}_1$ no depende de cómo se hace la diferenciación, ya que se trata de una simple reorganización.

Cuando las variables explicativas se añaden a la ecuación (13.10) (para controlar el hecho de que las poblaciones muestreadas pueden diferir sistemáticamente a lo largo de dos periodos), la estimación por MCO de δ_1 ya no tiene la forma simple de la ecuación (13.11), pero su interpretación es parecida.

Ejemplo 13.4

[Efecto de las leyes de indemnización a los trabajadores sobre las semanas sin trabajo]

Meyer, Viscusi y Durbin (1995) (de aquí en adelante, MVD) estudiaron la cobertura temporal (en semanas) de la indemnización que recibe un trabajador lesionado. El 15 de julio de 1980, Kentucky aumentó el tope de ingresos semanales cubiertos por la indemnización de los trabajadores. Un aumento en el tope no tiene efecto sobre los beneficios para los trabajadores de bajos ingresos, pero se vuelve menos costoso para los trabajadores de altos ingresos estar recibiendo la indemnización. Por tanto, el grupo de control son los trabajadores de bajos ingresos y el grupo de tratamiento son los trabajadores con altos ingresos; estos últimos se definen como aquellos que estuvieron sujetos al tope previo al cambio de política. Usando muestras aleatorias tanto antes como después del cambio de política, MVD pudieron probar si una indemnización más generosa a los trabajadores provoca que la gente permanezca sin trabajar más tiempo (si todos los demás factores se mantienen fijos). Empezaron con un análisis de diferencia en diferencias, usando log(*durat*) como la variable dependiente. Sea *afchnge* la variable binaria para observaciones después del cambio de política y *highearn* la variable binaria para los asalariados con altos ingresos. Utilizando la base de datos INJURY.RAW, la ecuación estimada, con errores estándar entre paréntesis, es

$$\begin{aligned}\widehat{\log(durat)} = {} & \underset{(0.031)}{1.126} + \underset{(.0447)}{.0077}\, afchnge + \underset{(.047)}{.256}\, highearn \\ & + \underset{(.069)}{.191}\, afchnge \cdot highearn \\ & n = 5{,}626,\ R^2 = .021.\end{aligned} \tag{13.12}$$

Por tanto, $\hat{\delta}_1 = .191$ ($t = 2.77$), lo cual implica que la duración promedio de la indemnización para los asalariados con altos ingresos aumentó alrededor de 19% debido al incremento en el tope de ingresos. El coeficiente de *afchnge* es pequeño y estadísticamente insignificante; como se esperaba, el aumento en el tope de ingresos no tiene efecto sobre la cobertura temporal para los trabajadores de bajos ingresos.

Este es un buen ejemplo de cómo es posible obtener una estimación muy precisa del efecto de un cambio de política, aun cuando no sea posible explicar gran parte del cambio en la variable dependiente. Las variables binarias de la ecuación (13.12) explican sólo 2.1% de la variación en log(*durat*). Esto tiene sentido, ya que claramente hay muchos factores, incluida la gravedad de la lesión, que influyen en la duración que tiene la indemnización laboral. Por fortuna, se cuenta con un tamaño de muestra muy grande, y esto permite obtener un estadístico t significativo.

MVD también agregaron una gran variedad de controles de género, estado civil, edad, industria y tipo de lesión. Esto permite considerar que las personas y las lesiones difieren de forma sistemática en los dos años. El control de estos factores resulta tener un efecto pequeño en la estimación de δ_1. (Vea el ejercicio para computadora C13.4.)

Pregunta 13.2

¿Cómo interpreta el coeficiente y el estadístico t de *highearn* en la ecuación (13.12)?

En ocasiones, los dos grupos están conformados por personas que viven en dos estados vecinos de Estados Unidos. Por ejemplo, para evaluar el impacto del cambio en el impuesto a los cigarros sobre su consumo, se pueden obtener muestras aleatorias de los dos estados para dos años. En el estado A, el grupo control, no hubo modificación en el impuesto a ese producto. En el estado B, el impuesto aumentó (o disminuyó) de un año a otro. La variable resultante debería ser una medida del consumo de los cigarros y la ecuación (13.10) puede estimarse para determinar el efecto del impuesto sobre el consumo de cigarros.

Para un estudio interesante sobre la metodología del experimento natural y varios ejemplos adicionales, vea Meyer (1995).

13.3 Análisis de datos de panel para un periodo de dos años

Vuelva a considerar el análisis de la clase más simple de datos de panel: un corte transversal de individuos, escuelas, empresas, ciudades o cualquier otra cosa para la cual tenga dos años, $t = 1$ y $t = 2$. No es forzoso que estos dos años sean adyacentes, pero $t = 1$ corresponde al primer año. Por ejemplo, el archivo CRIME2.RAW contiene datos de las tasas de delitos y desempleo (entre otros) de 46 ciudades para 1982 y 1987. Por consiguiente, $t = 1$ corresponde a 1982, y $t = 2$ a 1987.

¿Qué pasa si se utiliza el corte transversal de 1987 y se efectúa una regresión simple de la tasa de delitos *crmrte* sobre el de desempleo *unem*? Se obtiene

$$\widehat{crmrte} = 128.38 - 4.16\ unem$$
$$(20.76)\quad (3.42)$$
$$n = 46,\ R^2 = .033.$$

Si se interpreta la ecuación estimada de forma causal, ésta implica que un aumento en la tasa de desempleo *disminuye* la tasa de delitos, y ciertamente no es lo que se esperaba. El coeficiente de *unem* no es significativo estadísticamente a niveles de significancia estándar: lo más que puede decirse es que no se ha encontrado una relación entre estas dos tasas.

Como se ha enfatizado a lo largo de este libro, es probable que esta ecuación de regresión simple sufra problemas de variables omitidas. Una solución posible es tratar de controlar más factores, como la distribución de la edad, la distribución de género, los niveles de educación, los esfuerzos para hacer cumplir la ley, etc., en un análisis de regresión múltiple. Pero es posible que existan muchos factores difíciles de controlar. En el capítulo 9 se mostró cómo la inclusión de *crmrte* de un año anterior, en este caso 1982, puede ayudar a controlar el hecho de que diferentes ciudades tengan tasas de delincuencia distintas. Esta es una manera de usar dos años de datos para la estimación de un efecto causal.

Una opción para usar datos de panel es clasificar los factores no observables, que influyen en la variable dependiente, en dos tipos: aquellos que son constantes y aquellos que varían con el tiempo. Considerando que i es la unidad de corte transversal y t el tiempo, se puede escribir un modelo con una sola variable explicativa observada como

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it},\ t = 1,2. \tag{13.13}$$

En la notación y_{it}, i denota la persona, empresa, ciudad, etc., y t indica el periodo. La variable $d2_t$ es una variable binaria que es igual a cero cuando $t = 1$ y a uno cuando $t = 2$; no cambia en i, razón por la cual no tiene subíndice i. Por consiguiente, el intercepto para $t = 1$ es β_0, y el intercepto para $t = 2$ es $\beta_0 + \delta_0$. Del mismo modo que cuando se utiliza una combinación de cortes transversales, permitir que el intercepto cambie con el tiempo es importante en la mayoría de las aplicaciones. En el ejemplo de la delincuencia, la tendencia secular en Estados Unidos provocará que las tasas de delincuencia de todas las ciudades estadounidenses cambien, quizá marcadamente, en un periodo de cinco años.

La variable a_i captura todos los factores inobservables, constantes en el tiempo, que influyen en y_{it}. (El hecho de que a_i no tenga subíndice t indica que no varía con el tiempo.) Genéricamente, a_i se conoce como **efecto inobservable**. También es común en la práctica encontrar que a_i es llamado un **efecto fijo**, lo cual nos ayuda a recordar que a_i es fijo en el tiempo. El modelo de la ecuación (13.13) se llama **modelo de efectos inobservables** o **modelo de efectos fijos**. En la práctica, usted puede ver que a_i también se refiere a la **heterogeneidad inobservable** (o *heterogeneidad individual, heterogeneidad de la empresa, heterogeneidad de la ciudad*, etcétera).

El error u_{it} con frecuencia se llama **error idiosincrático** o error variable con el tiempo, debido a que representa factores inobservables que cambian con el tiempo e influyen en y_{it}. Estos errores son muy parecidos a los errores en la ecuación de regresión de series de tiempo.

Un modelo de efectos inobservables simple para las tasas de delincuencia para 1982 y 1987 es

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \tag{13.14}$$

donde *d87* es una variable binaria para 1987. Dado que i denota diferentes ciudades, llámese a_i un *efecto inobservable de ciudad* o un *efecto fijo de ciudad*: representa todos los factores que afectan las tasas de delincuencia de una ciudad que no cambian con el tiempo. Las características geográficas, como la ubicación de la ciudad en Estados Unidos, se incluyen en a_i. Es posible que muchos otros factores no sean exactamente constantes, pero podrían ser aproximadamente constantes durante un periodo de cinco años. Estos factores podrían incluir ciertas características demográficas de la población (edad, raza y nivel educativo). Las distintas ciudades pueden tener sus propios métodos para reportar crímenes, y las personas que viven en las ciudades podrían tener diferentes actitudes respecto a la delincuencia; por lo general estos factores cambian lentamente. Por razones históricas, las ciudades pueden tener tasas de delincuencia muy diferentes, y el efecto inobservable a_i captura los factores históricos de manera eficiente.

¿Cómo se debe estimar el parámetro de interés, β_1, con dos años de datos de panel? Una posibilidad es tan solo combinar los dos años y utilizar MCO fundamentalmente como en la sección 13.1. Este método presenta dos inconvenientes, el más importante es que para hacer que MCO generen un estimador consistente de β_1, tendría que suponer que el efecto inobservable a_i, no se correlaciona con x_{it}. Es posible apreciar esto con sólo escribir la ecuación (13.13) como

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}, \quad t = 1,2, \tag{13.15}$$

donde $v_{it} = a_i + u_{it}$ a menudo se denomina **error compuesto**. De lo que se sabe de MCO se debe suponer que v_{it} no se correlaciona con x_{it}, donde $t = 1$ o 2, para que MCO estimen de manera consistente β_1 (y los otros parámetros). Esto es verdadero, ya sea que se utilice un corte transversal o una combinación de dos cortes transversales. Por tanto, aun cuando se supone que el error idiosincrático u_{it} no se correlaciona con x_{it}, la estimación combinada por MCO es sesgada e inconsistente si a_i y x_{it} están correlacionados. Al sesgo resultante en la estimación combinada por MCO a veces se le llama sesgo de heterogeneidad, pero en realidad es sólo un sesgo ocasionado por la omisión de una variable constante en el tiempo.

Pregunta 13.3

Suponga que a_i, u_{i1} y u_{i2} tienen media cero y que no se correlacionan serialmente por parejas. Muestre que $\text{Cov}(v_{i1}, v_{i2}) = \text{Var}(a_i)$, de modo que los errores compuestos se correlacionan serialmente en el tiempo de manera positiva a menos que $a_i = 0$. ¿Qué consecuencia tiene esto en los errores estándar de MCO de la estimación combinada por MCO?

Para ilustrar lo que sucede, se utilizan los datos de CRIME2.RAW con el fin de obtener la estimación combinada por MCO de la ecuación (13.14). Dado que se tienen 46 ciudades y dos años para cada ciudad, existen 92 observaciones en total:

$$\begin{aligned} \widehat{crmrte} &= \underset{(12.74)}{93.42} + \underset{(7.98)}{7.94}\, d87 + \underset{(1.188)}{.427}\, unem \\ n &= 92,\ R^2 = .012. \end{aligned} \tag{13.16}$$

(Cuando se reporta la ecuación estimada, por lo común se omiten los subíndices i y t.) El coeficiente de *unem*, aunque es positivo en la ecuación (13.16), cuenta con un estadístico t muy pequeño. De ahí que el uso de estimadores combinados por MCO para los dos años no difiera de manera sustancial del uso de un corte transversal solo. Esto no sorprende debido a que los estimadores combinados por MCO no resuelven el problema de variables omitidas. (Los errores estándar en esta ecuación son incorrectos por la correlación serial descrita en la pregunta 13.3, pero se ignora esto ya que la estimación combinada por MCO no es el tema de interés aquí.)

En la mayoría de las aplicaciones, la razón principal para reunir datos de panel es permitir que el efecto inobservable, a_i, se correlacione con las variables explicativas. Por ejemplo, en la ecuación de la delincuencia se debe permitir que los factores urbanos no contemplados contenidos en a_i que influyen en la tasa de delincuencia, también se correlacionen con la tasa de desempleo. Resulta que es muy fácil propiciar esto: como a_i es constante en el tiempo se pueden diferenciar los datos a lo largo de los dos años. De manera más precisa, para la observación de corte transversal i, escriba los dos años como

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

Si se resta la *segunda* ecuación de la *primera*, se obtiene

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1}),$$

o

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i, \qquad \textbf{13.17}$$

donde Δ denota el cambio de $t = 1$ a $t = 2$. El efecto inobservable, a_i, no aparece en la ecuación (13.17): se ha "eliminado por diferenciación". Además, el intercepto en la ecuación (13.17) en realidad es el *cambio* en el intercepto de $t = 1$ a $t = 2$.

La ecuación (13.17), la cual es llamada **ecuación en primera diferencia**, es muy simple. Es sólo una ecuación de corte transversal, pero cada variable se diferencia respecto al tiempo. Es posible analizar esta ecuación usando los métodos desarrollados en la parte 1, siempre y cuando se cumplan los supuestos clave. El más importante de ellos es que Δu_i no está serialmente correlacionado con Δx_i. Este supuesto es válido si el error idiosincrático en cada tiempo t, u_{it}, no se correlaciona con la variable explicativa en *ambos* periodos. Se trata de otra versión del supuesto de **exogeneidad estricta** tratado en el capítulo 10 para los modelos de series de tiempo. En particular, este supuesto descarta el caso donde x_{it} es la variable dependiente rezagada, $y_{i,t-1}$. A diferencia del capítulo 10, se permite que x_{it} se correlacione con los efectos inobservables que son constantes en el tiempo. Cuando se obtiene el estimador de MCO de β_1 de (13.17), al estimador resultante se le llama **estimador de primera diferencia**.

En el ejemplo de la delincuencia, suponer que Δu_i y $\Delta unem_i$ no están serialmente correlacionados tal vez sea razonable, pero también puede fallar. Por ejemplo, suponga que el esfuerzo por hacer cumplir la ley (que está en el error idiosincrático) aumenta más en ciudades donde la tasa de desempleo disminuye. Esto puede provocar una correlación negativa entre Δu_i y $\Delta unem_i$, la cual conduciría después a un sesgo en el estimador de MCO. Como es natural, este problema puede resolverse hasta cierto punto al incluir más factores en la ecuación, situación que se tratará posteriormente. Como de costumbre, siempre es posible que no se hayan considerado suficientes factores que varían con el tiempo.

Otra condición crucial es que Δx_i debe tener cierta variación en i. Este requisito no se cumple si la variable explicativa no cambia en el tiempo para cualquier observación de corte transversal, o si cambia la misma cantidad para cada observación. Este no es un problema en el ejemplo de la tasa de delincuencia debido a que la tasa de desempleo cambia en casi todas las ciudades. Pero, si i denota una persona y x_{it} es una variable binaria para el género, $\Delta x_i = 0$ para toda i; desde luego no es posible estimar la ecuación (13.17) por MCO en este caso. En realidad, esto tiene perfecto sentido: dado que se ha permitido que a_i se correlacione con x_{it}, no se puede esperar que el efecto de a_i sobre y_{it} se separe del efecto de cualquier variable que no cambia con el tiempo.

El otro único supuesto necesario para aplicar a los estadísticos usuales de MCO es que la ecuación (13.17) satisfaga el supuesto de homocedasticidad. Esto es razonable en muchos casos, y si no es válido, se sabe cómo probar y corregir la heterocedasticidad usando los métodos estudiados en el capítulo 8. En ocasiones es correcto suponer que la ecuación (13.17) cumple con todos los supuestos del modelo lineal clásico. Los estimadores de MCO son insesgados y toda la inferencia estadística es exacta en estos casos.

Cuando se estima la ecuación (13.17) para el ejemplo de la tasa de delincuencia, se obtiene

$$\Delta\widehat{crmrte} = \underset{(4.70)}{15.40} + \underset{(.88)}{2.22}\,\Delta unem \qquad \textbf{13.18}$$
$$n = 46,\ R^2 = .127,$$

que ahora proporciona una relación positiva y estadísticamente significativa entre las tasas de delincuencia y de desempleo. De modo que hacer la diferenciación para eliminar los efectos constantes en el tiempo hace una gran distinción en este ejemplo. El intercepto de la ecuación (13.18) también revela algo interesante. Aun cuando $\Delta unem = 0$, se proyecta un aumento en la tasa de delincuencia (delitos por cada 1,000 personas) de 15.40. Esto refleja un aumento secular de las tasas de delincuencia en todo Estados Unidos de 1982 a 1987.

Aun cuando no se comenzó con el modelo de los efectos inobservables (13.13), tiene sentido desde un punto de vista intuitivo utilizar las diferencias en el tiempo. En lugar de estimar una relación de corte transversal estándar, que puede tener variables omitidas, dificultando, por ende, las conclusiones *ceteris paribus*, en la ecuación (13.17) se considera de forma explícita cómo los cambios en el tiempo de la variable explicativa influyen en el cambio en y durante el mismo periodo. No obstante, sigue siendo muy útil tener en mente la ecuación (13.13), ya que muestra de forma explícita que es posible estimar el efecto de x_{it} en y_{it}, con a_i fija.

Si bien diferenciar dos años de datos de panel es una forma eficaz de controlar los efectos inobservables, no está exenta de costos. En primer lugar, resulta más difícil reunir conjuntos de datos de panel que un solo conjunto de corte transversal, en especial cuando se trata de personas. Se debe utilizar una encuesta y no perder la pista de las personas. A menudo es difícil ubicar a algunas de ellas para una segunda encuesta. En el caso de unidades como empresas, algunas compañías entran en bancarrota o se fusionan con otras. Es mucho más sencillo obtener datos de panel para escuelas, ciudades, municipios, estados y países.

Aun si se ha reunido un conjunto de datos de panel, la diferenciación empleada para eliminar a_i reduce en gran medida la variación en las variables explicativas. Aunque x_{it} con frecuencia tiene una variación sustancial en el corte transversal para cada t, es posible que Δx_i no cuente con mucha variación. Se sabe del capítulo 3 que una variación pequeña en Δx_i conduce a un error estándar grande para $\hat{\beta}_1$ al estimar la ecuación (13.17) por MCO. Si bien esto puede arreglarse mediante un corte transversal grande, no siempre es posible. Además, la utilización de diferencias más grandes en el tiempo a veces es mejor que servirse de cambios de un año a otro.

Como ejemplo, considere el problema de la estimación del rendimiento de la educación, usando ahora un panel de datos de las personas para dos años. El modelo para la persona i es

$$\log(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}, \quad t = 1, 2,$$

donde a_i contiene la capacidad innata inobservable, que probablemente se correlaciona con $educ_{it}$. Una vez más, se deja que los diferentes interceptos en el tiempo capten las ganancias agregadas en la productividad (y en la inflación, si $wage_{it}$ está en términos nominales). Dado que, por definición, la capacidad innata no cambia con el tiempo, los métodos de datos de panel parecen adecuarse de manera ideal para estimar la rentabilidad de la educación. La ecuación en primera diferencia es

$$\Delta\log(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i, \qquad \textbf{13.19}$$

y se estima por MCO. El problema es que se está interesado en adultos que trabajan y, en el caso de la mayoría de los empleados, la educación no cambia con el tiempo. Si sólo una pequeña fracción de nuestra muestra tiene $\Delta educ_i$ distinta de cero, será difícil obtener un estimador preciso de β_1 a partir de la ecuación (13.19), a menos que se cuente con un tamaño de muestra muy grande. En teoría, el uso de una ecuación en primera diferencia para estimar la rentabilidad de la educación es una buena idea, pero no es muy adecuado para la mayor parte de los conjuntos de datos de panel disponibles.

La adición de diversas variables explicativas no genera dificultades. Se comienza con el modelo de efectos inobservables

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \tag{13.20}$$

para $t = 1$ y 2. Esta ecuación parece más compleja de lo que en realidad es, ya que cada variable explicativa cuenta con tres subíndices. El primero denota el número de observación del corte transversal, el segundo el periodo y el tercero es sólo el número de la variable.

Ejemplo 13.5

[Dormir o trabajar]

Se utilizan los dos años de datos de panel en SLP75_81.RAW, de Biddle y Hamermesh (1990), para estimar el intercambio entre el tiempo dedicado a dormir y a trabajar. En el problema 3.3 se utiliza el corte transversal de 1975. La base de datos de panel de 1975 y 1981 tiene 239 personas, lo que es más reducido que el corte transversal de 1975, el cual comprende más de 700 personas. Un modelo de efectos inobservables para el número total de minutos de sueño por semana (*slpnap*) es

$$\begin{aligned} slpnap_{it} = {} & \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} + \beta_3 marr_{it} \\ & + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + a_i + u_{it}, \quad t = 1, 2. \end{aligned}$$

El efecto inobservable, a_i, se llamaría *efecto individual inobservable* o *efecto individual fijo*. Es potencialmente importante dejar que a_i pueda estar correlacionada con $totwrk_{it}$: es probable que los mismos factores (algunos biológicos) que hacen que la gente duerma más o menos (captados en a_i) estén correlacionados con la cantidad de tiempo dedicado a trabajar. Algunas personas simplemente tienen más energía y esto provoca que duerman menos y trabajen más. La variable *educ* son los años de educación, *marr* es una variable binaria de matrimonio, *yngkid* es una variable binaria que indica la presencia de un hijo de corta edad y *gdhlth* es una variable binaria de "buena salud". Observe que no se incluyó género o raza (como se hizo en el análisis de corte transversal), ya que esto no se modifica con el paso del tiempo; son parte de a_i. El principal interés se centra en β_1.

La diferenciación entre los dos años da la ecuación estimable

$$\begin{aligned} \Delta slpnap_i = {} & \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i \\ & + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i. \end{aligned}$$

Si se supone que el cambio en el error idiosincrático, Δu_i, no se correlaciona con los cambios en todas las variables explicativas, se obtienen estimadores consistentes por MCO. Esto da

$$\begin{aligned} \Delta \widehat{slpnap} = {} & \underset{(45.87)}{-92.63} - \underset{(.036)}{.227}\ \Delta totwrk - \underset{(48.759)}{.024}\ \Delta educ \\ & + \underset{(92.86)}{104.21}\ \Delta marr + \underset{(87.65)}{94.67}\ \Delta yngkid + \underset{(76.60)}{87.58}\ \Delta gdhlth \\ n = {} & 239,\ R^2 = .150. \end{aligned} \tag{13.21}$$

El coeficiente de $\Delta totwrk$ indica el intercambio entre horas de dormir y trabajar: con los otros factores fijos, una hora más de trabajo se asocia con .227(60) = 13.62 minutos menos de sueño. El estadístico t (−6.31) es muy significativo. Ningún otro estimador, salvo el del intercepto, es estadísticamente diferente de cero. La prueba F de significancia conjunta para todas las variables, excepto $\Delta totwrk$ da un valor-p = .49, lo que quiere decir que no son conjuntamente significativas a cualquier nivel razonable de significancia y que, por tanto, podrían omitirse de la ecuación.

El error estándar de $\Delta educ$ es grande en especial con relación a la estimación. Este fenómeno fue descrito antes en la ecuación del salario. En la muestra de 239 personas, 183 (76.6%) no han cambiado su nivel educativo durante el periodo de seis años; el 90% tiene un cambio en la escolaridad de cuando más un año. Como lo refleja el error estándar sumamente grande de $\hat{\beta}_2$, no hay la suficiente variación en la educación para estimar β_2 con cierta precisión. De cualquier modo, $\hat{\beta}_2$ prácticamente es muy pequeña.

Los datos de panel también se pueden utilizar para estimar modelos de rezagos distribuidos finitos. Aun cuando se especifique la ecuación para únicamente dos años, es necesario recabar más años de datos para obtener las variables explicativas rezagadas. El siguiente es un ejemplo sencillo de esto.

Ejemplo 13.6

[Rezagos distribuidos de la tasa de delincuencia sobre la tasa de casos resueltos]

Eide (1994) utiliza datos de panel de distritos policiacos en Noruega para estimar un modelo de rezagos distribuidos para las tasas de delincuencia. La variable explicativa única es el "porcentaje de casos resueltos" (*clrprc*), es decir, el porcentaje de delitos que dieron por resultado una condena. Los datos sobre la tasa de delincuencia pertenecen a los años 1972 y 1978. Siguiendo a Eide, se hace el rezago de *clrprc* para uno y dos años: es probable que la tasa de casos resueltos pasados ejerza un efecto disuasivo en los delitos actuales. Esto conduce al modelo de efectos inobservables siguiente para los dos años:

$$\log(crime_{it}) = \beta_0 + \delta_0 d78_t + \beta_1 clrprc_{i,t-1} + \beta_2 clrprc_{i,t-2} + a_i + u_{it}.$$

Cuando se diferencia la ecuación y se estima con los datos de CRIME3.RAW, se obtiene

$$\begin{aligned}\Delta\widehat{\log(crime)} &= \underset{(.064)}{.086} - \underset{(.0047)}{.0040}\,\Delta clrprc_{-1} - \underset{(.0052)}{.0132}\,\Delta clrprc_{-2} \\ n &= 53,\ R^2 = .193,\ \bar{R}^2 = .161.\end{aligned} \tag{13.22}$$

El segundo rezago es negativo y significativo, lo que implica que un mayor porcentaje de casos resueltos hace dos años disuadiría los delitos de este año. En concreto, un aumento de 10 puntos porcentuales en *clrprc* hace dos años conduciría a una disminución estimada de 13.2% en la tasa de delincuencia de este año. Esto sugiere que el uso de más recursos para resolver delitos y recibir condenas reduce la delincuencia a futuro.

Organización de los datos de panel

Al utilizar datos de panel en un estudio econométrico, es importante conocer la forma en que éstos deben almacenarse. Se debe ser cuidadoso al ordenarlos de modo que los distintos periodos para la misma unidad de corte transversal (persona, empresa, ciudad, etc.) se vinculen con facilidad. En concreto, suponga que el conjunto de datos es sobre ciudades para dos años diferentes. Para la mayoría de los propósitos, la mejor manera de introducir los datos es contar con dos registros para cada ciudad, uno para cada año: el primer registro corresponde al primer año y el segundo al último. Estos dos registros deben ser contiguos. Por tanto, un conjunto de datos para 100 ciudades y dos años contendrá 200 registros. Los dos primeros registros son para la primera ciudad de la muestra, los dos siguientes para la segunda, etc. (En la tabla 1.5 del capítulo 1 se proporciona un ejemplo.) Esto facilita la elaboración de diferencias para almacenarlas en el segundo registro de cada ciudad y realizar un análisis de una combinación de corte transversal, que puede compararse con la estimación obtenida de la diferenciación.

La mayor parte de las bases de datos de panel de dos periodos que acompañan este libro están almacenadas de esta manera (por ejemplo, CRIME2.RAW, CRIME3.RAW, GPA3.RAW, LOWBRTH.RAW y RENTAL.RAW). Se utiliza una extensión directa de este esquema para bases de datos de panel con más de dos periodos.

Una segunda manera de organizar dos periodos de datos de panel es contar sólo con un registro por unidad de corte transversal. Esto exige dos entradas para cada variable, una para cada periodo. Los datos de panel de SLP75_81.RAW están organizados de este modo. Cada persona tiene datos en las variables *slpnap75*, *slpnap81*, *totwrk75*, *totwrk81*, etc. Crear las diferencias de 1975 a 1981 es fácil. Otras bases de datos de panel con esta estructura son TRAFFIC1.RAW y VOTE2.RAW. No obstante, un inconveniente de este tipo de registro es que no permite un análisis combinado mediante MCO de los dos periodos sobre los datos originales. Además, este método de organización no funciona para bases de datos de panel con más de dos periodos, caso que se considera en la sección 13.5.

13.4 Análisis de políticas con datos de panel de dos periodos

Los conjuntos de datos de panel son muy útiles para el análisis de políticas, en particular para la evaluación de programas. En el sistema de evaluación de programas más sencillo, se obtiene una muestra de personas, empresas, ciudades, etc. en el primer periodo. Algunas de estas unidades toman parte luego en un determinado programa en un periodo posterior; aquellas que no lo hacen pertenecen al grupo de control. Esto es similar al estudio del experimento natural presentado antes, pero con una diferencia importante: en cada periodo aparecen las *mismas* unidades de corte transversal.

A manera de ejemplo, suponga que se desea evaluar el efecto del programa de capacitación laboral de Michigan sobre la productividad de los trabajadores de las empresas manufactureras (vea también el ejercicio para computadora C9.3). Sea $scrap_{it}$ la tasa de desperdicio de la industria i durante el año t (el número de artículos, por cada 100, que deben desecharse por estar defectuosos). Sea $grant_{it}$ un indicador binario igual a uno si la empresa i en el año t recibió un subsidio para capacitación laboral. Para los años 1987 y 1988, el modelo es

$$scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \quad t = 1, 2, \tag{13.23}$$

donde $y88_t$ es una variable binaria de 1988 y a_i es el *efecto inobservable de la empresa* o *efecto fijo de la empresa*. El efecto inobservable contiene factores como la capacidad promedio del empleado, el capital y las habilidades administrativas, que son aproximadamente constantes durante un periodo de dos años. La preocupación es que a_i se relacione de forma sistemática con el hecho de si la empresa recibe o no un subsidio. Por ejemplo, los administradores del programa podrían dar prioridad a las empresas cuyos trabajadores tienen habilidades menores. O bien, puede ocurrir el problema contrario: para que el programa de capacitación parezca eficiente, los administradores tal vez den subsidios a los empleadores que tengan a su cargo a los trabajadores más productivos. En realidad, en este programa particular, los subsidios se otorgan con base en que el primero en llegar es el primero en ser atendido. Pero el hecho de que una empresa haya solicitado antes un subsidio podría correlacionarse con la productividad de los trabajadores. En ese caso, un análisis que usa un solo corte transversal o sólo una combinación de cortes transversales producirá estimadores sesgados e inconsistentes.

La diferenciación para eliminar a_i da

$$\Delta scrap_i = \delta_0 + \beta_1 \Delta grant_i + \Delta u_i. \tag{13.24}$$

Por consiguiente, sencillamente se efectúa la regresión del cambio en la tasa de desperdicio sobre el cambio en el indicador de subsidios. Como ninguna empresa recibió subsidios en 1987, $grant_{i1} = 0$ para toda i, y por tanto $\Delta grant_i = grant_{i2} - grant_{i1} = grant_{i2}$, lo cual tan solo indica si la empresa recibió un subsidio en 1988. Sin embargo, por lo general es importante diferenciar todas las variables (variables binarias incluidas) ya que es necesario para eliminar a_i en el modelo de efectos inobservables (13.23).

La estimación de la ecuación en primera diferencia usando la base de datos JTRAIN.RAW es

$$\begin{aligned}\Delta \widehat{scrap} &= -.564 - .739\ \Delta grant \\ &\quad\ (.405)\quad (.683) \\ n &= 54,\ R^2 = .022.\end{aligned}$$

Así, se estima que haber tenido un subsidio para capacitación laboral redujo en promedio $-.739$ la tasa de desperdicio. Pero la estimación no es estadísticamente diferente de cero.

Se obtienen resultados más contundentes al usar log($scrap$) y estimar el efecto en porcentaje:

$$\begin{aligned}\widehat{\Delta \log(scrap)} &= -.057 - .317\ \Delta grant \\ &\quad\ (.097)\quad (.164) \\ n &= 54,\ R^2 = .067.\end{aligned}$$

Se estima que tener un subsidio de capacitación laboral reduce la tasa de desperdicio en cerca de 27.2%. [Esta estimación se obtiene a partir de la ecuación (7.10): $\exp(-.317) - 1 \approx -.272$.] El estadístico t es alrededor de -1.93, lo cual es marginalmente significativo. En comparación, la estimación combinada por MCO de log($scrap$) sobre $y88$ y $grant$ da $\hat{\beta}_1 = .057$ (error estándar $=$ .431). Por consiguiente, no se encuentra una relación significativa entre la tasa de desperdicio y el subsidio para capacitación laboral. Dado que esto difiere en gran medida de las estimaciones en primera diferencia, lo anterior sugiere que es más probable que las empresas con empleados menos capaces reciban un subsidio.

Es conveniente estudiar el modelo de evaluación de un programa de una manera más general. Sea y_{it} una variable resultante y $prog_{it}$ una variable binaria de participación en el programa. El modelo de efectos inobservables más simple es

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it}. \tag{13.25}$$

Si la participación en el programa sólo ocurrió en el segundo periodo, entonces el estimador de MCO de β_1 en la ecuación diferenciada se representa de una manera muy sencilla:

$$\hat{\beta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}. \tag{13.26}$$

Es decir, se calcula el cambio promedio en y entre los dos periodos para los grupos de tratamiento y control. Entonces, $\hat{\beta}_1$ es la diferencia de estos grupos. Esta es la versión de datos de panel del estimador de diferencia en diferencias de la ecuación (13.11) para la combinación de dos cortes transversales. Con los datos de panel, se cuenta con una ventaja potencialmente importante: es posible diferenciar y en el tiempo para las *mismas* unidades de corte transversal. Esto permite controlar los efectos específicos de la persona, empresa o ciudad, como lo deja en claro el modelo de la ecuación (13.25).

Si la participación en el programa tiene lugar en ambos periodos, $\hat{\beta}_1$ no se escribe como en la ecuación (13.26), pero se interpreta de la misma manera: se trata del cambio en el valor medio de y debido a la participación en el programa.

Controlar los factores que varían con el tiempo no modifica nada respecto a la significancia. Simplemente se diferencian tales variables y se incluyen con $\Delta prog$. Esto permite controlar las variables que cambian con el tiempo, que podrían estar correlacionadas con la participación en el programa.

El mismo método de diferenciación funciona para analizar los efectos de cualquier política que varíe de una ciudad o estado a otro. Véase el siguiente ejemplo sencillo.

Ejemplo 13.7

[Efecto de las leyes de conducir en estado de ebriedad sobre las muertes en accidentes de tráfico]

Muchos estados de Estados Unidos han adoptado diferentes políticas que intentan poner freno a los conductores en estado de ebriedad. Dos tipos de leyes que se estudiarán aquí son las *leyes sobre recipientes abiertos*, que vuelve ilegal para los pasajeros portar recipientes destapados de bebidas alcohólicas, y las *leyes administrativas per se*, que permiten que los tribunales suspendan las licencias de manejo después de que un conductor es arrestado por conducir en estado de ebriedad, pero antes de que se le dicte una sentencia. Un posible análisis es utilizar un solo corte transversal de los estados para hacer la regresión de las muertes por accidentes de tráfico (o las relacionadas con la conducción en estado de ebriedad) sobre variables binarias que indiquen la presencia de cada ley. Es probable que esto no funcione adecuadamente, porque los estados deciden, mediante procesos legislativos, si necesitan tales leyes. Por ende, la presencia de las leyes quizá se relacione con el promedio de muertes por conducir en estado de ebriedad en los últimos años. En un análisis más convincente se utilizan datos de panel en un periodo en el que algunos estados adoptaron nuevas leyes (y otros tal vez hayan revocado las existentes). El archivo TRAFFIC1.RAW contiene datos de 1985 y 1990 para los 50 estados y el Distrito de Columbia. La variable dependiente es el número de muertes por accidentes de tránsito por cada 100 millones de millas conducidas (*dthrte*). En 1985, 19 estados contaban con leyes sobre recipientes abiertos, mientras que 22 tenían dichas leyes en 1990. En 1985, 21 estados contaban con leyes *per se*; la cifra creció a 29 para 1990.

Al utilizar MCO luego de hacer una primera diferenciación se obtiene lo siguiente

$$
\begin{aligned}
\widehat{\Delta dthrte} &= -.497 - .420\ \Delta open - .151\ \Delta admn \\
&\quad\ \ (.052)\quad (.206) \qquad\qquad (.117) \\
n &= 51,\ R^2 = .119.
\end{aligned}
\tag{13.27}
$$

Las estimaciones sugieren que la adopción de una ley sobre recipientes abiertos disminuyó la tasa de muertes por accidentes de tránsito en .42, un efecto no trivial dado que la tasa promedio de muertes en 1985 fue de 2.7 con una desviación estándar de alrededor de .6. La estimación es estadísticamente significativa al nivel de 5% contra una alternativa de dos colas. La ley administrativa *per se* tiene un efecto menor y su estadístico t es de sólo -1.29; pero la estimación tiene el signo que se esperaba. El intercepto en esta ecuación muestra que las muertes por accidentes de tráfico disminuyen considerablemente para todos los estados durante el periodo de cinco años, ya sea que hubiera cambios en alguna ley o no. Los estados que adoptaron una ley de recipientes abiertos

Pregunta 13.4

En el ejemplo 13.7, $\Delta admn = -1$ para el estado de Washington. Explique lo que esto significa.

durante este periodo vieron una disminución mucho mayor, en promedio, en las tasas de muertes por accidentes de tráfico.

Existen otras leyes que también podrían afectar las muertes por accidentes de tráfico, como las leyes del uso del cinturón de seguridad, las leyes del uso del casco al conducir una motocicleta y los límites de velocidad máxima. Asimismo, se podría querer controlar las distribuciones de edad y género, así como las mediciones de qué tan influyente es una organización como Madres contra la Conducción en Estado de Ebriedad en el estado.

13.5 Diferenciación con más de dos periodos

También es posible utilizar la diferenciación con más de dos periodos. Como ejemplo, suponga que se tienen N personas y $T = 3$ periodos para cada persona. Un modelo general de efectos fijos es

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad \mathbf{13.28}$$

para $t = 1$, 2 y 3. (El número total de observaciones es por tanto $3N$.) Advierta que ahora se incluyeron dos variables binarias de tiempo además del intercepto. Es buena idea permitir un intercepto separado para cada periodo, en particular cuando se tiene un número pequeño de ellos. El periodo base, como siempre, es $t = 1$. El intercepto para el segundo periodo es $\delta_1 + \delta_2$, etc. Lo de principal interés es $\beta_1, \beta_2, \ldots, \beta_k$. Si el efecto inobservable a_i se correlaciona con cualquiera de las variables explicativas, entonces la estimación combinada por MCO en los tres años de datos da como resultado estimaciones sesgadas e inconsistentes.

El supuesto clave es que los errores idiosincráticos no están serialmente correlacionados con la variable explicativa en cada periodo:

$$\text{Cov}(x_{itj}, u_{is}) = 0, \text{ para toda } t, s \text{ y } j. \quad \mathbf{13.29}$$

Esto significa que las variables explicativas son *estrictamente exógenas* después de que se elimina el efecto inobservable, a_i. (El supuesto de exogeneidad estricta establecido en términos de una esperanza condicional igual a cero se proporciona en el apéndice del capítulo). El supuesto (13.29) descarta casos donde las variables explicativas futuras reaccionan a los cambios actuales en los errores idiosincráticos, como debe ser el caso si x_{itj} es una variable dependiente rezagada. Si se ha omitido una variable importante que varía con el tiempo, entonces por lo general se viola (13.29). El error de medición en una o más variables explicativas puede provocar que (13.29) sea falso, precisamente como en el capítulo 9. En los capítulos 15 y 16, se estudiará lo que puede hacerse en tales casos.

Si a_i se correlaciona con x_{itj}, entonces x_{itj} se correlacionará con el error *compuesto*, $v_{it} = a_i + u_{it}$, bajo (13.29). Puede eliminarse a_i al diferenciar los periodos adyacentes. En el caso de $T = 3$ se resta el periodo uno del periodo dos y el periodo dos del periodo tres. Esto da

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad \mathbf{13.30}$$

para $t = 2$ y 3. No se tiene una ecuación diferenciada para $t = 1$ debido a que no hay nada que restar de la ecuación en $t = 1$ Ahora bien, la ecuación (13.30) representa *dos* periodos para cada persona de la muestra. Si esta ecuación satisface los supuestos del modelo lineal clásico, entonces la estimación combinada por MCO arroja estimadores insesgados, y los estadísticos t y F son válidos para probar hipótesis. También es posible recurrir a resultados asintóticos. El requisito importante para que los estimadores de MCO sean consistentes es que Δu_{it} no se correlacione con Δx_{itj} para toda j y $t = 2$ y 3. Esta es la extensión natural del caso de dos periodos.

Observe cómo la ecuación (13.30) contiene las diferencias en las variables binarias anuales, $d2_t$ y $d3_t$. Para $t = 2$, $\Delta d2_t = 1$ y $\Delta d3_t = 0$; para $t = 3$, $\Delta d2_t = -1$ y $\Delta d3_t = 1$. Por consiguiente, la ecuación (13.30) no contiene un intercepto, lo cual es un inconveniente para ciertos propósitos, incluido el cálculo de la R-cuadrada. A menos que los interceptos temporales del modelo original (13.28) sean de interés directamente, lo cual es raro, es mejor estimar la ecuación en primera diferencia con un intercepto y una sola variable binaria de periodo, por lo general para el tercero. En otras palabras, la ecuación se vuelve

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + \dots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad \text{para } t = 2 \text{ y } 3.$$

Las estimaciones de β_j son idénticas en cada formulación.

Con más de tres periodos, las cosas son similares. Si se cuenta con los mismos periodos T para cada una de las unidades de corte transversal N, se dice que el conjunto de datos es un **panel balanceado**: se tienen los mismos periodos para todas las personas, empresas, ciudades, etc. Cuando T es pequeña en relación con N, se debe incluir una variable binaria para cada periodo, para representar los cambios seculares que no hayan sido considerados. Por tanto, luego de hacer la primera diferenciación, la ecuación se parece a lo siguiente

$$\begin{aligned}\Delta y_{it} = {} & \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + \dots + \alpha_T dT_t + \beta_1 \Delta x_{it1} + \dots \\ & + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad t = 2, 3, \dots, T,\end{aligned} \tag{13.31}$$

donde se tienen $T - 1$ periodos para cada unidad i en la ecuación en primera diferencia. El número total de observaciones es $N(T - 1)$.

Es sencillo obtener la estimación combinada por MCO de la ecuación (13.31), siempre y cuando las observaciones se hayan organizado de forma adecuada y la diferenciación se haya realizado con cuidado. Para facilitar la primera diferenciación, el archivo de datos debe contar con NT registros. Los primeros T registros son para la primera observación de corte transversal, ordenada de manera cronológica; los segundos T registros son para la segunda observación de corte transversal, dispuesta en forma cronológica; y así sucesivamente. Luego, se calculan las diferencias, con el cambio de $t - 1$ a t almacenado en el registro de tiempo t. Por consiguiente, las diferencias para $t = 1$ deben carecer de valores para todas las observaciones de corte transversal. De no hacer esto, se corre el riesgo de utilizar observaciones falsas en el análisis de regresión. Una observación no válida se genera cuando la última observación de la persona, por ejemplo para la persona $i - 1$, se sustrae de la primera observación de la persona i. Si se realiza la regresión sobre los datos diferenciados, y se reportan NT o $NT - 1$ observaciones, entonces se olvidó establecer las observaciones $t = 1$ como faltantes.

Cuando se utilizan más de dos periodos, se debe suponer que Δu_{it} no se correlaciona en el tiempo para que los errores estándar usuales y los estadísticos de prueba sean válidos. Esta premisa a veces es razonable, pero no se aplica cuando se supone que los errores idiosincráticos originales, u_{it}, no se correlacionan serialmente en el tiempo (supuesto que será utilizado en el capítulo 14). De hecho, si se supone que u_{it} no se correlaciona serialmente y tiene varianza constante, entonces puede demostrarse que la correlación entre Δu_{it} y $\Delta u_{i,t+1}$ es $-.5$. Si u_{it} sigue un modelo AR(1) estable, entonces Δu_{it} tendrá correlación serial. Sólo cuando u_{it} sigue una caminata aleatoria, Δu_{it} no se correlacionará serialmente.

Es fácil probar la correlación serial en la ecuación en primera diferencia. Sea $r_{it} = \Delta u_{it}$ la primera diferencia del error original. Si r_{it} sigue el modelo AR(1): $r_{it} = \rho r_{i,t-1} + e_{it}$, entonces es posible probar con facilidad H_0: $\rho = 0$. En primer lugar, se obtienen las estimaciones combinadas por MCO de la ecuación (13.31) y se obtienen los residuales, $\hat{r}_{it}$.

Luego, se efectúa la regresión combinada por MCO de $\hat{r}_{it}$ sobre $\hat{r}_{i,t-1}$, para $t = 3, ..., T$, $i = 1, ..., N$, y se calcula una prueba t estándar para el coeficiente de $\hat{r}_{i,t-1}$. (O se puede hacer el estadístico t robusto a la heterocedasticidad.) El coeficiente $\hat{\rho}$ de $\hat{r}_{i,t-1}$ es un estimador consistente de ρ. Dado que se ha utilizado el residual rezagado, se pierde otro periodo. Por ejemplo, si se inicia con $T = 3$, la ecuación diferenciada tiene dos periodos y la prueba para la correlación serial es sólo una regresión de corte transversal de los residuales del tercer periodo sobre los residuales del segundo periodo. Más adelante se dará un ejemplo.

Se puede corregir la presencia de la correlación serial AR(1) en r_{it} utilizando estimadores de MCG factibles. En esencia, dentro de cada observación de corte transversal se utilizaría la transformación de Prais-Winsten con base en el $\hat{\rho}$ descrito en el párrafo anterior. (Desde luego, es preferible Prais-Winsten a Cochrane-Orcutt aquí, ya que omitir ahora el primer periodo significaría perder N observaciones de corte transversal.) Por desgracia, los paquetes estándar que realizan las correcciones AR(1) para las regresiones de series de tiempo no funcionarán. Los métodos estándar de Cochrane-Orcutt o Prais-Winsten tratarán las observaciones como si siguieran un proceso AR(1) a lo largo de i y t; esto no tiene sentido, ya que se está suponiendo que las observaciones son independientes entre las i. La corrección de los errores estándar de MCO que permiten formas arbitrarias de correlación serial (y de heterocedasticidad) se pueden calcular cuando N es grande (y N debe ser notablemente mayor que T). Un tratamiento detallado de estos temas está más allá del alcance de este libro [vea Wooldridge (2002, capítulo 10)], pero son fáciles de calcular en ciertos paquetes de regresión.

Pregunta 13.5

¿La correlación serial en Δu_{it} hace que el estimador de primera diferencia sea sesgado e inconsistente? ¿Por qué la correlación serial es un motivo de preocupación?

Si no hay correlación serial en los errores, los métodos usuales para tratar la heterocedasticidad son válidos. Se pueden utilizar las pruebas de Breusch-Pagan y White del capítulo 8 para la heterocedasticidad y también calcular los errores estándar robustos.

La diferenciación de más de dos años de los datos de panel es muy útil para el análisis de políticas, como se aprecia en el siguiente ejemplo.

Ejemplo 13.8

[Efecto de las zonas empresariales en los reclamos de desempleo]

Papke (1994) estudió el efecto del programa de zonas empresariales (EZ) de Indiana en los reclamos del seguro de desempleo. Analizó 22 ciudades de Indiana de 1980 a 1988. Se designaron seis zonas empresariales en 1984 y cuatro más en 1985. Doce de las ciudades de la muestra no recibieron ninguna zona empresarial en este periodo; sirvieron como grupo control.

Un modelo sencillo de evaluación de políticas es

$$\log(uclms_{it}) = \theta_t + \beta_1 ez_{it} + a_i + u_{it},$$

donde $uclms_{it}$ es el número de reclamos del seguro de desempleo presentados durante el año t en la ciudad i. El parámetro θ_t denota sencillamente un intercepto distinto para cada periodo. En general, los reclamos del seguro de desempleo descendieron en todos los estados en este periodo y esto debía reflejarse en los diferentes interceptos anuales. La variable binaria ez_{it} es igual a uno si la ciudad i en el momento t era una zona empresarial; β_1 es la variable de interés. El efecto inobservable a_i representa los factores fijos que influyen en el ambiente económico de la ciudad i. Como la designación de la zona empresarial no se determinó aleatoriamente —estas zonas por lo general están económicamente deprimidas—, es probable que ez_{it} y a_i se correlacionen positivamente (una a_i elevada significa un mayor número de reclamos del seguro de desempleo, lo que conduce a una mayor probabilidad de que se le asigne una zona empresarial). Por esta razón, se debe diferenciar la ecuación para eliminar a_i:

$$\Delta\log(uclms_{it}) = \alpha_0 + \alpha_1 d82_t + \ldots + \alpha_7 d88_t + \beta_1 \Delta ez_{it} + \Delta u_{it}. \tag{13.32}$$

La variable dependiente en esta ecuación, el cambio en $\log(uclms_{it})$, es la tasa de crecimiento anual aproximada de los reclamos del seguro de desempleo del año $t - 1$ a t. Es posible estimar esta ecuación para los años 1981 a 1988 utilizando la base de datos EZUNEM.RAW; el tamaño total de la muestra es $22 \cdot 8 = 176$. La estimación de β_1 es $\hat{\beta}_1 = -.182$ (error estándar $= .078$). Por tanto, parece que la presencia de una EZ provoca que los reclamos del seguro de desempleo se reduzcan aproximadamente 16.6% $[\exp(-.182) - 1 \approx -.166]$. Este es un efecto económicamente grande y estadísticamente significativo.

No hay evidencia de heterocedasticidad en la ecuación: la prueba F de Breusch-Pagan produce $F = .85$, valor-$p = .557$. Sin embargo, cuando se agregan los residuales de MCO rezagados a la ecuación diferenciada (y se pierde el año 1981), se obtiene $\hat{\rho} = -.197$ ($t = -2.44$), de modo que hay una mínima evidencia de correlación serial negativa en los errores en primera diferencia. A diferencia de la correlación serial positiva, los errores estándar usuales de MCO tal vez no subestimen en gran medida a los errores estándar correctos cuando los errores se correlacionan negativamente (vea la sección 12.1). Así, probablemente no se vea afectada la significancia de la variable binaria de la zona empresarial.

Ejemplo 13.9

[Tasas de delincuencia por condado en Carolina del Norte]

Cornwell y Trumbull (1994) utilizaron datos de 90 condados en Carolina del Norte, de los años 1981 a 1987, para estimar un modelo de efectos inobservables de la delincuencia; los datos se hallan en CRIME4.RAW. Aquí, se estima una versión más sencilla de su modelo y se diferencia la ecuación en el tiempo para eliminar a_i, el efecto inobservable. (Cornwell y Trumbull emplean una transformación diferente, que se tratará en el capítulo 14.) Diversos factores, entre los que se cuentan la ubicación geográfica, las actitudes hacia la delincuencia, los registros históricos y las normas de denuncia del crimen, podrían estar contenidos en a_i. La tasa de delincuencia (*crmrte*) es el número de delitos por persona, *prbarr* es la probabilidad estimada de arresto, *prbconv* es la probabilidad estimada de condena (dado un arresto), *prbpris* es la probabilidad de cumplir una sentencia en prisión (dada una condena), *avgsen* es la duración promedio de la sentencia, y *polpc* es el número de policías *per cápita*. Como es usual en los estudios econométricos, se utilizan los logaritmos de todas las variables con la finalidad de estimar las elasticidades. También se incluye un conjunto completo de variables binarias anuales para controlar las tendencias estatales en las tasas de delincuencia. Es posible utilizar los años de 1982 a 1987 para estimar la ecuación en diferencias. Las cantidades entre paréntesis son los errores estándar usuales de MCO y las que están entre corchetes son los errores estándar robustos tanto a la correlación serial como a la heterocedasticidad:

$$
\begin{aligned}
\widehat{\Delta\log(crmrte)} = {} & \underset{[.014]}{\underset{(.017)}{.008}} - \underset{[.022]}{\underset{(.024)}{.100}}\, d83 - \underset{[.020]}{\underset{(.024)}{.048}}\, d84 - \underset{[.025]}{\underset{(.023)}{.005}}\, d85 \\
& + \underset{[.021]}{\underset{(.024)}{.028}}\, d86 + \underset{[.024]}{\underset{(.024)}{.041}}\, d87 - \underset{[.056]}{\underset{(.030)}{.327}}\, \Delta\log(prbarr) \\
& - \underset{[.040]}{\underset{(.018)}{.238}}\, \Delta\log(prbconv) - \underset{[.046]}{\underset{(.026)}{.165}}\, \Delta\log(prbpris) \\
& - \underset{[.026]}{\underset{(.022)}{.022}}\, \Delta\log(avgsen) + \underset{[.103]}{\underset{(.027)}{.398}}\, \Delta\log(polpc) \\
n = {} & 540,\ R^2 = .433,\ \bar{R}^2 = .422.
\end{aligned}
\tag{13.33}
$$

Las tres variables de probabilidad (arresto, condena y tiempo cumplido en prisión) tienen el signo esperado y todas son estadísticamente significativas. Por ejemplo, se predice que un aumento de 1% en la probabilidad de arresto disminuirá la tasa de delincuencia en alrededor de .33%. La variable de sentencia promedio muestra un efecto disuasivo modesto, pero este no es estadísticamente significativo.

El coeficiente de la variable de policías *per cápita* es algo sorprendente y es una característica de la mayor parte de los estudios que buscan explicar las tasas de delincuencia. Si se interpreta de manera causal, indica que un aumento de 1% en los policías *per cápita incrementa* las tasas de delincuencia aproximadamente en .4%. (El estadístico t usual es muy grande, de casi 15.) Resulta difícil creer que contar con más policías genera más delincuencia. ¿Qué está pasando aquí? Hay al menos dos posibilidades. En primer lugar, la variable de la tasa de delincuencia se calcula a partir de los delitos *reportados*. Podría ser que, cuando hay policía adicional, se reportan más delitos. En segundo lugar, la variable de los policías podría ser endógena en la ecuación por otras razones: los condados tal vez amplíen la fuerza policiaca cuando esperan que aumenten las tasas de delincuencia. En este caso, la ecuación (13.33) no puede interpretarse de manera causal. En los capítulos 15 y 16, se tratarán modelos y métodos de estimación que dan cuenta de esta forma adicional de endogeneidad.

El caso especial de la prueba de White para heterocedasticidad de la sección 8.3 da $F = 75.48$ y un valor-$p = .0000$, de modo que hay evidencia contundente de heterocedasticidad. (Técnicamente, la prueba no es válida si existe también correlación serial, pero resulta muy sugerente.) La prueba de la correlación serial AR(1) produce $\hat{\rho} = -.233$, $t = -4.77$, de modo que existe correlación serial negativa. Los errores estándar entre corchetes se ajustan para la correlación serial y la heterocedasticidad. [No se darán los pormenores de esto, los cálculos son parecidos a aquellos descritos en la sección 12.5 y los realizan muchos paquetes econométricos. Vea Wooldridge (2002, capítulo 10) para un análisis más detallado.] Ninguna variable pierde su significancia estadística, pero los estadísticos t sobre las variables de disuasión se vuelven notablemente más pequeños. Por ejemplo, el estadístico t de la probabilidad de condena va de -13.22 utilizando los errores estándar usuales de MCO, a -6.10 empleando el error estándar completamente robusto. De manera equivalente, los intervalos de confianza elaborados con ayuda de los errores estándar robustos serán mucho más amplios, como es propio, que aquéllos basados en los errores estándar usuales de MCO.

Como es natural, es posible aplicar la prueba de Chow a los modelos de datos de panel estimados en primera diferencia. Al igual que en el caso de los cortes transversales combinados, rara vez se querrá probar si los interceptos son diferentes. Es mucho más interesante probar si los parámetros de pendiente han cambiando con el tiempo y se pueden realizar estas pruebas fácilmente al hacer interactuar las variables explicativas de interés con las variables binarias de tiempo. Resulta interesante que aun cuando no es posible estimar las pendientes en las variables que no cambian con el tiempo, es posible probar si los efectos parciales de las variables constantes en el tiempo han cambiado en el transcurso del tiempo. Como ejemplo, suponga que se observan tres años de datos en una muestra aleatoria de personas que trabajan en 2000, 2002 y 2004, y se especifica el modelo (para el logaritmo del salario, *lwage*),

$$
\begin{aligned}
lwage_{it} = \beta_0 + \delta_1 d02_t + \delta_2 d04_t + \beta_1 female_i + \gamma_1 d02_t\, female_i \\
+ \gamma_2 d04_t\, female_i + \mathbf{z}_{it}\boldsymbol{\lambda} + a_i + u_{it},
\end{aligned}
$$

donde $\mathbf{z}_{it}\boldsymbol{\lambda}$ representa otras variables explicativas incluidas en el modelo y sus coeficientes. Cuando se hace la primera diferenciación, se elimina el intercepto para 2000, β_0, y también la diferencia de salario por género para 2000, β_1. Sin embargo, el cambio en $d02_t\, female_i$ es $(\Delta d01_t) female_i$, el cual no se elimina. Por consiguiente, se puede estimar cómo la diferencia de salario por género ha cambiado en 2002 y 2004 respecto a 2000, y es posible probar si $\gamma_1 = 0$ o $\gamma_2 = 0$, o ambas. También se podría preguntar si la prima salarial por afiliación a un sindicato ha cambiado con el tiempo, en cuyo caso se incluye en el modelo $union_{it}$, $d02_t union_{it}$ y $d04_t union_{it}$. Los coeficientes de todas estas variables explicativas pueden estimarse debido a que se supone que $union_{it}$ tendría alguna variación en el tiempo.

Si uno trata de estimar un modelo que contiene interacciones calculando las diferencias manualmente puede tener dificultades. Por ejemplo, en la ecuación anterior con el estado de afiliación al sindicato, sencillamente se deben diferenciar los términos de interacción, $d02_t union_{it}$ y $d04_t union_{it}$. *No se pueden* calcular las diferencias apropiadas como, por ejemplo, $d02_t \Delta union_{it}$ y $d04_t \Delta union_{it}$, o incluso remplazar $d02_t$ y $d04_t$ con sus primeras diferencias.

Como un comentario general, es importante volver al modelo original y recordar que la diferenciación se usa para eliminar a_i. Es más fácil usar un comando integrado que permita la primera diferenciación como una opción en el análisis de datos de panel. (Se verán algunas de las otras opciones en el capítulo 14.)

Posibles dificultades con la primera diferenciación en los datos de panel

En esta sección y en las anteriores, se ha mencionado que la diferenciación de datos de panel en el tiempo, con el propósito de eliminar un efecto inobservable constante en el tiempo, es un método valioso para obtener efectos causales. No obstante, la diferenciación no está libre de dificultades. Ya se han comentado los problemas potenciales con el método cuando las variables explicativas no varían en gran medida con el tiempo (y el método no sirve de nada para las variables explicativas que nunca varían). Por desgracia, aun cuando se tenga suficiente variación en el tiempo en las x_{itj}, la estimación de primera diferencia (PD) puede estar sujeta a sesgos serios. Ya se ha mencionado que la exogeneidad estricta de los regresores es un supuesto crítico. Por desgracia, como se ve en Wooldridge (2002, sección 11.1), el hecho de tener más periodos, por lo general, no reduce la inconsistencia en el estimador de PD cuando los regresores no son estrictamente exógenos (por ejemplo, si $y_{i,t-1}$ se incluye entre las x_{itj}).

Otra desventaja importante del estimador de PD es que puede ser peor que el estimador combinado por MCO si una o más variables explicativas están sujetas a errores de medición, en particular el modelo de errores clásicos en las variables estudiado en la sección 9.3. La diferenciación de un regresor mal medido reduce su variación en relación con su correlación con el error diferenciado causado por el error clásico de medición, lo que da como resultado un sesgo potencialmente importante. La solución a estos problemas puede ser muy difícil de encontrar. Vea la sección 15.8 y Wooldridge (2002, capítulo 11).

RESUMEN

Se han estudiado métodos para analizar la combinación independiente de cortes transversales y los conjuntos de datos de panel. Las combinaciones independientes de cortes transversales surgen cuando se obtienen diferentes muestras aleatorias en distintos periodos (por lo general años). MCO que utiliza datos combinados es el método principal de estimación, y los procedimientos de inferencia usuales están disponibles, incluidas las correcciones de la heterocedasticidad. (La correlación serial no es un problema debido a que las muestras son independientes en el tiempo.) Debido a la dimensión de las series de tiempo, con frecuencia se permiten diferentes interceptos en el tiempo. También se podrían relacionar las variables binarias temporales con ciertas variables clave para ver cómo han cambiado respecto al tiempo. Esto es particularmente importante en la literatura sobre la evaluación de políticas para experimentos naturales.

Los conjuntos de datos de panel se usan en la práctica cada vez con mayor frecuencia, en especial para el análisis de políticas. Éstos son conjuntos de datos donde las mismas unidades de corte transversal se siguen en el transcurso del tiempo. Los conjuntos de datos de panel son más útiles cuando se controlan aspectos inobservables constantes en el tiempo de personas, empresas, ciudades, etc., los cuales se piensa que podrían estar correlacionados con las variables explicativas en el modelo. Una manera de eliminar el efecto inobservable es diferenciar los datos de periodos adyacentes. Luego, se puede usar un análisis de MCO estándar sobre las diferencias. Si se utilizan dos periodos de datos se obtiene como resultado una regresión de corte transversal de los datos diferenciados. Los procedimientos de inferencia usuales son asintóticamente válidos bajo la homocedasticidad; la inferencia exacta existe bajo el supuesto de normalidad.

Para más de dos periodos, es posible utilizar la estimación combinada de MCO en los datos diferenciados; se pierde el primer periodo debido a la diferenciación. Además de la homocedasticidad, se debe dar por sentado que los errores *diferenciados* no están serialmente correlacionados con el fin de aplicar los estadísticos t y F usuales. (El apéndice del capítulo contiene un listado detallado de los supuestos.) Naturalmente, cualquier variable que sea constante en el tiempo queda fuera del análisis.

TÉRMINOS CLAVE

Combinación independiente de cortes transversales
Cuasiexperimento
Datos de panel
Datos longitudinales
Ecuación en primera diferencia
Efecto fijo
Efecto inobservable
Efecto promedio del tratamiento
Error compuesto
Error idiosincrático
Estimador de diferencia en diferencias
Estimador de primera diferencia
Exogeneidad estricta
Experimento natural
Heterogeneidad inobservable
Modelo de efectos fijos
Modelo de efectos inobservables
Panel balanceado
Sesgo de heterogeneidad
Variables binarias anuales

PROBLEMAS

13.1 En el ejemplo 13.1 se hizo el supuesto que los promedios de todos los factores distintos de *educ* han permanecido constantes en el tiempo y que el nivel promedio de educación es 12.2 para la muestra de 1972 y 13.3 para la muestra de 1984. Utilizando las estimaciones de la tabla 13.1, calcule el cambio estimado en la fertilidad promedio entre 1972 y 1984. (Asegúrese de considerar el cambio del intercepto y el cambio en el nivel medio de educación.)

13.2 Las siguientes ecuaciones se estimaron usando la base de datos KIELMC.RAW para los años 1978 y 1981:

$$\widehat{\log(price)} = \underset{(.26)}{11.49} - \underset{(.058)}{.547}\, nearinc + \underset{(.080)}{.394}\, y81 \cdot nearinc$$

$$n = 321,\ R^2 = .220$$

y

$$\widehat{\log(price)} = \underset{(.27)}{11.18} + \underset{(.044)}{.563}\, y81 - \underset{(.067)}{.403}\, y81 \cdot nearinc$$

$$n = 321,\ R^2 = .337.$$

Compare las estimaciones para el término de interacción *y81·nearinc* con aquéllas de la ecuación (13.9). ¿Por qué son tan diferentes las estimaciones?

13.3 ¿Por qué no se pueden usar primeras diferencias cuando se tienen cortes transversales independientes en dos años (contrariamente a los datos de panel)?

13.4 Si considera que β_1 es positivo en la ecuación (13.14) y que Δu_i y $\Delta unem_i$ están correlacionados negativamente, ¿cuál es el sesgo en el estimador de MCO de β_1 en la ecuación en primera diferencia? [*Sugerencia:* revise la ecuación (5.4.).]

13.5 Suponga que se desea estimar el efecto de varias variables sobre el ahorro anual y se tiene una base de datos de panel sobre personas, recolectada el 31 de enero de 1990 y el 31 de enero de 1992. Si se incluye una variable binaria para 1992 y se utiliza la primera diferenciación, ¿es posible incluir además la edad en el modelo original? Explique.

13.6 En 1985, ni Florida ni Georgia tenían leyes que prohibían portar recipientes abiertos con bebidas alcohólicas en los compartimentos de los vehículos de pasajeros. Para 1990, Florida había aprobado esta ley, pero Georgia no.

i) Suponga que puede reunir muestras aleatorias de población en edad de conducir en los dos estados, para 1985 y 1990. Sea *arrest* una variable binaria igual a la unidad si una persona fue arrestada por conducir en estado de ebriedad durante el año. Sin controlar ningún otro factor, elabore un modelo de probabilidad lineal que le permita demostrar si la ley de recipientes abiertos redujo la probabilidad de ser arrestado por conducir en estado de ebriedad. ¿Cuál coeficiente de su modelo mide el efecto de la ley?

ii) ¿Por qué querría usted controlar otros factores del modelo? ¿Cuáles serían estos factores?

iii) Ahora, suponga que sólo puede reunir datos para 1985 y 1990 a nivel de condado para los dos estados. La variable dependiente sería la fracción de los conductores con licencia arrestados por conducir en estado de ebriedad durante el año. ¿Cómo difiere esta estructura de datos de los datos sobre personas descritos en el inciso i)? ¿Qué método econométrico usaría usted?

13.7 i) Usando la base de datos INJURY.RAW para Kentucky, la ecuación estimada cuando *afchnge* se omite en la ecuación (13.12) es

$$\widehat{\log(durat)} = \underset{(0.022)}{1.129} + \underset{(.042)}{.253}\, highearn + \underset{(.052)}{.198}\, afchnge \cdot highearn$$

$$n = 5{,}626,\ R^2 = .021.$$

¿Es de sorprender que la estimación de la interacción sea muy cercana a aquella de la ecuación (13.12)? Explique por qué.

ii) Cuando *afchnge* se incluye pero *highearn* se omite, el resultado es

$$\widehat{\log(durat)} = \underset{(0.023)}{1.233} - \underset{(.040)}{.100}\, afchnge + \underset{(.050)}{.447}\, afchnge \cdot highearn$$
$$n = 5{,}626,\ R^2 = .016.$$

¿Por qué el coeficiente del término de interacción ahora es mucho mayor que en la ecuación (13.12)? [*Sugerencia:* en la ecuación (13.10), ¿cuál es el supuesto que se hace sobre los grupos de tratamiento y control si $\beta_1 = 0$?]

EJERCICIOS EN COMPUTADORA

C13.1 Utilice la base de datos FERTIL1.RAW para este ejercicio.

i) En la ecuación estimada del ejemplo 13.1, pruebe si las condiciones de vida a los 16 años influyen sobre la fertilidad. (El grupo base es una ciudad grande.) Informe el valor del estadístico *F* y el valor-*p*.

ii) Pruebe si la región del país a los 16 años (el Sur es el grupo base) influye sobre la fertilidad.

iii) Sea *u* el término de error en la ecuación de la población. Suponga que piensa que la varianza de *u* cambia con el tiempo (pero no con *educ*, *age*, etc.). Un modelo que captura esto es

$$u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + \ldots + \gamma_6 y84 + v.$$

Usando este modelo, pruebe si hay heterocedasticidad en *u*. (*Sugerencia:* su prueba *F* debe tener 6 y 1,122 grados de libertad.)

iv) Añada los términos de interacción *y74·educ*, *y76·educ*, …, *y84·educ* al modelo estimado de la tabla 13.1. Explique qué representan estos términos. ¿Son conjuntamente significativos?

C13.2 Utilice la base de datos CPS78_85.RAW para este ejercicio.

i) ¿Cómo interpreta el coeficiente de *y85* en la ecuación (13.2)? ¿Tiene una interpretación interesante? (Tenga cuidado aquí; debe considerar los términos de interacción *y85·educ* y *y85·female*.)

ii) Si se mantienen otros factores fijos, ¿cuál es el incremento porcentual estimado en el salario nominal para un hombre con 12 años de educación? Proponga una regresión para obtener un intervalo de confianza para esta estimación. [*Sugerencia:* para obtener el intervalo de confianza, remplace *y85·educ* con *y85·*(*educ* − 12); remítase al ejemplo 6.3.]

iii) Vuelva a estimar la ecuación (13.2) pero mida todos los salarios en dólares de 1978. En particular, defina el salario real como *rwage* = *wage* para 1978 y como *rwage* = *wage*/1.65 para 1985. Ahora, use log(*rwage*) en lugar de log(*wage*) al estimar la ecuación (13.2). ¿Cuáles coeficientes difieren de aquéllos de la ecuación (13.2)?

iv) Explique por qué la *R*-cuadrada de su regresión del inciso iii) no es igual a aquélla de la ecuación (13.2). (*Sugerencia:* los residuales y, por ende, la suma de los residuales cuadrados, de las dos regresiones *son* idénticos.)

v) Describa cómo la afiliación a un sindicato cambió de 1978 a 1985.

vi) A partir de la ecuación (13.2), pruebe si el diferencial de salarios por afiliación a un sindicato cambió con el tiempo. (Debe ser una prueba *t* simple.)

vii) ¿Sus hallazgos de los incisos v) y vi) entran en conflicto?

C13.3 Use los datos de KIELMC.RAW para este ejercicio.

i) La variable *dist* es la distancia de cada casa al incinerador, en pies. Considere el modelo

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u.$$

Si la construcción del incinerador reduce el valor de las casas más cercanas a su ubicación, ¿cuál es el signo de δ_1? ¿Qué significa si $\beta_1 > 0$?

ii) Estime el modelo del inciso i) e informe los resultados de la manera acostumbrada. Interprete el coeficiente de $y81 \cdot \log(dist)$. ¿Qué concluye a partir de esto?

iii) Añada *age*, age^2, *rooms*, *baths*, log(*intst*), log(*land*) y log(*area*) a la ecuación. Ahora, ¿cuál es su conclusión acerca del efecto del incinerador sobre los valores de las casas?

iv) ¿Por qué el coeficiente de log(*dist*) es positivo y estadísticamente significativo en el inciso ii) pero no en el inciso iii)? ¿Qué dice esto sobre los controles usados en el inciso iii)?

C13.4 Utilice la base de datos INJURY.RAW para este ejercicio.

i) Con los datos de Kentucky, vuelva a estimar la ecuación (13.12) añadiendo como variables explicativas *male*, *married* y un conjunto completo de variables binarias de la industria y el tipo de lesión. ¿Cómo cambia la estimación de $afchnge \cdot highearn$ cuando se controlan estos otros factores? ¿El estimador sigue siendo estadísticamente significativo?

ii) ¿Qué significa la pequeña *R*-cuadrada del inciso i)? ¿Quiere decir esto que la ecuación no es útil?

iii) Estime la ecuación (13.12) con los datos de Michigan. Compare las estimaciones del término de interacción para Michigan y Kentucky. ¿El estimador de Michigan es estadísticamente significativo? ¿Qué significa esto?

C13.5 Utilice los datos de RENTAL.RAW para este ejercicio. Los datos para los años 1980 y 1990 incluyen precios de arrendamiento (*rent*) y otras variables de las ciudades universitarias. La idea es ver si una mayor presencia de los estudiantes influye en las tasas de arrendamiento. El modelo de efectos inobservables es

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

donde *pop* es la población de la ciudad, *avginc* es el ingreso promedio y *pctstu* es la población estudiantil expresada como un porcentaje de la población de la ciudad (durante el año escolar).

i) Obtenga la estimación combinada por MCO y reporte los resultados en la forma acostumbrada. ¿Qué significa el coeficiente estimado de la variable binaria de 1990? ¿Qué obtiene para $\hat{\beta}_{pctstu}$?

ii) ¿Son válidos los errores estándar que reportó en el inciso i)? Explique las razones.

iii) Ahora diferencie la ecuación y estime por MCO. Compare el estimador de β_{pctstu} con el estimador del inciso ii). ¿El tamaño relativo de la población estudiantil parece influir en el precio de la renta?

iv) Obtenga los errores estándar robustos a la heterocedasticidad de la ecuación en primera diferencia del inciso iii). ¿Cambian sus conclusiones?

C13.6 Utilice la base de datos CRIME3.RAW para este ejercicio.

i) En el modelo del ejemplo 13.6, pruebe la hipótesis H_0: $\beta_1 = \beta_2$. (*Sugerencia:* defina $\theta_1 = \beta_1 - \beta_2$ y escriba β_1 en función de θ_1 y β_2. Sustituya esto en la ecuación y ordene los términos. Haga una prueba t para θ_1.)

ii) Si $\beta_1 = \beta_2$, demuestre que la ecuación en diferencia puede escribirse como

$$\Delta\log(crime_i) = \delta_0 + \delta_1\Delta avgclr_i + \Delta u_i,$$

donde $\delta_1 = 2\beta_1$ y $avgclr_i = (clrprc_{i,-1} + clrprc_{i,-2})/2$ es el promedio del porcentaje de casos resueltos en los dos años anteriores.

iii) Estime la ecuación del inciso ii). Compare la R-cuadrada ajustada con aquélla de la ecuación (13.22). ¿Qué modelo utilizaría finalmente?

C13.7 Utilice la base de datos GPA3.RAW para este ejercicio. La base de datos es para 366 estudiantes atletas de una universidad grande, para los semestres de otoño y primavera. [Un análisis similar se halla en Maloney y McCormick (1993), pero aquí se utiliza un verdadero conjunto de datos de panel]. Como se cuenta con datos de dos semestres para cada estudiante, resulta adecuado un modelo de efectos inobservables. La primera interrogante de interés es la siguiente: ¿los estudiantes atletas tienen un menor aprovechamiento en la escuela durante el semestre en que su deporte está en temporada?

i) Utilice MCO combinados para estimar un modelo con el promedio de calificaciones del semestre (*trmgpa*) como variable dependiente y como variables explicativas *spring*, *sat*, *hsperc*, *female*, *black*, *white*, *frstsem*, *tothrs*, *crsgpa* y *season*. Interprete el coeficiente de *season* (variable binaria igual a uno si el deporte está en temporada). ¿Es estadísticamente significativo?

ii) La mayoría de los atletas que realiza su deporte sólo en el otoño son jugadores de fútbol americano. Suponga que sus niveles de capacidad innata difieren de forma sistemática de los de otros atletas. Si la capacidad no se captura adecuadamente por medio de las puntuaciones de la prueba de admisión a la universidad (*sat*) y el percentil que ocupó en el bachillerato (*hsperc*), explique por qué los estimadores obtenidos estarán sesgados.

iii) Utilice los datos diferenciados entre los dos semestres. ¿Qué variables se eliminan? Ahora pruebe el efecto de que el deporte esté en temporada.

iv) ¿Considera que hay una o más variables potencialmente importantes que varían en el tiempo y que se hayan omitido del análisis?

C13.8 La base de datos VOTE2.RAW incluye datos de panel sobre las elecciones para la Cámara de Representantes de los Estados Unidos en 1988 y 1990. Sólo los ganadores de 1988 que también compitieron en 1990 aparecen en la muestra; se trata de titulares del cargo. Un modelo de efectos inobservables que explica el porcentaje del voto obtenido por los titulares del cargo (*vote*) en términos de gastos de los dos candidatos es

$$vote_{it} = \beta_0 + \delta_0 d90_t + \beta_1\log(inexp_{it}) + \beta_2\log(chexp_{it}) + \beta_3 incshr_{it} + a_i + u_{it},$$

donde $incshr_{it}$ es la parte del total de gastos de campaña correspondiente a los titulares del cargo (en forma de porcentaje), *inexp* y *chexp* son los gastos en dólares del candidato titular del cargo y del candidato opositor, respectivamente. El efecto inobservable a_i contiene características del titular del cargo, como la "calidad", además de aspectos sobre el distrito que son constantes. El género y el partido del titular son constantes en el tiempo, de manera que están contenidos en a_i. Lo que interesa es el efecto de los gastos de campaña en los resultados de las elecciones.

i) Diferencie la ecuación dada a lo largo de los dos años y estime por MCO la ecuación en diferencias. ¿Qué variables son significativas individualmente al nivel de 5%, contra una alternativa de dos colas?

ii) En la ecuación del inciso i), haga una prueba de significancia conjunta para $\Delta\log(inexp)$ y $\Delta\log(chexp)$. Reporte el valor-p.

iii) Vuelva a estimar la ecuación del inciso i) con $\Delta incshr$ como la única variable independiente. Interprete el coeficiente de $\Delta incshr$. Por ejemplo, si la parte de gastos del titular del cargo aumenta 10 puntos porcentuales, según su predicción ¿cómo se afecta el porcentaje de votos obtenido por el titular?

iv) Vuelva a realizar el inciso iii), pero ahora sólo use los datos en los que en ambas elecciones se repite el mismo candidato opositor ($rptchall = 1$). [Esto nos permite controlar características de los candidatos opositores también, que estarían en a_i. Levitt (1994) realiza un análisis mucho más exhaustivo.]

C13.9 Utilice la base de datos CRIME4.RAW para este ejercicio.

i) A la ecuación del ejemplo 13.9, añada los logaritmos de cada variable de salario en la base de datos y estime el modelo en primera diferencia. ¿Cómo afecta la inclusión de estas variables a los coeficientes de las variables de justicia penal del ejemplo 13.9?

ii) ¿Tienen todas las variables del inciso i) el signo esperado? ¿Son conjuntamente significativas? Explique por qué.

C13.10 Para este ejercicio se utiliza la base de datos JTRAIN.RAW con el propósito de determinar el efecto de las subvenciones para la capacitación laboral (*grant*) sobre las horas de capacitación para el trabajo por empleado (*hrsemp*). El modelo básico para los tres años es

$$\begin{aligned} hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} \\ + \beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}. \end{aligned}$$

i) Estime la ecuación utilizando las primeras diferencias. ¿Cuántas empresas se emplean en la estimación? ¿Cuántas observaciones totales se podrían usar si cada empresa tuviera datos para todas las variables (en particular, para *hrsemp*) para los tres periodos?

ii) Interprete el coeficiente de *grant* (variable binaria igual a uno si recibió subvención) y comente su significancia.

iii) ¿Le extraña que $grant_{-1}$ sea insignificante? Explique sus razones.

iv) ¿Las empresas más grandes (*employ* es número de empleados) capacitan más o menos en promedio a sus empleados? ¿Son muy grandes las diferencias en la capacitación?

C13.11 La base de datos MATHPNL.RAW contiene datos de panel sobre los distritos escolares en Michigan para los años 1992 a 1998. Es el análogo a nivel distrital de los datos a nivel escolar usados por Papke (2005). La variable de respuesta que interesa en esta pregunta es *math4*, el porcentaje de estudiantes de cuarto grado en un distrito que recibe una calificación aprobatoria en un examen estandarizado de matemáticas. La variable explicativa clave es *rexpp*, que son los gastos reales por alumno en el distrito. Los montos están expresados en dólares de 1997. La variable del gasto aparecerá de forma logarítmica.

i) Considere el modelo estático de efectos inobservables

$$\begin{aligned} math4_{it} = \delta_1 y93_t + \ldots + \delta_6 y98_t + \beta_1 \log(rexpp_{it}) \\ + \beta_2 \log(enrol_{it}) + \beta_3 lunch_{it} + a_i + u_{it}, \end{aligned}$$

donde $enrol_{it}$ es la matrícula total del distrito y $lunch_{it}$ es el porcentaje de estudiantes en el distrito elegibles para el programa de almuerzos escolares. (De modo que $lunch_{it}$ es una medida bastante aceptable de la tasa de pobreza en todo el distrito). Sostenga que $\beta_1/10$ es el cambio en puntos porcentuales en $math4_{it}$ cuando el gasto real por estudiante aumenta aproximadamente 10%.

ii) Use la primera diferencia para estimar el modelo del inciso i). El método más sencillo es permitir un intercepto en la ecuación en primeras diferencias e incluir variables binarias para los años de 1994 a 1998. Interprete el coeficiente de la variable del gasto.

iii) Ahora, añada un rezago de la variable del gasto y vuelva a estimar usando las primeras diferencias. Observe que perdió otro año de datos, así que sólo está usando los cambios a partir de 1994. Comente los coeficientes y la significancia de las variables del gasto actual y rezagado.

iv) Obtenga los errores estándar robustos a la heterocedasticidad para la regresión en primeras diferencias del inciso iii). ¿Cómo se comparan estos errores estándar con aquéllos del inciso iii) para las variables del gasto?

v) Ahora, obtenga los errores estándar robustos a la heterocedasticidad y a la correlación serial. ¿Cómo afecta esto a la significancia de la variable del gasto rezagado?

vi) Verifique que los errores diferenciados $r_{it} = \Delta u_{it}$ tengan correlación serial negativa al realizar una prueba de correlación serial AR(1).

vii) Con base en la prueba de hipótesis conjunta completamente robusta, ¿parece necesario incluir las variables de la matrícula y la comida en el modelo?

C13.12 Use la base de datos MURDER.RAW para este ejercicio.

i) Utilizando los años 1990 y 1993 haga una estimación combinada de MCO de la ecuación

$$mrdrte_{it} = \delta_0 + \delta_1 d93_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it},\ t = 1, 2$$

y reporte los resultados como acostumbra. No se preocupe de que los errores estándar sean inapropiados debido a la presencia de a_i. ¿Estima un efecto disuasivo de la pena capital (*exec*) sobre la tasa de homicidios (*mrdrte*)?

ii) Calcule las estimaciones de PD (use sólo las diferencias de 1990 a 1993; debe tener 51 observaciones en la regresión por PD). ¿Qué concluye ahora sobre un efecto disuasivo?

iii) En la regresión por PD del inciso ii), obtenga los residuales, $\hat{e}_i$. Haga la regresión de Breusch-Pagan de $\hat{e}_i^2$ sobre $\Delta exec_i$, $\Delta unem_i$ y calcule la prueba F para la heterocedasticidad. Haga lo mismo para el caso especial de la prueba de White [es decir, la regresión de $\hat{e}_i^2$ *sobre* $\hat{y}_i$, $\hat{y}_i^2$, donde los valores ajustados son del inciso ii)]. ¿Qué concluye acerca de la heterocedasticidad en la ecuación de PD?

iv) Efectúe la misma regresión del inciso ii), pero obtenga estadísticos t robustos a la heterocedasticidad. ¿Qué ocurre?

v) ¿Cuál estadístico t sobre $\Delta exec_i$ considera más adecuado: el usual o el robusto a la heterocedasticidad? ¿Por qué?

C13.13 Para este ejercicio utilice la base de datos WAGEPAN.RAW.

i) Considere el modelo de efectos inobservables

$$\begin{aligned} lwage_{it} = {} & \beta_0 + \delta_1 d81_t + \ldots + \delta_7 d87_t + \beta_1 educ_i \\ & + \gamma_1 d81_t educ_i + \ldots + \delta_7 d87_t educ_i + \beta_2 union_{it} + a_i + u_{it}, \end{aligned}$$

donde se permite que a_i se correlacione con $educ_i$ y $union_{it}$ (variable binaria igual a uno si pertenece a un sindicato). ¿Cuáles parámetros puede estimar usando la primera diferencia?

ii) Estime la ecuación del inciso i) por PD, y pruebe la hipótesis nula de que la rentabilidad de la educación no ha cambiado con el tiempo.

iii) Pruebe la hipótesis del inciso ii) usando una prueba completamente robusta, es decir, una que permita heterocedasticidad arbitraria y correlación serial en los errores en PD, Δu_{it}. ¿Cambia su conclusión?

iv) Ahora permita que el diferencial salarial debido a la afiliación sindical cambie con el tiempo (junto con la educación) y estime la ecuación por PD. ¿Cuál es el diferencial salarial por la afiliación sindical en 1980? ¿Y el de 1987? ¿La diferencia es estadísticamente significativa?

v) Pruebe la hipótesis nula de que el diferencial salarial por la afiliación sindical no ha cambiado con el tiempo y comente sus resultados según su respuesta al inciso iv).

C13.14 Utilice los datos de JTRAIN3.RAW para esta pregunta.

i) Estime el modelo de regresión simple $re78 = \beta_0 + \beta_1 train + u$, e informe los resultados como acostumbra. Con base en esta regresión, ¿parece que la capacitación laboral (*train*), la cual ocurrió en 1976 y 1977, tuvo un efecto positivo en los ingresos laborales reales en 1978 (*re78*)?

ii) Ahora utilice el cambio en los ingresos laborales reales, $cre = re78 - re75$, como variable dependiente. (No es necesario diferenciar *train* debido a que se da por sentado que no hubo capacitación laboral antes de 1975. Esto significa que si se define $ctrain = train78 - train75$ entonces $ctrain = train78$ ya que $train75 = 0$.) ¿Cuál es el efecto estimado de la capacitación ahora? Comente cómo se compara con la estimación del inciso i).

iii) Calcule el intervalo de confianza de 95% para el efecto de la capacitación usando el error estándar usual de MCO y el error estándar robusto a la heterocedasticidad, y describa sus hallazgos.

Apéndice 13A

Supuestos para las estimaciones combinadas de MCO usando las primeras diferencias

En este apéndice se proporcionan enunciados cuidadosos de los supuestos necesarios para obtener el estimador de primeras diferencias. La verificación de estas afirmaciones es un poco complicada, pero se halla en Wooldridge (2002, capítulo 10).

Supuesto PD.1

Para cada i, el modelo es

$$y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, \dots, T,$$

donde las β_j son los parámetros que hay que estimar y a_i es el efecto inobservable.

Supuesto PD.2

Se tiene una muestra aleatoria de corte transversal.

Supuesto PD.3

Cada variable explicativa cambia con el tiempo (para al menos alguna i) y no existe una relación lineal perfecta entre las variables explicativas.

En el siguiente supuesto resulta de utilidad hacer que $\mathbf{X}_i$ denote las variables explicativas de todos los periodos, para la observación de corte transversal i; así, $\mathbf{X}_i$ contiene a las x_{itj}, $t = 1, \ldots, T$, $j = 1, \ldots, k$.

Supuesto PD.4

Para cada t, el valor esperado del error idiosincrático, dadas las variables explicativas en *todos* los periodos y el efecto inobservable, es cero: $\mathrm{E}(u_{it}|\mathbf{X}_i, a_i) = 0$.

Cuando el supuesto PD.4 es válido, en ocasiones se dice que las x_{itj} son *estrictamente exógenas condicionadas al efecto inobservable*. La idea es que, una vez que se controlan las a_i, no hay correlación entre las x_{isj} y el error idiosincrático restante, u_{it}, para toda s y t.

Como se estableció, el supuesto PD.4 es más fuerte de lo necesario. Se utiliza esta forma de supuesto porque hace énfasis en que nos interesa la ecuación

$$\mathrm{E}(y_{it}|\mathbf{X}_i, a_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it}, a_i) = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i,$$

de modo que las β_j miden los efectos parciales de las variables explicativas observables manteniendo fijo, o "controlando por", el efecto inobservable, a_i. Sin embargo, una implicación importante de PD.4, que es suficiente para el insesgamiento del estimador de PD, es que $\mathrm{E}(\Delta u_{it}|\mathbf{X}_i) = 0$, para $t = 2, \ldots, T$ y $j = 1, \ldots, k$. De hecho, por consistencia se puede sencillamente suponer que Δx_{itj} no se correlaciona con Δu_{it} para toda $t = 2, \ldots, T$ y $j = 1, \ldots, k$. Véase Wooldridge (2002, capítulo 10) para un análisis más detallado.

De acuerdo con estos primeros cuatro supuestos, los estimadores de la ecuación en primeras diferencias son insesgados. El supuesto clave es PD.4, es decir, la exogeneidad estricta de las variables explicativas. Bajo los mismos supuestos, también se demuestra que el estimador de primeras diferencias es consistente cuando T es fija y conforme $N \to \infty$ (y tal vez de manera más general).

Los siguientes dos supuestos aseguran que los errores estándar y los estadísticos de prueba que resultan de la estimación combinada de MCO en primeras diferencias sean válidos (asintóticamente).

Supuesto PD.5

La varianza de los errores diferenciados, condicional en todas las variables explicativas, es constante: $\mathrm{Var}(\Delta u_{it}|\mathbf{X}_i) = \sigma^2$, $t = 2, \ldots, T$.

Supuesto PD.6

Para toda $t \neq s$, las *diferencias* de los errores idiosincráticos no se correlacionan (condicional en todas las variables explicativas): $\mathrm{Cov}(\Delta u_{it}, \Delta u_{is}|\mathbf{X}_i) = 0$, $t \neq s$.

El supuesto PD.5 asegura que los errores diferenciados, Δu_{it}, son homocedásticos y el supuesto PD.6 establece que los errores diferenciados no se correlacionan serialmente, lo que significa que las u_{it} siguen una caminata aleatoria en el tiempo (vea el capítulo 11). Bajo los supuestos PD.1 a PD.6, el estimador de PD de las β_j es el mejor estimador lineal insesgado (condicional en todas las variables explicativas).

Supuesto PD.7

Condicionales en $\mathbf{X}_i$, las Δu_{it} son variables aleatorias independientes e idénticamente distribuidas normales.

Cuando se agrega el supuesto PD.7, los estimadores de PD se distribuyen en forma normal y los estadísticos t y F de la estimación combinada de MCO sobre las diferencias tienen distribuciones exactas t y F. Sin el supuesto PD.7, se emplean las aproximaciones asintóticas usuales.

CAPÍTULO 14

Métodos avanzados para datos de panel

En este capítulo se cubren dos métodos que sirven para estimar modelos de datos de panel de efectos inobservables que son al menos tan comunes como aplicar la primera diferencia. Aunque estos métodos son hasta cierto punto difíciles de describir e implementar existen varios paquetes econométricos que los soportan.

En la sección 14.1 se estudia el estimador de efectos fijos, el cual, al igual que se hace con la primera diferencia, utiliza una transformación para eliminar el efecto inobservable a_i antes de la estimación. Cualquier variable explicativa que sea constante en el tiempo se elimina junto con a_i.

El estimador de efectos aleatorios de la sección 14.2 es atractivo cuando se piensa que el efecto inobservable no está correlacionado con ninguna de las variables explicativas. Si se tienen buenos controles en la ecuación, se podría considerar que cualquier heterogeneidad sobrante olvidada sólo induce correlación serial en el término del error compuesto, pero no genera correlación entre los errores compuestos y las variables explicativas. La estimación de los modelos de efectos aleatorios mediante mínimos cuadrados generalizados es muy sencilla y se lleva a cabo de manera rutinaria en muchos paquetes de econometría.

En la sección 14.3 se muestra cómo se aplican los métodos de datos de panel a otras estructuras de datos, entre las que se incluyen las muestras de datos apareados y las muestras de agrupamientos.

14.1 Estimación de efectos fijos

Aplicar la primera diferencia es sólo una de las muchas formas de eliminar el efecto fijo, a_i. Un método alterno, que funciona mejor bajo ciertos supuestos, es la **transformación de efectos fijos**. Para ver lo que este método realiza, considere un modelo con una sola variable explicativa: para cada i,

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad t = 1, 2, \ldots, T. \tag{14.1}$$

Ahora, para cada i, se promedia la ecuación en el tiempo y se obtiene

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \tag{14.2}$$

donde $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, y así sucesivamente. Como a_i permanece constante en el tiempo, aparece tanto en la ecuación (14.1) como en la (14.2). Si se resta la ecuación (14.2) de la (14.1) para cada t, se obtiene

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i, \quad t = 1, 2, \ldots, T,$$

o

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it} + \ddot{u}_{it}, \quad t = 1, 2, \ldots, T, \tag{14.3}$$

donde $\ddot{y}_{it} = y_{it} - \bar{y}_i$ son los **datos con el tiempo deducido** sobre y, y del mismo modo para $\ddot{x}_{it}$ y $\ddot{u}_{it}$. La transformación de efectos fijos también se conoce como **transformación intragrupal (*within*)**. Lo importante respecto a la ecuación (14.3) es que el efecto inobservable, a_i, ha desaparecido. Esto sugiere que se debe hacer una estimación combinada de MCO de la ecuación (14.3). Un estimador de MCO combinados que se basa en las variables con el tiempo deducido se llama **estimador de efectos fijos** o **estimador intragrupal (*within*)**. Este último nombre proviene del hecho de que en la ecuación (14.3) MCO utiliza la variación en el tiempo en y y x *dentro* (*within*) de cada observación de corte transversal.

El *estimador intragrupal* (*between*) se obtiene a partir del estimador MCO en la ecuación de corte transversal (14.2) (donde se incluye un intercepto, β_0): se utilizan los promedios a lo largo del tiempo tanto de y como de x para luego efectuar una regresión de corte transversal. No se estudiará el estimador intragrupal con detalle porque está sesgado cuando a_i se correlaciona serialmente con $\bar{x}_i$ (vea el problema 14.2). Si se cree que a_i no está correlacionado con x_{it}, es mejor utilizar el estimador de efectos aleatorios, el cual se estudia en la sección 14.2. El estimador intragrupal ignora información importante sobre cómo cambian las variables con el tiempo.

La adición a la ecuación de más variables explicativas suscita pocos cambios. El **modelo de efectos inobservables** original es

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, 2, \ldots, T. \tag{14.4}$$

Sencillamente se aplica la deducción del tiempo a cada variable explicativa, incluyendo a factores como las variables binarias temporales, y luego se realiza una regresión combinada de MCO utilizando todas las variables con el tiempo deducido. La ecuación general con el tiempo deducido para cada i es

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it1} + \beta_2\ddot{x}_{it2} + \ldots + \beta_k\ddot{x}_{itk} + \ddot{u}_{it}, \quad t = 1, 2, \ldots, T, \tag{14.5}$$

la cual se estima por MCO combinados.

Bajo el supuesto de exogeneidad estricta sobre las variables explicativas, el estimador de efectos fijos es insesgado: aproximadamente, el error idiosincrático u_{it} no debe correlacionarse serialmente con ninguna variable explicativa en *todos* los periodos. (Vea el apéndice del capítulo para los enunciados precisos de los supuestos.) El estimador de efectos fijos permite la correlación arbitraria entre a_i y las variables explicativas en cualquier periodo, al igual que con las primeras diferencias. Debido a esto, cualquier variable explicativa que sea constante en el tiempo para toda i queda erradicada por la transformación de efectos fijos: $\ddot{x}_{it} = 0$ para toda i y t, si x_{it} es constante en t. Por consiguiente, no es posible incluir variables como el género o la distancia de una ciudad a un río.

Pregunta 14.1

Suponga que en una ecuación de ahorros familiares, para los años 1990, 1991 y 1992, $kids_{it}$ denota el número de hijos en la familia i para el año t. Si el número de hijos es constante durante este periodo de tres años para la mayoría de las familias de la muestra, ¿qué problemas podría causar esto para estimar el efecto que tiene el número de hijos en los ahorros?

El otro supuesto necesario para que un análisis por MCO directo sea válido es que los errores u_{it} sean homocedásticos y no estén serialmente correlacionados (en t); consulte el apéndice de este capítulo.

Existe un punto sutil en la determinación de los grados de libertad para el estimador de efectos fijos. Cuando se hace una estimación combinada de MCO de la ecuación con el tiempo deducido (14.5), se tienen NT observaciones totales y k variables independientes. (Observe que no existe intercepto en la ecuación (14.5); se elimina por la transformación de efectos fijos.) Por tanto, en apariencia se deben tener $NT - k$ grados de libertad. Este cálculo es incorrecto. Para cada observación de corte transversal i se pierde un gl debido al tiempo deducido. En otras palabras, para cada i, los errores deducidos $\ddot{u}_{it}$ suman en total cero cuando se suman a lo largo de t, de manera que se pierde un grado de libertad. (No existe esta restricción para los errores idiosincráticos originales u_{it}.) Por tanto, los grados de libertad apropiados son $gl = NT - N - k = N(T - 1) - k$. Por suerte, los paquetes de regresión modernos que manejan la estimación de efectos fijos también permiten un cálculo adecuado de gl. Pero si se debe determinar el tiempo deducido y la estimación combinada de MCO por cuenta propia, es necesario corregir los errores estándar y los estadísticos de prueba.

Ejemplo 14.1

[Efecto de la capacitación laboral en las tasas de desperdicios industriales]

Se utilizan los datos para tres años, 1987, 1988 y 1989, sobre 54 empresas que reportaron tasas de desperdicio en cada año. Ninguna empresa recibió subsidios antes de 1988; en 1988, 19 empresas recibieron subsidios; en 1989, 10 empresas diferentes recibieron subsidios. Por ende, se debe considerar también la posibilidad de que la capacitación laboral adicional en 1988 hizo que los empleados fueran más productivos en 1989. Esto se hace fácilmente al incluir un valor rezagado del indicador de subsidios. También se incluyen variables binarias anuales para 1988 y 1989. Los resultados se proporcionan en la tabla 14.1.

Se presentan los resultados en una forma que destaca la necesidad de interpretar las estimaciones a la luz del modelo de efectos inobservables (14.4). Se controlan, de forma explícita, los efectos inobservables constantes en el tiempo en a_i. La deducción del tiempo permite estimar las β_j, pero la ecuación (14.5) no es la mejor para interpretar las estimaciones.

Resulta interesante que el efecto estimado del rezago del subsidio a la capacitación es sustancialmente más grande que el efecto contemporáneo: la capacitación laboral ejerce un efecto al menos un año después. Dado que la variable dependiente está en forma logarítmica, se pronostica que obtener un subsidio en 1988 disminuye la tasa de desperdicios industriales en 1989 aproximadamente 34.4% [$\exp(-.422) - 1 \approx -.344$]; el coeficiente de $grant_{-1}$ es significativo al 5% contra la alternativa de dos colas. El coeficiente de $grant$ es significativo al 10%, y su magnitud no es pequeña. Observe que los gl se obtienen como $N(T - 1) - k = 54(3 - 1) - 4 = 104$.

El coeficiente de $d89$ indica que la tasa de desperdicio fue sustancialmente menor en 1989 que en el año base, 1987, aun en ausencia de los subsidios para la capacitación laboral. Así pues, resulta importante considerar estos efectos agregados. Si se omiten las binarias anuales, el aumento secular en la productividad de los trabajadores se atribuiría a los subsidios para la capacitación laboral. En la tabla 14.1 se aprecia que, aun después de tomar en cuenta las tendencias agregadas en la productividad, los subsidios para la capacitación tuvieron un efecto estimado grande.

Pregunta 14.2

De acuerdo con el programa de Michigan, si una empresa recibía un subsidio en un año, no era elegible para el año siguiente. ¿Qué implica esto sobre la correlación entre $grant$ y $grant_{-1}$?

TABLA 14.1

Estimación de efectos fijos de la ecuación de la tasa de desperdicio

Variable dependiente: log(*scrap*)	
Variables independientes	**Coeficiente (error estándar)**
d88	−.080 (.109)
d89	−.247 (.133)
grant	−.252 (.151)
$grant_{-1}$	−.422 (.210)
Observaciones Grados de libertad *R*-cuadrada	162 104 .201

Por último, resulta crucial considerar el efecto rezagado en el modelo. Si se omite $grant_{-1}$, entonces se está suponiendo que el efecto de la capacitación laboral no perdura hasta el año siguiente. La estimación de *grant* cuando se elimina $grant_{-1}$ es $-.082$ ($t = -.65$); lo cual resulta más pequeño y estadísticamente insignificante.

Cuando se estima un modelo de efectos inobservables mediante efectos fijos, no queda claro cómo se debe calcular una medida de bondad de ajuste. La *R*-cuadrada que se presenta en la tabla 14.1 se basa en la transformación intragrupal: es la *R*-cuadrada obtenida a partir de la estimación (14.5). Así, se interpreta como la cantidad en la variación temporal en la y_{it} que se explica por la variación temporal en las variables explicativas. Existen otras formas posibles de calcular la *R*-cuadrada y después se analizará una de ellas.

Aunque las variables constantes en el tiempo no pueden incluirse por sí solas en un modelo de efectos fijos, *pueden* interactuar con variables que cambian con el tiempo y, en particular, con variables binarias anuales. Por ejemplo, en una ecuación de salario donde la educación es constante en el tiempo para cada individuo de la muestra, se puede hacer que la educación interactúe con cada variable binaria anual para ver cómo el rendimiento de la educación ha cambiado en el tiempo. Pero no es posible utilizar efectos fijos para estimar el rendimiento de la educación en el periodo base, lo que significa que no es posible estimar este efecto en cualquier periodo, sólo se puede observar cómo difiere en cada año del correspondiente al periodo base.

Cuando se incluye un conjunto completo de variables binarias anuales —es decir, para todos los años salvo el primero— no es posible estimar el efecto de ninguna variable cuyo *cambio* en el tiempo sea constante. Un ejemplo de ello son los años de experiencia en un conjunto de datos de panel en el que cada persona trabaja cada año de modo que la experiencia siempre aumenta en uno anualmente, para cada individuo de la muestra. La presencia de a_i explica las diferencias entre las personas en cuanto a sus años de experiencia durante el periodo inicial. Pero, entonces, el efecto del aumento de un año en la experiencia no se distingue de los efectos de tiempo agregados (pues la experiencia se incrementa en la misma cantidad para todos). Esto también sería verdadero si, en lugar de variables binarias anuales separadas, se utilizara una tendencia lineal en el tiempo: para cada persona, la experiencia no se distingue de una tendencia lineal.

Ejemplo 14.2

[¿El rendimiento ha cambiado de la educación con el tiempo?]

Los datos de WAGEPAN.RAW son de Vella y Verbeek (1998). Cada uno de los 545 hombres de la muestra trabajó cada año de 1980 a 1987. Algunas variables del conjunto de datos varían con el tiempo: la experiencia, el estado civil y la afiliación sindical son las tres más importantes. Otras variables no cambian: raza y educación son los ejemplos clave. Si se utilizan efectos fijos (o las primeras diferencias), no es posible incluir raza, educación o experiencia en la ecuación. Sin embargo, es posible incluir interacciones de *educ* con variables binarias anuales de 1981 a 1987 para probar si el rendimiento de la educación fue constante en este periodo. Se utiliza log(*wage*) como variable dependiente, variables binarias anuales para el estado civil y la afiliación sindical, un conjunto completo de binarias anuales y los términos de interacción *d81·educ*, *d82·educ*, …, *d87·educ*.

Las estimaciones sobre los términos de interacción son positivas y, por lo general, mayores en los años más recientes. El coeficiente más grande, de .030 es el de *d87·educ*, con $t = 2.48$. En otras palabras, se estima que el rendimiento de la educación sea aproximadamente tres puntos porcentuales más grande en 1987 que en el año base 1980. (No se tiene una estimación del rendimiento de la educación en el año base, por las razones ya expuestas.) El otro término de interacción significativo es el de *d86·educ* (coeficiente = .027, $t = 2.23$). Las estimaciones sobre los primeros años son más reducidas e insignificantes al 5% contra la alternativa de dos colas. Si se realiza una prueba F de significancia conjunta para los siete términos de interacción, se obtiene un valor-p = .28: éste es un ejemplo de un conjunto de variables que son insignificantes de manera conjunta aun cuando algunas de ellas sean significativas individualmente. [Los *gl* para la prueba F son 7 y 3,799; el segundo de éstos proviene de $N(T - 1) - k = 545(8 - 1) - 16 = 3{,}799$.] En general, los resultados son consistentes con un incremento en el rendimiento de la educación durante este periodo.

Regresión de variables binarias

Un punto de vista tradicional respecto del modelo de efectos fijos es suponer que el efecto inobservable, a_i, es un parámetro que debe estimarse para cada i. Por tanto, en la ecuación (14.4), a_i es el intercepto para la persona i (o la empresa i, la ciudad i, etc.) que debe estimarse junto con las β_j. (Es claro que no es posible hacer lo anterior con un único corte transversal: habría que estimar $N + k$ parámetros con sólo N observaciones; se requieren por lo menos dos periodos.) La forma en que se estima un intercepto para cada i es asignar una variable binaria para cada observación de corte transversal, junto con las variables explicativas (y probablemente variables binarias para cada periodo). A este método se le llama comúnmente **regresión de variables binarias**. Aun cuando N no es muy grande (por ejemplo, $N = 54$ como en el ejemplo 14.1), esto da por resultado muchas variables explicativas —en la mayoría de los casos, demasiadas para hacer la regresión de manera explícita—. Así pues, el método de las variables binarias no es muy práctico para conjuntos de datos de panel con numerosas observaciones de corte transversal.

No obstante, la regresión de variables binarias posee algunas características interesantes. Lo más importante es que provee *exactamente* las mismas estimaciones de las β_j que se obtendrían de la regresión con datos con el tiempo deducido, y los errores estándar y otros estadísticos importantes también son idénticos. En consecuencia, el estimador de efectos fijos se obtiene por medio de la regresión de variables binarias. Un beneficio de esta regresión es que calcula adecuadamente los grados de libertad de forma directa (ventaja menor ahora que muchos paquetes econométricos han programado opciones para los efectos fijos.

La R-cuadrada de la regresión de variables binarias, por lo general, tiene un valor muy alto. Esto se debe a que se está incluyendo una variable binaria para cada unidad de corte transversal, lo que explica buena parte de la variación en los datos. Por ejemplo, si se estima el método de efectos inobservables del ejemplo 13.8 por medio de efectos fijos utilizando la regresión de variables binarias (lo que es posible con $N = 22$), entonces $R^2 = .933$. No debería provocar exaltación esta R-cuadrada grande: no sorprende que sea posible explicar buena parte de la variación en los reclamos de los seguros de desempleo con la ayuda de binarias para los años y para las ciudades. Como en el ejemplo 13.8, la estimación del coeficiente de la variable binaria EZ (zonas empresariales) es más importante que la R^2.

La R-cuadrada de la regresión de variables binarias se utiliza para calcular las pruebas F a la usanza tradicional, suponiendo desde luego que los supuestos del modelo lineal clásico son válidos (vea el apéndice del capítulo). En particular, se puede probar la significancia conjunta de todas las binarias de corte transversal ($N - 1$, ya que se elige una unidad como el grupo base). La R-cuadrada no restringida se obtiene de la regresión con todas las binarias de corte transversal, y la R-cuadrada restringida las omite. En la mayoría de las aplicaciones, las variables binarias son conjuntamente significativas.

En ocasiones, son de interés los interceptos estimados $\hat{a}_i$. Esto sucede si se desea estudiar la distribución de las $\hat{a}_i$ a lo largo de i, o si se quiere tomar una determinada empresa o ciudad para ver si su $\hat{a}_i$ está por encima o por debajo del valor promedio de la muestra. Se dispone de forma directa de estas estimaciones en la regresión de variables binarias, pero pocas veces las reportan los paquetes que tienen rutinas de efectos fijos (por la razón práctica de que hay demasiadas $\hat{a}_i$). Después de la estimación de los efectos fijos con N de cualquier tamaño, es muy sencillo calcular las $\hat{a}_i$:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1\bar{x}_{i1} - \ldots - \hat{\beta}_k\bar{x}_{ik},\ i = 1, \ldots, N, \tag{14.6}$$

donde la barra superior alude a los promedios a lo largo del tiempo y las $\hat{\beta}_j$ son las estimaciones de efectos fijos. Por ejemplo, si se ha estimado un modelo de delincuencia controlando diversos factores que varían con el tiempo, se puede obtener $\hat{a}_i$ para una ciudad y ver si los efectos fijos inobservados que contribuyen a la delincuencia están por encima o por debajo del promedio.

Algunos paquetes econométricos que soportan la estimación de efectos fijos reportan un "intercepto", lo cual puede crear confusión en vista de nuestra afirmación anterior de que el tiempo deducido elimina todas las variables constantes en el tiempo. [Vea la ecuación (14.5).] Los reportes de un intercepto general en la estimación de efectos fijos (EF) surgen de considerar las a_i como parámetros a estimar. Comúnmente, el intercepto reportado es el promedio en i de las $\hat{a}_i$. En otras palabras, el intercepto general es, en realidad, el promedio de los interceptos específicos individuales, el cual es un estimador insesgado y consistente de $\alpha = \mathrm{E}(a_i)$.

En la mayor parte de los estudios, las $\hat{\beta}_j$ son de interés y, por tanto, se utilizan ecuaciones con variables con el tiempo deducido para obtener estas estimaciones. Además, por lo común resulta mejor considerar las a_i como variables omitidas que se controlan mediante la transformación intragrupal. En general, el sentido en que se estima a_i es débil. De hecho, aun cuando $\hat{a}_i$ sea insesgada (de acuerdo con los supuestos EF.1 a EF.4 del apéndice del capítulo), no es consistente para un T fijo cuando $N \to \infty$. La razón es que, a medida que se agrega cada observación de corte transversal, se suma una nueva a_i. No se acumula información acerca de cada a_i cuando T es fijo. Con T grande, se obtienen mejores estimaciones de a_i, pero la mayoría de los conjuntos de datos de panel son del tipo en que N es grande y T es pequeño.

¿Efectos fijos o primera diferencia?

Hasta ahora se han visto dos métodos comparables para estimar modelos de efectos inobservables. Uno comprende la diferenciación de los datos y el otro la deducción del tiempo. ¿Cómo saber cuál debe utilizarse?

Es posible eliminar un caso de inmediato: cuando $T = 2$, las estimaciones de EF (efectos fijos) y PD (primera diferencia), así como todos los estadísticos de prueba, son *idénticos* y, por tanto, no importa a cuál se recurra. Desde luego, la equivalencia entre las estimaciones de EF y PD requiere que se estime el mismo modelo en cada caso. En particular, como se vio en el capítulo 13, es natural incluir un intercepto en la ecuación de PD; este intercepto en realidad es el intercepto para el segundo periodo en el modelo original escrito para los dos periodos. Por consiguiente, la estimación de EF debe incluir una variable binaria para el segundo periodo con el fin de que sea idéntica a la estimación de PD que incluye un intercepto.

Cuando $T = 2$, la PD cuenta con la ventaja de ser fácil de realizar en cualquier paquete econométrico o estadístico que soporte la manipulación de datos básica y resulta fácil calcular los estadísticos robustos a la heterocedasticidad después de la estimación de PD (debido a que cuando $T = 2$, la estimación de PD es sólo una regresión de corte transversal).

Cuando $T \geq 3$, los estimadores de EF y PD no son los mismos. En virtud de que ambos son insesgados, con base en los supuestos EF.1 a EF.4, no es posible utilizar el insesgamiento como un criterio. Además, ambos son consistentes (con T fijo cuando $N \rightarrow \infty$) bajo los supuestos EF.1 a EF.4. Para N grande y T pequeño, la elección entre estimadores de EF y PD depende de su eficiencia relativa, y esto está determinado por la correlación serial de los errores idiosincráticos, u_{it}. (Se supondrá homocedasticidad de las u_{it}, ya que las comparaciones de eficiencia exigen errores homocedásticos.)

Cuando las u_{it} no se correlacionan serialmente, los estimadores de efectos fijos son más eficientes que los de primera diferencia (y los errores estándar de efectos fijos son válidos). Dado que el modelo de efectos fijos casi siempre se establece con errores idiosincráticos no correlacionados serialmente, el estimador de EF se emplea más a menudo. Pero se debe recordar que este supuesto puede ser falso. En varias aplicaciones se puede esperar que los factores inobservables que se modifican en el tiempo estén correlacionados. Si u_{it} sigue una caminata aleatoria, significa que existe correlación serial positiva sustancial, entonces, la diferencia Δu_{it} no está serialmente correlacionada y es mejor la primera diferencia. En varios casos u_{it} muestra algo de correlación serial positiva, pero quizá no tanta como la de la caminata aleatoria. Por tanto, no es posible comparar con facilidad la eficiencia de los estimadores de EF y PD.

Resulta difícil probar que las u_{it} no se correlacionan serialmente después de la estimación de EF: se pueden estimar los errores con el tiempo deducido, $\ddot{u}_{it}$, pero no las u_{it}. Sin embargo, en la sección 13.3 se mostró cómo probar que los errores diferenciados, Δu_{it}, no están serialmente correlacionados. Si éste fuera el caso, se emplea la primera diferencia. Si existe correlación serial negativa sustancial en las Δu_{it}, probablemente sea mejor usar los EF. A menudo es recomendable probar ambos métodos; si los resultados no son sensibles, tanto mejor.

Cuando T es grande, y en especial cuando N no es muy grande (por ejemplo, $N = 20$ y $T = 30$), se debe extremar precauciones al utilizar el estimador de efectos fijos. Aunque los resultados de la distribución exacta son válidos para cualquier N y T bajo los supuestos clásicos de los efectos fijos, la inferencia puede ser muy sensible a las violaciones de los supuestos cuando N es pequeña y T es grande. En particular, cuando se utilizan los procesos de raíces unitarias (vea el capítulo 11) puede surgir el problema de la regresión espuria. La primera diferencia tiene la ventaja de convertir un proceso integrado de series de tiempo en un proceso débilmente dependiente. Por consiguiente, si se aplica la primera diferencia, es posible apelar al teorema del límite central, incluso en casos en que T sea mayor que N. La normalidad en los errores idiosincráticos no es necesaria, y la heterocedasticidad y la correlación serial pueden tratarse como se estudió en el capítulo 13. La inferencia con el estimador de efectos fijos es potencialmente más sensible a la no normalidad, la heterocedasticidad y la correlación serial en los errores idiosincráticos.

Al igual que el estimador de primera diferencia, el estimador de efectos fijos puede ser muy sensible al error de medición clásico en una o más variables explicativas. Sin embargo, si cada x_{itj} no está correlacionada con u_{it}, pero se viola el supuesto de exogeneidad estricta, por ejemplo, cuando una variable dependiente rezagada se incluye entre los regresores o existe retroalimentación entre u_{it} y los futuros resultados de la variable explicativa, entonces es probable que el estimador de EF sea considerablemente menos sesgado que el estimador de PD (a menos que $T = 2$). El hecho teórico importante es que en el estimador de PD el sesgo no depende de T, mientras que el sesgo en el estimador de EF tiende a cero a razón de $1/T$. Vea Wooldridge (2002, sección 11.1) para obtener detalles.

Por lo general, resulta difícil elegir entre los estimadores de EF y PD cuando generan resultados sustancialmente distintos. Es conveniente pues, reportar ambos conjuntos de resultados y tratar de determinar por qué difieren.

Efectos fijos con paneles no balanceados

Algunos conjuntos de datos de panel, en especial sobre personas o empresas, carecen de ciertos años, al menos en algunas unidades de corte transversal de la muestra. En este caso se llama al conjunto de datos un **panel no balanceado**. La mecánica de la estimación de efectos fijos con un panel no balanceado no es mucho más difícil que con uno balanceado. Si T_i es el número de periodos para la unidad de corte transversal i, sencillamente se utilizan las observaciones T_i al realizar la deducción del tiempo. El número total de observaciones es entonces $T_1 + T_2 + \ldots + T_N$. Como en el caso balanceado, se pierde un grado de libertad para cada observación de corte transversal debido a la deducción del tiempo. Cualquier paquete de regresión que procesa efectos fijos realiza el ajuste adecuado por esta pérdida. La regresión de variables binarias también se lleva a cabo exactamente de la misma manera que con un panel balanceado y los *gl* se obtienen de forma apropiada.

Es fácil ver que las unidades para las que se cuenta sólo con un periodo no desempeñan función alguna en un análisis de efectos fijos. La deducción del tiempo para tales observaciones genera sólo ceros, que no se emplean en la estimación. (Si T_i es a lo sumo dos para toda i, es posible aplicar la primera diferencia: si $T_i = 1$ para alguna i, no se tienen dos periodos a diferenciar.)

El problema más difícil que se presenta con un panel no balanceado es determinar por qué está fuera de balance. Con ciudades y estados, por ejemplo, se carece de datos sobre las variables clave de ciertos años. Mientras no se correlacione la ausencia de datos de una determinada i con los errores idiosincráticos, u_{it}, el panel no balanceado no genera problemas. Cuando se cuenta con datos sobre personas, familias o empresas, las cosas se complican. Imagine, por ejemplo, que se tiene una muestra aleatoria de empresas manufactureras en 1990 y que es de interés probar la forma en que la sindicalización influye en la rentabilidad de las empresas. Idealmente, es posible aprovechar un análisis de datos de panel para tomar en cuenta características inobservables del personal y administrativas que influyan en la rentabilidad, y que tal vez pudieran estar correlacionadas con la fracción del personal de la empresa que está sindicalizado. Si se recaban de nuevo datos de años subsiguientes, tal vez se pierdan algunas empresas por haberse salido del mercado o por haberse fusionado con otras. De ser así, es probable que se tenga una muestra no aleatoria en los periodos posteriores. La pregunta es: si se aplican efectos fijos al panel no balanceado, ¿cuándo serán insesgados los estimadores (o al menos consistentes)?

Si la razón por la que una empresa deja la muestra (denominada *desaparición*) se correlaciona con el error idiosincrático, es decir, aquellos factores inobservables que cambian en el tiempo e influyen en las ganancias, entonces el problema resultante de la selección muestral (vea el capítulo 9) puede causar estimadores sesgados. Se trata de una consideración seria en este ejemplo. No obstante, algo útil acerca de un análisis de efectos fijos es que *permite* que la desaparición se correlacione con a_i, el efecto inobservable. La idea es que, con el muestreo inicial, algunas unidades tengan más probabilidad de abandonar la muestra y esto lo capta a_i.

Ejemplo 14.3

[Efecto de la capacitación laboral en la tasa de desperdicios industriales]

Se agregan dos variables al análisis de la tabla 14.1: $\log(sales_{it})$ y $\log(employ_{it})$, donde *sales* son las ventas anuales de la empresa y *employ* es el número de empleados. Tres de las 54 empresas abandonaron el análisis debido a que no cuentan con datos de las ventas o de los empleados. Cinco observaciones adicionales se perdieron porque faltaban datos sobre una o dos de estas variables durante algunos años, quedando $n = 148$. El uso de efectos fijos en el panel no balanceado no cambia el planteamiento básico, aun cuando el efecto estimado del subsidio se vuelva más grande: $\hat{\beta}_{grant} = -.297$, $t_{grant} = -1.89$; $\hat{\beta}_{grant_{-1}} = -.536$, $t_{grant_{-1}} = -2.389$.

Resolver problemas generales de desaparición con datos de panel es complicado y fuera del alcance de este libro. [Vea, por ejemplo, Wooldridge (2002, capítulo 17).]

14.2 Modelos de efectos aleatorios

Se comienza con el mismo modelo de efectos inobservables que antes,

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad \textbf{14.7}$$

donde se incluye, de manera explícita, un intercepto de modo que se puede suponer que el efecto inobservable, a_i, tiene media cero (sin pérdida de generalidad). Por lo común, también se permitirían las variables binarias temporales entre las variables explicativas. Cuando se utilizan efectos fijos o las primeras diferencias, el objetivo es eliminar a_i porque se considera que está correlacionada con una o más de las x_{itj}. Pero suponga que a_i *no está correlacionada* con ninguna variable explicativa en todos los periodos. Entonces, el uso de una transformación para eliminar a_i da como resultado estimadores ineficientes.

La ecuación (14.7) se vuelve un **modelo de efectos aleatorios** cuando se da por sentado que el efecto inobservable a_i no se correlaciona con ninguna variable explicativa:

$$\mathrm{Cov}(x_{itj}, a_i) = 0, \quad t = 1, 2, \ldots, T; j = 1, 2, \ldots, k. \quad \textbf{14.8}$$

De hecho, los supuestos ideales de los efectos aleatorios incluyen todos los supuestos de efectos fijos más el requisito adicional de que a_i es independiente de todas las variables explicativas en todos los periodos. (En el apéndice del capítulo se proporcionan los supuestos que en realidad se utilizan.) Si se piensa que el efecto inobservable a_i se correlaciona con alguna o algunas variables explicativas, se debe utilizar las primeras diferencias o los efectos fijos.

Bajo la ecuación (14.8), y junto con los supuestos de efectos aleatorios, ¿cómo se deben estimar las β_j? Es importante darse cuenta de que si se cree que a_i no se correlaciona con las variables explicativas, las β_j pueden estimarse de manera consistente al usar un solo corte transversal: no hay necesidad de un panel de datos en lo absoluto. Pero al utilizar un solo corte transversal se ignora mucha información útil de los otros periodos. También es posible usar los datos en un procedimiento combinado de MCO: tan solo realice MCO de y_{it} sobre las variables explicativas y probablemente las variables binarias temporales. Esto también produce estimadores consistentes de las β_j bajo el supuesto de efectos aleatorios. Pero ignora una característica fundamental del modelo. Si se define el **término de error compuesto** como $v_{it} = a_i + u_{it}$, entonces la ecuación (14.7) puede escribirse como

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + v_{it}. \tag{14.9}$$

Como a_i está en el error compuesto en cada periodo, las v_{it} se correlacionan serialmente en cada periodo de tiempo. De hecho, bajo los supuestos de efectos aleatorios,

$$\mathrm{Corr}(v_{it}, v_{is}) = \sigma_a^2/(\sigma_a^2 + \sigma_u^2), \quad t \neq s,$$

donde $\sigma_a^2 = \mathrm{Var}(a_i)$ y $\sigma_u^2 = \mathrm{Var}(u_{it})$. Esta correlación serial (necesariamente) positiva en el término de error puede ser sustancial y, debido a que los errores estándar usuales de MCO combinados ignoran esta correlación, serán incorrectos, como lo serán los estadísticos de prueba usuales. En el capítulo 12 se mostró cómo los mínimos cuadrados generalizados se pueden utilizar para estimar modelos con correlación serial autorregresiva. También es posible usar MCG aquí para resolver los problemas de correlación serial. Para que el procedimiento tenga buenas propiedades, se debe tener una N grande y un T relativamente pequeño. Se supone que se tiene un panel balanceado, aunque el método puede extenderse a paneles no balanceados.

Para obtener la transformación de MCG que elimina la correlación serial en los errores se requiere de álgebra matricial compleja [vea, por ejemplo, Wooldridge (2002, capítulo 10)]. Pero la transformación en sí es sencilla. Defínase

$$\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}, \tag{14.10}$$

que está entre cero y uno. Por tanto, la ecuación transformada resulta ser

$$\begin{aligned} y_{it} - \lambda\bar{y}_i &= \beta_0(1 - \lambda) + \beta_1(x_{it1} - \lambda\bar{x}_{i1}) + \ldots \\ &\quad + \beta_k(x_{itk} - \lambda\bar{x}_{ik}) + (v_{it} - \lambda\bar{v}_i), \end{aligned} \tag{14.11}$$

donde la barra superior vuelve a indicar los promedios a lo largo del tiempo. Esta ecuación es muy interesante, ya que involucra **datos cuasi deducidos** en cada variable. El estimador de efectos fijos resta los promedios de los periodos de la variable correspondiente. La transformación de efectos aleatorios resta una fracción de ese promedio a lo largo del tiempo, donde la fracción depende de σ_u^2, σ_a^2, y el número de periodos, T. El estimador de MCG es simplemente el estimador combinado de MCO de la ecuación (14.11). No es muy evidente que los errores en la ecuación (14.11) ya no están serialmente correlacionados, pero en efecto no lo están. (Vea el problema 14.3.)

La transformación de la ecuación (14.11) permite variables explicativas que son constantes en el tiempo, y ésta es una ventaja de los efectos aleatorios (EA) sobre, ya sea los efectos fijos o las primeras diferencias. Esto es posible debido a que EA supone que el efecto inobservable no está correlacionado con ninguna de las variables explicativas, ya sea que las variables explicativas estén fijas en el tiempo o no. Por tanto, en una ecuación de salario, es posible incluir una variable como la educación, incluso si ésta no cambia con el tiempo. Pero se está dando por sentado que la educación no se correlaciona con a_i, la cual contiene la capacidad innata y los antecedentes familiares. En muchas aplicaciones, la razón para utilizar datos de panel es permitir que el efecto inobservable se correlacione con las variables explicativas.

El parámetro λ nunca se conoce en la práctica, pero siempre puede estimarse. Existen diferentes maneras de hacerlo, mismas que pueden basarse en MCO combinados o en efectos fijos, por ejemplo. Por lo general, $\hat{\lambda}$ toma la forma $\hat{\lambda} = 1 - \{1/[1 + T(\hat{\sigma}_a^2/\hat{\sigma}_u^2)]\}^{1/2}$, donde $\hat{\sigma}_a^2$ es un estimador consistente de σ_a^2 y $\hat{\sigma}_u^2$ es un estimador consistente de σ_u^2. Estos estimadores pueden basarse en los residuales de MCO combinados o de efectos fijos. Una posibilidad es que

$\hat{\sigma}_a^2 = [NT(T-1)/2 - (k+1)]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it}\hat{v}_{is}$, donde los $\hat{v}_{it}$ son los residuales de la estimación combinada de MCO (14.9). Dado esto, es posible estimar σ_u^2 mediante $\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_a^2$, donde $\hat{\sigma}_v^2$ es el cuadrado del error estándar usual de la regresión combinada de MCO. [Vea Wooldridge (2002, capítulo 10) para obtener un análisis adicional de estos estimadores.]

Numerosos paquetes econométricos soportan la estimación de los modelos de efectos aleatorios y calculan de forma automática alguna variación de $\hat{\lambda}$. El estimador de MCG factibles que utiliza $\hat{\lambda}$ en lugar de λ se denomina **estimador de efectos aleatorios**. De acuerdo con los supuestos de los efectos aleatorios del apéndice de este capítulo, el estimador es consistente (no insesgado) y tiene una distribución normal asintótica a medida que N aumenta con T fijo. Se desconocen en buena parte las propiedades del estimador de EA cuando N es pequeña y T es grande, aunque desde luego se ha usado en tales situaciones.

La ecuación (14.11) permite relacionar el estimador de EA tanto con un estimador de MCO combinados como con uno de efectos fijos. El primero se obtiene cuando $\lambda = 0$, y el segundo cuando $\lambda = 1$. En la práctica, la estimación $\hat{\lambda}$ nunca es cero ni uno. Pero si $\hat{\lambda}$ es cercano a cero, las estimaciones de EA estarán próximas a las estimaciones combinadas de MCO. Esto ocurre cuando el efecto inobservable, a_i, tiene relativamente poca importancia (ya que su varianza es pequeña respecto a σ_u^2). Es más común que σ_a^2 sea grande respecto a σ_u^2, en cuyo caso $\hat{\lambda}$ estará mucho más cerca de uno. A medida que T es mayor, $\hat{\lambda}$ tiende a uno, y esto hace que las estimaciones de EA y de EF sean muy parecidas.

Se puede lograr una mejor comprensión de los méritos relativos de los efectos aleatorios, en comparación con los efectos fijos, al escribir el error cuasi deducido de la ecuación (14.11) como $v_{it} - \lambda\bar{v}_i = (1-\lambda)a_i + u_{it} - \lambda\bar{u}_i$. Esta simple expresión hace más evidente que los errores en la ecuación transformada que se utilizó en la estimación de los efectos aleatorios pondera el efecto inobservable por $(1-\lambda)$. Aunque la correlación entre a_i y uno o más x_{itj} provoca inconsistencia en la estimación de los efectos aleatorios, se observa que la correlación se atenúa por el factor $(1-\lambda)$. Conforme $\lambda \to 1$, el término del sesgo tiende a cero, como debe ser, ya que el estimador de EA tiende al estimador de EF. Si λ está cerca de cero, se está dejando una fracción más grande del efecto inobservable en el término de error y, en consecuencia, el sesgo asintótico del estimador de EA será mayor.

En las aplicaciones de EF y EA por lo general también es informativo calcular las estimaciones combinadas de MCO. La comparación de tres conjuntos de estimaciones puede ayudar a determinar la naturaleza de los sesgos provocados por dejar el efecto inobservable, a_i, enteramente en el término de error (como lo hacen MCO combinados) o parcialmente en el término de error (como lo hace la transformación de EA). Pero se debe recordar que, incluso si a_i no se correlaciona con ninguna de las variables explicativas en todos los periodos, los errores estándar y estadísticos de prueba de MCO combinados por lo común no son válidos: ignoran la, a menudo, importante correlación serial en los errores compuestos, $v_{it} = a_i + u_{it}$. Como se mencionó en el capítulo 13 (vea el ejemplo 13.9), es posible calcular los errores estándar y los estadísticos de prueba que son robustos a la correlación serial arbitraria (y a la heterocedasticidad) en v_{it}, y los paquetes de estadística conocidos con frecuencia cuentan con esta opción. [Vea, por ejemplo, Wooldridge (2002, capítulo 10).]

Ejemplo 14.4

[Una ecuación de salario que utiliza datos de panel]

Se utilizan de nuevo los datos de WAGEPAN.RAW para estimar la ecuación de salario para los hombres. Se emplean tres métodos de estimación: combinada de MCO, de efectos aleatorios y de efectos fijos. En los dos primeros es posible incluir *educ* y binarias de raza (*black* e *hispan*), pero estas variables están fuera del ámbito del análisis de efectos fijos. Las variables que cambian con el tiempo son *exper*, *exper*^{2}, *union*

TABLA 14.2

Tres estimadores diferentes de una ecuación de salario

Variable dependiente: log(*wage*)			
Variables independientes	**MCO combinados**	**Efectos aleatorios**	**Efectos fijos**
educ	.091 (.005)	.092 (.011)	———
black	−.139 (.024)	−.139 (.048)	———
hispan	.016 (.021)	.022 (.043)	———
exper	.067 (.014)	.106 (.015)	———
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

y *married*. Como se planteó en la sección 14.1, *exper* (experiencia) se elimina en el análisis de EF (pero $exper^2$ permanece). Cada regresión contiene también un conjunto completo de binarias anuales. Los resultados de la estimación se presentan en la tabla 14.2.

Los coeficientes de *educ*, *black* e *hispan* son similares en las estimaciones de MCO combinados y de efectos aleatorios. Los errores estándar de MCO combinados son los errores estándar usuales de MCO, y éstos subestiman los verdaderos errores estándar, ya que ignoran la correlación serial positiva; aquí se reportan únicamente para compararlos. El perfil del efecto de la experiencia es un poco diferente, y tanto el impacto del matrimonio (*married*) como del sindicato (*union*) en particular disminuyen notablemente en la estimación de efectos aleatorios. Cuando se elimina el efecto inobservable por completo mediante efectos fijos, el impacto del matrimonio decrece aproximadamente a 4.7%, aunque sigue siendo estadísticamente significativo. La disminución en el impacto del matrimonio es consistente con la idea de que los hombres más capaces —según lo capta el efecto inobservable más alto, a_i— tienen más posibilidades de estar casados. Por tanto, en la estimación combinada de MCO, una gran parte del efecto del matrimonio se refleja en el hecho de que los varones que están casados ganarían más que si no lo estuvieran. El restante 4.7% tuvo al menos dos explicaciones posibles: 1) el matrimonio realmente hace que los hombres sean más productivos o 2) los empleadores pagan a los hombres casados una prima debido a que el matrimonio es una señal de mayor estabilidad. No es posible distinguir entre estas dos hipótesis.

La λ propuesta para la estimación de efectos aleatorios es $\hat{\lambda} = .643$, lo que explica por qué, en las variables que se modifican con el tiempo, las estimaciones de EA se encuentran más cerca de las estimaciones de EF que de las de MCO combinados.

Pregunta 14.3

El impacto estimado de la afiliación a un sindicato mediante efectos fijos es de alrededor de 10 puntos porcentuales menor que el impacto de la estimación por MCO. ¿Qué sugiere esto marcadamente respecto a la correlación entre *union* y el efecto inobservable?

¿Efectos aleatorios o efectos fijos?

Dado que los efectos fijos permiten una correlación arbitraria entre a_i y las x_{itj}, mientras que los efectos aleatorios no, se considera ampliamente que los EF constituyen una herramienta más convincente para la estimación de los efectos *ceteris paribus*. No obstante, los efectos aleatorios se aplican en ciertas situaciones. Evidentemente, si la variable explicativa clave es constante en el tiempo, no es posible usar EF para estimar su efecto sobre y. Por ejemplo, en la tabla 14.2, la estimación del rendimiento de la educación se debe basar en EA (o MCO combinados). Desde luego, sólo se puede utilizar efectos aleatorios debido a que se está dispuesto a suponer que el efecto inobservable no se correlaciona con ninguna de las variables explicativas. Por lo común, si uno utiliza efectos aleatorios, se incluyen tantos controles constantes en el tiempo como sea posible entre las variables explicativas. (En un análisis de EF, no es necesario incluir estos controles). EA es preferible a una estimación combinada de MCO debido a que, por lo general, EA es más eficiente.

Si lo que interesa está en una variable explicativa que cambia con el tiempo, ¿existe un caso en el que se utilice EA en vez de EF? Sí, pero las situaciones en las cuales $\text{Cov}(x_{itj}, a_i) = 0$ deben considerarse más la excepción que la regla. Si la variable de política clave se establece de forma experimental, por ejemplo, cada año los niños se asignan aleatoriamente a clases de diferente tamaño, entonces los efectos aleatorios serían apropiados para estimar el efecto del tamaño de la clase sobre el desempeño. Por desgracia, en la mayoría de los casos los regresores son por sí mismos resultado de los procesos de elección y es probable que se correlacionen con las preferencias y capacidades individuales capturadas en a_i.

Sigue siendo muy común ver a investigadores que aplican tanto los efectos fijos como los efectos aleatorios, y luego prueban de manera formal las diferencias estadísticamente significativas en los coeficientes de las variables explicativas que cambian con el tiempo. (Así que en la tabla 14.2 éstos serían los coeficientes de *exper*^{2}, *married* y *union*.) Hausman (1978) propuso por primera vez este tipo de prueba, y algunos paquetes econométricos calculan automáticamente la prueba de Hausman bajo el conjunto completo de los supuestos de efectos aleatorios, el cual aparece en forma de lista en el apéndice del capítulo. La idea es utilizar las estimaciones de efectos aleatorios a menos que la prueba de Hausman lo rechace (14.8). En la práctica, si no hay rechazo, significa que las estimaciones de EA y de EF están lo suficientemente cerca para que no importe cuál usar, o bien que la variación de muestreo es tan grande en las estimaciones de EF que no se puede concluir que las diferencias que son significativas desde el punto de vista práctico son estadísticamente significativas. En el último caso, a uno le resta preguntarse si existe suficiente información en los datos para proporcionar estimaciones precisas de los coeficientes. Un rechazo mediante la prueba de Hausman significa que el supuesto clave de EA, (14.8), es falso y por tanto, se usan las estimaciones de EF. (Como es natural, al igual que en todas las aplicaciones de la inferencia estadística, se debe hacer una distinción entre una diferencia significativa en la práctica y una diferencia estadísticamente significativa.) [Vea Wooldridge (2002, sección 10.7) para un análisis más a fondo.]

Una advertencia final, al leer trabajos empíricos tal vez se encuentre con que algunos autores prefieren la estimación de EF a la de EA con base en si los a_i se consideran propiamente como parámetros o como variables aleatorias. Estas consideraciones, por lo general, son desatinadas. En este capítulo se han tratado a los a_i como variables aleatorias en el modelo de efectos inobservables (14.7), sin importar cómo se decida estimar las β_j. Como se ha recalcado, el principal aspecto para determinar si se utiliza EF o EA es si es posible suponer convincentemente que a_i no se correlaciona con todas las x_{itj}. Sin embargo, en algunas aplicaciones de los métodos de datos de panel, a nuestra muestra no puede tratarse como una muestra aleatoria de una población grande, en especial cuando la unidad de observación es una unidad geográfica extensa (por ejemplo, estados o provincias). Así que tiene sentido pensar en cada a_i como un intercepto separado a estimar para cada unidad de corte transversal. En este caso se utilizan efectos fijos: recuerde, el uso de EF es mecánicamente el mismo que permitir un intercepto diferente para cada unidad de corte transversal. Por suerte, ya sea que se participe o no en el debate filosófico sobre la naturaleza de a_i, EF casi siempre es mucho más convincente que EA para el análisis de políticas que utilizan datos agregados.

14.3 Aplicación de métodos de datos de panel a otras estructuras de datos

Los diversos métodos de datos de panel pueden aplicarse a ciertas estructuras de datos que no involucran el tiempo. Por ejemplo, en demografía es común el uso de hermanos (en ocasiones gemelos) para considerar características inobservables relacionadas con la familia y los antecedentes. Por lo general, se desea permitir que el "efecto familiar" inobservable, que es común a todos los gemelos de una familia, se correlacione con las variables explicativas observadas. Si dichas variables explicativas cambian entre los hermanos de una familia, tomar la diferencia entre las parejas de hermanos —o, en términos más generales, utilizar la transformación intragrupal dentro de una familia— es preferible como método de estimación. Al eliminar el efecto inobservable, se eliminan los posibles sesgos provocados por confundir las características de los antecedentes familiares. La implementación de efectos fijos en estas estructuras de datos es muy sencilla en los paquetes de regresión que soportan la estimación de EF.

A manera de ejemplo, Geronimus y Korenman (1992) utilizaron parejas de hermanas para estudiar los efectos de la maternidad entre adolescentes sobre variables económicas futuras. Cuando la variable es el ingreso relacionado con las necesidades, algo que depende del número de hijos, el modelo es

$$\log(incneeds_{fs}) = \beta_0 + \delta_0 sister2_s + \beta_1 teenbrth_{fs} + \beta_2 age_{fs} + otros\ factores + a_f + u_{fs}, \tag{14.12}$$

donde f indica la familia y s a una hermana dentro de la familia. El intercepto para la primera hermana es β_0, y para la segunda es $\beta_0 + \delta_0$. La variable de interés es $teenbrth_{fs}$, que es una variable binaria igual a uno si la hermana s de la familia f tuvo un hijo en la adolescencia. La variable age_{fs} es la edad actual de la hermana s en la familia f; Geronimus y Korenman también usaron otros controles. La variable inobservable a_f, que se modifica sólo de una familia a otra, es un *efecto familiar inobservable* o un *efecto familiar fijo*. La preocupación principal en el análisis es que *teenbrth* se correlacione con el efecto familiar. De ser así, un análisis de MCO que combine a familias y hermanas da un estimador sesgado del efecto de la maternidad en la adolescencia sobre los resultados económicos. La resolución de este problema es sencilla: dentro de cada familia, se aplican diferencias a (14.12) entre las hermanas para obtener

$$\Delta\log(incneeds) = \delta_0 + \beta_1\Delta teenbrth + \beta_2\Delta age + \ldots + \Delta u; \tag{14.13}$$

esto elimina el efecto familiar, a_f, y la ecuación resultante se estima por MCO. Advierta que aquí no existe un elemento temporal: la diferencia es entre hermanas dentro de una familia. También se han permitido diferencias en los interceptos entre las hermanas en la ecuación (14.12), lo cual conduce a un intercepto diferente de cero en la ecuación diferenciada (14.13). Si al introducir los datos, el orden de las hermanas en cada familia es en esencia aleatorio, el intercepto estimado debe estar cerca de cero. Pero incluso en tales casos no perjudica incluir un intercepto en la ecuación (14.13) y hacer que el intercepto permita el hecho de que, por ejemplo, la primera hermana listada sea siempre la más necesitada.

Pregunta 14.4

Cuando se utiliza el método de diferenciación, ¿tiene sentido incluir variables binarias para la raza del padre y de la madre en la ecuación (14.12)? Explique por qué.

Geronimus y Korenman, usando 129 parejas de hermanas de la Encuesta Nacional Longitudinal de Mujeres Jóvenes de 1982, estimaron primero β_1 por MCO combinados para obtener $-.33$ o $-.26$, donde la segunda estimación se deriva de controlar las variables de antecedentes familiares (como la educación de los padres); las dos estimaciones son estadísticamente significativas [vea la tabla 3 en Geronimus y Korenman (1992)]. Por tanto, la maternidad en la adolescencia tiene un impacto bastante grande en los ingresos familiares futuros. Sin embargo, cuando se estima la ecuación diferenciada, el coeficiente de *teenbrth* es $-.08$, una cifra pequeña y estadísticamente insignificante. Esto sugiere que en gran parte son los antecedentes familiares de una mujer los que afectan sus ingresos futuros y no la maternidad en la adolescencia.

Geronimus y Korenman estudiaron algunos otros resultados y otras dos bases de datos; en ciertos casos las estimaciones familiares intragrupales fueron en sentido económico grandes y estadísticamente significativas. También mostraron cómo desaparecen por completo los efectos cuando los niveles de educación de las hermanas se incluyen en la regresión.

Ashenfelter y Krueger (1994) utilizaron la metodología de la diferenciación para estimar el rendimiento de la educación. Obtuvieron una muestra de 149 gemelos idénticos y reunieron información sobre los ingresos, la educación y otras variables. Consideraron gemelos idénticos porque éstos deben tener la misma capacidad subyacente, la que se puede eliminar utilizando las diferencias entre los gemelos, en vez de usar MCO sobre los datos combinados. Como los gemelos idénticos tienen la misma edad, género y raza, todos estos factores quedan fuera de la ecuación diferenciada. Por consiguiente, Ashenfelter y Krueger realizaron una regresión de la diferencia en log(*earnings*) sobre la diferencia en la educación y estimaron que el rendimiento de la educación es aproximadamente de 9.2% ($t = 3.83$). Es interesante que en realidad esto sea *mayor* que la estimación combinada de MCO de 8.4% (la cual toma en cuenta el género, la edad y la raza). Ashenfelter y Krueger también estimaron la ecuación por efectos aleatorios y obtuvieron 8.7% como rendimiento de la educación. (Vea la tabla 5 en su artículo.) El análisis de efectos aleatorios es mecánicamente el mismo que el caso de los datos de panel con dos periodos.

Las muestras utilizadas por Geronimus y Korenman (1992) y Ashenfelter y Krueger (1994) son ejemplos de **muestras de datos apareados**. Por lo común, los métodos de efectos fijos y de efectos aleatorios pueden aplicarse a una **muestra de agrupamientos**. los cuales se componen de conjuntos de datos de corte transversal, pero cada observación pertenece a un agrupamiento bien definido. En los ejemplos anteriores, cada familia es un agrupamiento. A manera de un ejemplo adicional, imagine que se tienen datos de la participación en diversos planes de pensión, y las empresas ofrecen más de un plan. De esta manera es posible considerar a una empresa como un agrupamiento y está muy claro que los efectos inobservables de la empresa serían un factor importante al determinar las tasas de participación en los planes de pensión dentro de la empresa.

Los datos sobre educación de estudiantes muestreados de numerosas escuelas conforman una muestra de agrupamientos, donde cada escuela es un agrupamiento. Dado que es probable que los resultados dentro de un agrupamiento estén correlacionados, por lo general, es importante tener en cuenta un efecto inobservable de los agrupamientos. La estimación de efectos fijos es preferida cuando se piensa que un **efecto de grupo** inobservable —un ejemplo del cual es a_f en la ecuación (14.12)— se correlaciona con una o más de las variables explicativas. Por tanto, sólo es posible incluir variables explicativas que cambien, cuando menos un poco, dentro de los agrupamientos. Los tamaños de agrupamiento rara vez son iguales, así que, por lo general, se requieren métodos de efectos fijos para los paneles no balanceados.

En algunos casos las variables explicativas —con frecuencia variables de políticas— sólo cambian en el nivel de agrupamiento, no dentro del agrupamiento. En estos casos, el método de efectos fijos no aplica. Por ejemplo, tal vez interesen los efectos de la calidad medida del profesor sobre el desempeño de los estudiantes, donde cada agrupamiento es un aula de una escuela elemental. Dado que todos los estudiantes dentro de un agrupamiento tienen el mismo profesor, la eliminación de un "efecto de clase" también elimina cualquier medida observada de la calidad del profesor. Si se tienen buenos controles en la ecuación, tal vez esté justificada la aplicación de efectos aleatorios al agrupamiento no balanceado. Al igual que con los datos de panel, el requisito fundamental para que EA produzca estimaciones convincentes es que las variables explicativas no estén correlacionadas con el efecto de grupo inobservable. La mayoría de los paquetes econométricos permiten la estimación de efectos aleatorios sobre agrupamientos no balanceados sin mucho esfuerzo.

También es común la aplicación de MCO combinados a las muestras de agrupamientos cuando la eliminación de un efecto de grupo, por medio de efectos fijos, no es viable ni deseable. Sin embargo, al igual que con los datos de panel, los errores estándar usuales de MCO son incorrectos a menos que no haya un efecto de grupo y, por tanto, se deben utilizar los errores estándar robustos que permiten la "correlación de agrupamientos" (y heterocedasticidad). Algunos paquetes de regresión cuentan con instrucciones sencillas para corregir los errores estándar y los estadísticos de prueba usuales para una correlación intragrupal general dentro del agrupamiento (así como la heterocedasticidad). Éstas son las mismas correcciones que funcionan para las estimaciones combinadas de MCO sobre conjuntos de datos de panel, mismos que se muestran en el ejemplo 13.9. A modo de ejemplo, Papke (1999) estima modelos de probabilidad lineal para la continuación de los planes de prestaciones para jubilados definidos con base en si las empresas adoptaron planes de contribución definidos. Ya que es posible tener un efecto de la empresa que induzca la correlación entre distintos planes dentro la misma, Papke corrige los errores estándar usuales de MCO para el muestreo de agrupamientos, así como para la heterocedasticidad en el modelo de probabilidad lineal.

RESUMEN

Se han estudiado dos métodos comunes para estimar modelos de datos de panel con efectos inobservables. En comparación con las primeras diferencias, el estimador de efectos fijos es eficiente cuando los errores idiosincráticos no se correlacionan serialmente (y cuando son homocedásticos) y no se hacen suposiciones sobre la correlación entre el efecto inobservable a_i y las variables explicativas. Como en el caso de las primeras diferencias, elimina del análisis cualquier variable explicativa constante en el tiempo. Los métodos de efectos fijos se aplican de inmediato a los paneles no balanceados, pero se debe dar por sentado que las razones por las que faltan algunos periodos no se relacionan de forma sistemática con los errores idiosincráticos.

El estimador de efectos aleatorios es adecuado cuando se considera que el efecto inobservable no se correlaciona con ninguna variable explicativa. Por tanto, a_i se deja en el término de error y se resuelve la correlación serial resultante en el tiempo mediante la estimación de mínimos cuadrados generalizados. Resulta conveniente que MCG factibles se obtengan por medio de una regresión combinada sobre datos cuasi deducidos. El valor del parámetro estimado de transformación, $\hat{\lambda}$, indica cuándo es probable que las estimaciones estén más cerca de las estimaciones combinadas de MCO o de efectos fijos. Si el conjunto completo de supuestos de efectos aleatorios es válido, el estimador de efectos aleatorios es asintóticamente más eficiente, a medida que N aumenta con T fijo, que los estimadores combinados de MCO, de primeras diferencias o de efectos fijos (todos ellos son insesgados, consistentes y asintóticamente normales).

Finalmente, los métodos de datos de panel estudiados en los capítulos 13 y 14 se pueden aplicar cuando se trabaja con muestras de datos apareados o de agrupamientos. La diferenciación o la transformación intragrupal eliminan el efecto de grupo. Si dicho efecto no se correlaciona con las variables explicativas, se emplean los estimadores combinados de MCO, aunque los errores estándar y los estadísticos de prueba deban ajustarse para la correlación de agrupamientos. La estimación vía efectos aleatorios también es una posibilidad.

TÉRMINOS CLAVE

Datos con el tiempo deducido
Datos cuasi deducidos
Efecto de grupo
Estimador de efectos aleatorios
Estimador de efectos fijos
Estimador intragrupal (*within*)
Muestra de agrupamientos
Modelo de efectos aleatorios
Modelo de efectos inobservables
Muestras de datos apareados
Panel no balanceado
Regresión de variables binarias
Término de error compuesto
Transformación de efectos fijos
Transformación intragrupal

PROBLEMAS

14.1 Imagine que los errores idiosincráticos en la ecuación (14.4), $\{u_{it}: t = 1, 2, \dots, T\}$, no se correlacionan serialmente y tienen varianza constante, σ_u^2. Demuestre que la correlación entre las diferencias contiguas, Δu_{it} y $\Delta u_{i,t+1}$, es $-.5$. Por tanto, con base en los supuestos ideales de EF, la primera diferencia induce a una correlación serial negativa de un valor conocido.

14.2 Con una sola variable explicativa, la ecuación empleada para obtener el estimador intergrupal es

$$\bar{y}_i = \beta_0 + \beta_1\bar{x}_i + a_i + \bar{u}_i,$$

donde la barra representa el promedio a lo largo del tiempo. Suponga que $E(a_i) = 0$ porque se ha incluido un intercepto en la ecuación. Suponga que $\bar{u}_i$ no se correlaciona con $\bar{x}_i$, pero la $Cov(x_{it}, a_i) = \sigma_{xa}$ para toda t (y para toda i debido al muestreo aleatorio en el corte transversal).

i) Utilizando $\tilde{\beta}_1$ como el estimador intragrupal, es decir, el estimador de MCO que utiliza los promedios a lo largo del tiempo, demuestre que

$$\text{plim}\,\tilde{\beta}_1 = \beta_1 + \sigma_{xa}/\text{Var}(\bar{x}_i),$$

donde el límite de probabilidad se define conforme $N \to \infty$. [*Sugerencia:* observe las ecuaciones (5.5) y (5.6).]

ii) Suponga además que las x_{it}, para toda $t = 1, 2, \dots, T$, no se correlacionan y tienen varianza constante σ_x^2. Demuestre que $\text{plim}\,\tilde{\beta}_1 = \beta_1 + T(\sigma_{xa}/\sigma_x^2)$.

iii) Si las variables explicativas no tienen una correlación marcada en el tiempo, ¿qué sugiere el inciso ii) respecto a si la inconsistencia del estimador intragrupal es menor cuando hay más periodos?

14.3 En un modelo de efectos aleatorios, defina el error compuesto $v_{it} = a_i + u_{it}$, en el que a_i no se correlaciona con u_{it} y las u_{it} tienen varianza constante σ_u^2 y no están correlacionadas serialmente. Defina $e_{it} = v_{it} - \lambda\bar{v}_i$, donde λ se proporciona en la ecuación (14.10).

i) Muestre que $E(e_{it}) = 0$.

ii) Muestre que $Var(e_{it}) = \sigma_u^2$, $t = 1, \dots, T$.

iii) Muestre que para $t \neq s$, $Cov(e_{it}, e_{is}) = 0$.

14.4 Para determinar los efectos del desempeño atlético sobre los aspirantes a ingresar a la universidad, se reúnen datos sobre las solicitudes para una muestra de las universidades de la División I de 1985, 1990 y 1995.

i) ¿Qué medidas del éxito atlético incluiría en una ecuación? ¿Qué problemas hay acerca de la temporalidad de los datos?

ii) ¿Qué otros factores se podrían incluir en la ecuación?

iii) Escriba una ecuación que le permita estimar los efectos del éxito atlético sobre el cambio porcentual en las solicitudes. ¿Cómo se estimaría esta ecuación? ¿Por que se elegiría este método?

14.5 Suponga que, para un semestre, se pueden recabar los siguientes datos en una muestra aleatoria de estudiantes universitarios pertenecientes al primer y último grado, para cada clase tomada: las calificaciones de un examen final estandarizado, el porcentaje de clases a las que asistieron, una variable binaria que indica si la clase está dentro del campo de especialización del estudiante, el promedio acumulado de calificaciones antes del inicio del semestre y las calificaciones del examen SAT de admisión a la universidad.

i) ¿Por qué clasificaría este conjunto de datos como una muestra de agrupamientos? Aproximadamente, ¿cuántas observaciones esperaría para el estudiante típico?

ii) Escriba un modelo parecido al de la ecuación (14.12), que explique el desempeño en el examen final en función de la asistencia y otras características. Utilice s como subíndice de estudiante y c como subíndice de clase. ¿Qué variables no cambian en un estudiante?

iii) Si se combinan todos los datos y se utilizan MCO, ¿qué se está suponiendo sobre las características no observadas de los estudiantes que afectan la tasa de desempeño y asistencia? ¿Qué papel juegan, a este respecto, las calificaciones del examen SAT y el promedio de calificaciones anterior?

iv) Si piensa que las calificaciones del examen SAT y el promedio de calificaciones anterior no son adecuados para capturar la capacidad de los estudiantes, ¿cómo se estimaría el efecto de la asistencia sobre el desempeño del examen final?

14.6 Utilizando la opción de "agrupamiento" (*cluster*) en el paquete econométrico Stata®, los errores estándar completamente robustos para las estimaciones combinadas de MCO de la tabla 14.2, es decir, robustos a la correlación serial y a la heterocedasticidad en los errores compuestos, $\{v_{it}: t = 1, \ldots, T\}$, se obtienen como ee($\hat{\beta}_{educ}$) = .011, ee($\hat{\beta}_{black}$) = .051, ee($\hat{\beta}_{hispan}$) = .039, ee($\hat{\beta}_{exper}$) = .020, ee($\hat{\beta}_{exper^2}$) = .0010, ee($\hat{\beta}_{married}$) = .026 y ee($\hat{\beta}_{union}$) = .027.

i) ¿Cómo se comparan estos errores estándar, por lo general, con aquéllos no robustos, y por qué?

ii) ¿Cómo se comparan los errores estándar robustos de MCO combinados con los errores estándar de EA? ¿Tiene importancia si la variable explicativa es constante en el tiempo o varía con el tiempo?

EJERCICIOS EN COMPUTADORA

C14.1 Los datos de RENTAL.RAW son utilizados para este ejercicio. Estos datos sobre los precios de la renta y otras variables para las ciudades universitarias son para los años 1980 y 1990. La idea es ver si una presencia más fuerte de estudiantes afecta las tasas de alquiler. El modelo de efectos inobservables es

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

donde *pop* es la población de la ciudad, *avginc* es el ingreso promedio y *pctstu* es la población estudiantil como un porcentaje de la población de la ciudad (durante el año escolar).

i) Estime la ecuación por MCO combinados e informe los resultados de la manera usual. ¿Qué conclusión se obtiene de la estimación de la variable binaria de 1990? ¿Qué se obtiene para $\hat{\beta}_{pctstu}$?

ii) ¿Son válidos los errores estándar que se reportaron en el inciso i)? Explique por qué.

iii) Ahora aplique las diferencias a la ecuación y realice una estimación combinada de MCO. Compare la estimación de β_{pctstu} con aquélla del inciso i). ¿El tamaño relativo de la población estudiantil parece afectar los precios de alquiler?

iv) Estime el modelo por efectos fijos para verificar que se obtienen estimaciones y errores estándar idénticos a aquéllos del inciso iii).

C14.2 Utilice los datos en CRIME4.RAW para este ejercicio.

i) Vuelva a estimar el modelo de efectos inobservables para la delincuencia del ejemplo 13.9, pero ahora use efectos fijos en vez de diferenciación. ¿Existe un cambio notable en el signo o en la magnitud de los coeficientes? ¿Qué sucede con la significancia estadística?

ii) Añada los logaritmos de cada variable de salarios en la base de datos y estime el modelo por efectos fijos. ¿Cómo afecta la inclusión de estas variables a los coeficientes de las variables de justicia penal del inciso i)?

iii) ¿Tienen todas las variables de salario del inciso ii) el signo esperado? Explique los resultados. ¿Son conjuntamente significativas?

C14.3 En este ejercicio se utiliza la base de datos JTRAIN.RAW para determinar el efecto del subsidio a la capacitación laboral en las horas de capacitación por empleado. El modelo básico para los tres años es

$$hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}.$$

i) Estime la ecuación mediante efectos fijos. ¿Cuántas empresas se utilizan en la estimación de EF? ¿Cuántas observaciones en total se usarían si cada empresa tuviera datos sobre todas las variables (en particular, *hrsemp*) para los tres años?

ii) Interprete el coeficiente de *grant* (binaria igual a uno si recibió el subsidio) y comente sobre su significancia.

iii) ¿Resulta sorprendente que $grant_{-1}$ sea insignificante? Explique por qué.

iv) ¿Las empresas más grandes dan más o menos capacitación a sus empleados, en promedio? ¿Qué tan grandes son las diferencias? (Por ejemplo, si una empresa tiene 10% más de empleados, ¿cuál es el cambio en el promedio de horas de capacitación?).

C14.4 En el ejemplo 13.8, se utilizaron los datos de los reclamos del seguro de desempleo de Papke (1994) para estimar el efecto que tienen sobre ellos las zonas empresariales. Papke también utiliza un modelo que permite que cada ciudad tenga su propia tendencia en el tiempo:

$$\log(uclms_{it}) = a_i + c_i t + \beta_1 ez_{it} + u_{it},$$

donde a_i y c_i son efectos inobservables. Esto permite que haya más heterogeneidad entre las ciudades.

i) Demuestre que cuando se aplica la primera diferencia a la ecuación anterior, se obtiene

$$\Delta\log(uclms_{it}) = c_i + \beta_1 \Delta ez_{it} + \Delta u_{it},\ t = 2, \ldots, T.$$

Observe que la ecuación diferenciada contiene un efecto fijo, c_i.

ii) Estime la ecuación diferenciada mediante efectos fijos. ¿Cuál es la estimación de β_1? ¿Es muy distinta de la obtenida en el ejemplo 13.8? ¿El efecto de las zonas empresariales sigue siendo estadísticamente significativo?

iii) Agregue un conjunto completo de variables binarias anuales a la estimación del inciso ii). ¿Qué pasa con la estimación de β_1?

C14.5 i) En la ecuación de salario del ejemplo 14.4, explique por qué las variables binarias de ocupación podrían ser variables omitidas para estimar el efecto de la prima salarial por estar afiliado a un sindicato.

ii) Si cada hombre en la muestra se mantuviera en la misma ocupación de 1981 a 1987, ¿necesitaría incluir las binarias de ocupación en una estimación de efectos fijos? Explique.

iii) Con los datos de WAGEPAN.RAW, incluya ocho de las variables binarias sobre ocupación en la ecuación y estímela por efectos fijos. (Las variables de ocupación son *occ1*, ..., *occ9*.) ¿Cambia por mucho el coeficiente de *union*? ¿Qué pasa con su significancia estadística?

C14.6 Agregue el término de interacción $union_{it} \cdot t$ a la ecuación estimada de la tabla 14.2 para ver si el *growth* de salario depende de la afiliación sindical. Estime la ecuación mediante efectos aleatorios y fijos, y compare los resultados.

C14.7 Utilice los datos a nivel de estado sobre las tasas de homicidios y ejecuciones de la base de datos MURDER.RAW para el ejercicio siguiente.

i) Considere el modelo de efectos inobservables

$$mrdrte_{it} = \theta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it},$$

donde θ_t denota sencillamente los interceptos de diferentes años y a_i es el efecto inobservable del estado. Si las ejecuciones pasadas de asesinos convictos tienen un efecto disuasivo, ¿cuál debe ser el signo de β_1? ¿Qué signo piensa usted que tendría β_2 Explique por qué.

ii) Usando sólo los años 1990 y 1993, haga una estimación combinada de MCO de la ecuación del inciso i). No tome en cuenta el problema de la correlación serial en los errores compuestos. ¿Existe evidencia de un efecto disuasivo?

iii) Ahora, usando 1990 y 1993, estime la ecuación por efectos fijos. Puede emplear las primeras diferencias debido a que sólo se están usando datos de dos años. ¿Existe evidencia de un efecto disuasivo? ¿Qué tan fuerte es?

iv) Calcule el error estándar robusto a la heterocedasticidad para la estimación del inciso iii). Será más fácil usar las primeras diferencias.

v) Encuentre el estado que tiene el número mayor de ejecuciones en 1993. (La variable *exec* son las ejecuciones totales de 1991, 1992 y 1993.) ¿Por cuánto rebasa este valor al valor siguiente más alto?

vi) Estime la ecuación usando las primeras diferencias, omitiendo a Texas del análisis. Calcule los errores estándar usuales y robustos a la heterocedasticidad. Ahora, ¿qué resultados se obtienen? ¿Qué está pasando?

vii) Use los tres años de datos y estime el modelo por efectos fijos. Incluya a Texas en el análisis. Comente el tamaño y la significancia estadística del efecto disuasivo en comparación con utilizar sólo dos años, 1990 y 1993.

C14.8 Utilice la base de datos MATHPNL.RAW para este ejercicio. El propósito es realizar una versión de efectos fijos del modelo de primeras diferencias hecho en el ejercicio para computadora C13.11? El modelo de interés es

$$\begin{aligned} math4_{it} = {} & \delta_1 y94_t + \ldots + \delta_5 y98_t + \gamma_1 \log(rexpp_{it}) + \gamma_2 \log(rexpp_{i,t-1}) \\ & + \psi_1 \log(enrol_{it}) + \psi_2 lunch_{it} + a_i + u_{it}, \end{aligned}$$

donde el primer año disponible (el año base) es 1993 debido a la variable de gastos rezagada.

i) Haga una estimación combinada de MCO del modelo e informe los errores estándar usuales. Debe incluirse un intercepto junto con las variables binarias anuales para permitir que a_i tenga un valor esperado distinto de cero. ¿Cuáles son los efectos estimados de las variables de gastos ($rexpp_{it}$, $rexpp_{i,t-1}$)? Obtenga los residuales de MCO, $\hat{v}_{it}$.

ii) ¿El signo del coeficiente de $lunch_{it}$ (porcentaje de alumnos elegibles para recibir el almuerzo gratuito) es lo que se esperaba? Interprete la magnitud del coeficiente. ¿Se podría decir que la tasa de pobreza del distrito tiene un efecto marcado sobre las tasas de aprobación de exámenes ($math4$)?

iii) Calcule una prueba para la correlación serial AR(1) usando la regresión $\hat{v}_{it}$ sobre $\hat{v}_{i,t-1}$. Se deben utilizar los años 1994 a 1998 en la regresión. Verifique que existe una fuerte correlación serial positiva y explique por qué.

iv) Ahora debe estimarse la ecuación por efectos fijos. ¿La variable de gastos rezagada sigue siendo significativa?

v) ¿Por qué cree que en la estimación de efectos fijos las variables de tamaño de la escuela (*enrol*) y programa de almuerzos (*lunch*) son conjuntamente significativas?

vi) Defina el efecto total, o de largo plazo, del gasto como $\theta_1 = \gamma_1 + \gamma_2$. Use la sustitución $\gamma_1 = \theta_1 - \gamma_2$ para obtener un error estándar para $\hat{\theta}_1$. [*Sugerencia:* la estimación de efectos fijos estándar usando $\log(rexpp_{it})$ y $z_{it} = \log(rexpp_{i,t-1}) - \log(rexpp_{it})$ como variables explicativas debe funcionar bien.]

C14.9 La base de datos PENSION.RAW contiene información sobre los planes de pensión dirigidos por los participantes para los empleados estadounidenses. Algunas de las observaciones son para parejas que pertenecen a la misma familia, de modo que este conjunto de datos constituye una pequeña muestra de agrupamientos (con tamaños de agrupamientos de dos).

i) Ignorando el agrupamiento por familias, utilice MCO para estimar el modelo

$$\begin{aligned} pctstck = {} & \beta_0 + \beta_1 choice + \beta_2 prftshr + \beta_3 female + \beta_4 age \\ & + \beta_5 educ + \beta_6 finc25 + \beta_7 finc35 + \beta_8 finc50 + \beta_9 finc75 \\ & + \beta_{10} finc100 + \beta_{11} finc101 + \beta_{12} wealth89 + \beta_{13} stckin89 \\ & + \beta_{14} irain89 + u, \end{aligned}$$

donde las variables se definen en la base de datos. La variable de mayor interés es *choice*, una variable binaria igual a uno si el empleado tiene una opción en la manera de asignar los fondos de pensión entre diferentes inversiones. ¿Cuál es el efecto estimado de *choice*? ¿Es estadísticamente significativo?

ii) ¿Son importantes las variables de control de ingresos (*finc25*, ..., *finc101*), riqueza (*wealth89*), posesión de acciones (*stckin89*) y posesión de cuenta de retiro (*irain89*)? Explique por qué.

iii) Determine cuántas familias diferentes hay en la base de datos.

iv) Ahora, obtenga los errores estándar para MCO que son robustos a la correlación de agrupamientos dentro de una familia. ¿Difieren mucho de los errores estándar usuales de MCO? ¿Esto sorprende?

v) Estime la ecuación por diferenciación sólo entre los esposos dentro de una familia. ¿Por qué las variables explicativas sobre las que se preguntó en el inciso ii) se omiten en la estimación de primeras diferencias?

vi) ¿Son significativas algunas de las variables explicativas restantes del inciso v)? ¿ Sorprende esto?

C14.10 Use la base de datos AIRFARE.RAW para este ejercicio. Se tiene interés en la estimación del modelo

$$\begin{aligned} \log(fare_{it}) = {} & \theta_t + \beta_1 concen_{it} + \beta_2 \log(dist_i) + \beta_3[\log(dist_i)]^2 \\ & + a_i + u_{it},\ t = 1, \ldots, 4, \end{aligned}$$

donde θ_t significa que se toman en cuenta diferentes interceptos anuales, *fare* es la tarifa promedio de un vuelo aéreo en cierta ruta, *concen* es la participación en el mercado de la aerolínea más grande en esa ruta, *dist* es la distancia en millas del vuelo.

i) Haga una estimación combinada de MCO de la ecuación anterior y asegúrese de incluir binarias anuales. Si $\Delta concen = .10$, ¿cuál es el incremento porcentual estimado en *fare*?

ii) ¿Cuál es el intervalo de confianza usual de MCO al 95% para β_1? ¿Por qué es probable que no sea confiable? Si se tiene acceso a un paquete estadístico que calcule errores estándar completamente robustos, calcule el intervalo de confianza completamente robusto al 95% para β_1. Compárelo con el intervalo usual y comente los resultados.

iii) Describa lo que ocurre con la función cuadrática en log(*dist*). En particular, ¿para qué valor de *dist* se vuelve positiva la relación entre log(*fare*) y *dist*? [*Sugerencia:* averigüe primero el valor del punto de inflexión para log(*dist*), y luego realice la exponenciación.] ¿El punto de inflexión está fuera del rango de los datos?

iv) Estime la ecuación por medio de efectos aleatorios. ¿Cómo cambia la estimación de β_1?

v) Ahora, se debe estimar la ecuación mediante efectos fijos. ¿Cuál es la estimación de EF de β_1? ¿Por qué se parece mucho a la estimación de EA? (*Sugerencia:* ¿Cuánto mide $\hat{\lambda}$ para la estimación de EA?).

vi) Nombre dos características de una ruta (diferentes de la distancia entre paradas) que son capturadas por a_i. ¿Podrían estar correlacionadas con $concen_{it}$?

vii) ¿Está usted convencido de que una mayor concentración en una ruta aumenta las tarifas aéreas? ¿Cuál es su mejor estimación?

C14.11 Esta ecuación supone que se cuenta con acceso a un paquete estadístico que calcula errores estándar robustos a la correlación serial arbitraria y a la heterocedasticidad para los métodos de panel de datos.

i) Para las estimaciones combinadas de MCO de la tabla 14.1, obtenga los errores estándar que toman en cuenta la correlación serial arbitraria (en los errores compuestos, $v_{it} = a_i + u_{it}$) y la heterocedasticidad. ¿Qué diferencia hay entre los errores estándar robustos para *educ*, *married*, *union* y los errores no robustos?

ii) Ahora, obtenga los errores estándar robustos para las estimaciones de efectos fijos que permiten correlación serial arbitraria y heterocedasticidad en los errores idiosincráticos, u_{it}. ¿Qué diferencia existe entre éstos y los errores estándar de EF no robustos?

iii) ¿Para cuál método, MCO combinados o EF, es más importante el ajuste de errores estándar para la correlación serial? ¿Por qué?

C14.12 Para responder a esta pregunta utilice la base de datos ELEM94_95.RAW. Los datos se refieren a las escuelas elementales de Michigan. En este ejercicio, se consideran los datos como una muestra de agrupamientos donde cada escuela es parte de un agrupamiento, identificado por el distrito (*distid*).

i) ¿Cuáles son el número mayor y el número menor de escuelas en un distrito? ¿Cuál es el número promedio de escuelas por distrito?

ii) Mediante MCO combinados (es decir, la combinación de las 1,848 escuelas), estime un modelo que relacione *lavgsal* con *bs*, *lenrol*, *lstaff* y *lunch*; vea también el ejercicio para computadora C9.11. ¿Cuáles son el coeficiente y el error estándar de *bs*?

iii) Obtenga los errores estándar que son robustos a la correlación de agrupamientos dentro del distrito (y también a la heterocedasticidad). ¿Qué ocurre con el estadístico *t* para *bs*?

iv) Utilizando todavía MCO combinados, omita las cuatro observaciones con $bs > .5$ y obtenga $\hat{\beta}_{bs}$ y su error estándar robusto a la correlación de agrupamiento. ¿Existe ahora suficiente evidencia para decir que hay un efecto de sustitución entre salario y prestaciones?

v) Estime la ecuación por efectos fijos, considerando un efecto de distrito común para las escuelas dentro de un distrito. De nuevo, omita las observaciones con $bs > .5$. Ahora bien, ¿cuál es su conclusión respecto al efecto de sustitución entre salario y prestaciones?

vi) Considerando las estimaciones de los incisos iv) y v), comente la importancia de permitir que las remuneraciones de los profesores varíen sistemáticamente a lo largo de los distritos por medio de un efecto fijo del distrito.

Apéndice 14A

Supuestos para los efectos fijos y aleatorios

En este apéndice se proporcionan los enunciados de los supuestos para la estimación de efectos fijos y la de efectos aleatorios. También se ofrece un análisis de las propiedades de los estimadores bajo distintos conjuntos de supuestos. La verificación de estas afirmaciones es un tanto complicada, pero puede encontrarse en Wooldridge (2002, capítulo 10).

Supuesto EF.1

Para cada i, el modelo es

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it},\ t = 1, \ldots, T,$$

donde las β_j son los parámetros a estimar y a_i es el efecto inobservable.

Supuesto EF.2

Se tiene una muestra aleatoria en la dimensión de corte transversal.

Supuesto EF.3

Cada variable explicativa cambia con el tiempo (al menos para alguna i), y no existe una relación lineal perfecta entre las variables explicativas.

Supuesto EF.4

Para cada t, el valor esperado del error idiosincrático, dadas las variables explicativas en *todos* los periodos y el efecto inobservable, es cero: $\mathrm{E}(u_{it}|\mathbf{X}_i, a_i) = 0$.

Con base en los primeros cuatro supuestos, los cuales son idénticos a los supuestos del estimador de primeras diferencias, el estimador de efectos fijos es insesgado. De nuevo, la clave es el supuesto de exogeneidad estricta, EF.4. Bajo estos mismos supuestos, el estimador de EF es consistente con un T fijo conforme $N \to \infty$.

Supuesto EF.5

$\mathrm{Var}(u_{it}|\mathbf{X}_i, a_i) = \mathrm{Var}(u_{it}) = \sigma_u^2$, para toda $t = 1, \ldots, T$.

Supuesto EF.6

Para toda $t \neq s$, los errores idiosincráticos no están correlacionados (de manera condicional en todas las variables explicativas y en a_i): $\text{Cov}(u_{it}, u_{is} | \mathbf{X}_i, a_i) = 0$.

Con base en los supuestos EF.1 a EF.6, el estimador de efectos fijos de las β_j es el mejor estimador lineal insesgado. Como el estimador de PD es lineal e insesgado, forzosamente es peor que el estimador de EF. El supuesto que hace que EF sea mejor que PD es EF.6, lo que implica que los errores idiosincráticos no se correlacionan serialmente.

Supuesto EF.7

De manera condicional en $\mathbf{X}_i$ y en a_i, los u_{it} son independientes e idénticamente distribuidos como Normal$(0, \sigma_u^2)$.

El supuesto EF.7 implica los supuestos EF.4, EF.5 y EF.6, pero es más fuerte, ya que supone una distribución normal para los errores idiosincráticos. Si se agrega EF.7, el estimador de EF se distribuye normalmente y los estadísticos t y F tienen distribuciones exactas t y F. Sin EF.7, se puede confiar en las aproximaciones asintóticas. Pero, sin hacer suposiciones especiales, estas aproximaciones requieren una N grande y un T pequeño.

Los supuestos ideales de efectos aleatorios comprenden los supuestos EF.1, EF.2, EF.4, EF.5 y EF.6. (EF.7 podría agregarse a la lista, pero es poco práctico porque se debe estimar λ). Como sólo se está restando una fracción de los promedios en el tiempo, ahora se pueden permitir variables explicativas constantes. Así, el EF.3 se remplaza con

Supuesto EA.3

No existen relaciones lineales perfectas entre las variables explicativas.

El costo de permitir regresores constantes en el tiempo es que se deben añadir supuestos acerca de cómo se relaciona el efecto inobservable, a_i, con las variables explicativas.

Supuesto EA.4

Además del supuesto EF.4, el valor esperado de a_i dadas todas las variables explicativas, es constante: $\text{E}(a_i | \mathbf{X}_i) = \beta_0$.

Este es el supuesto que rige la correlación entre el efecto inobservable y las variables explicativas, y es la distinción fundamental entre los efectos fijos y los efectos aleatorios. Como se supone que a_i no se correlaciona con ninguno de los elementos de $\mathbf{x}_{it}$, es posible incluir variables explicativas constantes en el tiempo. (Técnicamente, los datos cuasi deducidos sólo eliminan una fracción y no todo el promedio a lo largo del tiempo.) Se toma en cuenta una esperanza diferente de cero para a_i al definir el supuesto EA.4, de modo que el modelo, bajo los supuestos de efectos aleatorios, contiene un intercepto, β_0, como en la ecuación (14.7). Recuerde que, por lo general, también se incluiría un conjunto de interceptos para cada periodo tomando como año base el primer año.

Asimismo, es necesario imponer la homocedasticidad en a_i como sigue:

Supuesto EA.5

Además de EF.5, la varianza de a_i dadas todas las variables explicativas, es constante: $\mathrm{Var}(a_i|\mathbf{X}_i) = \sigma_a^2$.

Con base en los seis supuestos de efectos aleatorios (EF.1, EF.2, EA.3, EA.4, EA.5 y EF.6), el estimador de efectos aleatorios es consistente y se distribuye asintóticamente como normal a medida que N aumenta para un T fijo. En realidad, la consistencia y la normalidad asintótica son consecuencia de los primeros cuatro supuestos, pero sin los dos últimos supuestos los errores estándar y los estadísticos de prueba de EA no serían válidos. Además, bajo los seis supuestos de EA, los estimadores de EA son asintóticamente eficientes. Esto significa que, en muestras grandes, los estimadores de EA tendrán errores estándar menores que aquellos que corresponden a los estimadores combinados de MCO (cuando se utilizan los errores estándar robustos apropiados para MCO combinados). Para los coeficientes de las variables explicativas que cambian en el tiempo (las únicas estimables por EF), el estimador de EA es más eficiente que el de EF y, con frecuencia, mucho más eficiente. Pero EF no es eficiente bajo los supuestos de EA; EF está pensado para ser robusto a la correlación entre a_i y las x_{itj}. Como sucede a menudo en la econometría, existe un equilibrio entre la robustez y la eficiencia. Vea Wooldridge (2002, capítulo 10) para la verificación de las afirmaciones hechas aquí.

CAPÍTULO 15

Estimación con variables instrumentales y mínimos cuadrados en dos etapas

En este capítulo se abundará más en el problema de las **variables explicativas endógenas** en los modelos de regresión múltiple. En el capítulo 3 se derivó el sesgo en los estimadores de MCO cuando se omite una variable importante; en el capítulo 5, se mostró que las MCO suelen ser inconsistentes cuando existen **variables omitidas**. En el capítulo 9 se demostró que el sesgo de variables omitidas se puede eliminar (o al menos mitigar) cuando existe una variable proxy adecuada para una variable explicativa no observada. Por desgracia, no siempre se cuenta con tales variables.

En los dos capítulos previos se explicó cómo se utiliza la estimación de efectos fijos o de primeras diferencias con datos de panel para estimar los efectos de variables independientes que cambian con el tiempo en presencia de variables omitidas *constantes en el tiempo*. Aunque tales métodos son muy útiles, no siempre se tiene acceso a los datos de panel. Aun si fuera posible obtenerlos, de poco sirve si lo que interesa es el efecto de una variable que no cambia con el paso del tiempo: la primera diferencia o la estimación de efectos fijos elimina las variables explicativas de tiempo constante. Además, los métodos de datos de panel que se han estudiado hasta ahora no resuelven el problema de variables omitidas que cambian con el tiempo y que están correlacionadas con las explicativas.

En este capítulo se aborda el problema de la endogeneidad desde una perspectiva diferente. Se observará cómo se puede utilizar el método de variables instrumentales (VI) para resolver el problema de la endogeneidad de una o más variables explicativas. El método de mínimos cuadrados en dos etapas (MC2E) es el segundo en popularidad en econometría para estimar ecuaciones lineales, sólo después del de mínimos cuadrados ordinarios.

Se iniciará este capítulo mostrando cómo se pueden utilizar los métodos de VI para obtener estimadores consistentes en presencia de variables omitidas. Las VI también se puede emplear para resolver el problema de **errores en las variables**, al menos bajo determinados supuestos. En el siguiente capítulo se demostrará cómo estimar modelos de ecuaciones simultáneas mediante los métodos de VI.

El tratamiento que aquí se da a la estimación con variables instrumentales sigue de cerca el desarrollo de mínimos cuadrados ordinarios de la parte 1, donde se supuso que se tenía una muestra aleatoria de una población subyacente. Este es un punto de partida positivo porque, además de simplificar la notación, hace énfasis en que los supuestos importantes para la estimación de VI están expresados en términos de la población subyacente (tal como con MCO). Como se mostró en la parte 2, MCO se pueden aplicar a datos de series de tiempo, y lo mismo es cierto para los métodos de variables instrumentales. La sección 15.7 analiza algunas cuestiones especiales que surgen al aplicar los métodos de VI a los datos de series de tiempo. En la sección 15.8 se abordarán las aplicaciones para cortes transversales combinados y datos de panel.

15.1 Justificación: variables omitidas en un modelo de regresión simple

Cuando existe la posibilidad de sesgo por variable omitida (o heterogeneidad inobservable), se han analizado hasta ahora tres opciones: 1) se puede ignorar el problema y sufrir las consecuencias de estimadores sesgados e inconsistentes; 2) se puede intentar encontrar y utilizar una variable proxy adecuada para la variable inobservable; o 3) se puede suponer que la variable omitida no cambia con el tiempo y utilizar los métodos de efectos fijos o de primeras diferencias de los capítulos 13 y 14. La primera respuesta puede ser satisfactoria si las estimaciones están asociadas a la dirección del sesgo para los parámetros clave. Por ejemplo, si se puede decir que el estimador de un parámetro positivo, por ejemplo, el efecto de la capacitación laboral en los salarios, está sesgado hacia cero y se encuentra una estimación positiva estadísticamente significativa, aún se ha aprendido algo: la capacitación laboral tiene un efecto positivo en los salarios y es probable que se haya subestimado el efecto. Por desgracia, el caso opuesto suele ocurrir, en el que las estimaciones son demasiado pequeñas, lo que hace muy difícil llegar a cualquier conclusión útil.

La solución de la variable proxy que se analizó en la sección 9.2 también puede producir resultados satisfactorios, pero no siempre es posible encontrar una proxy buena. Este método intenta resolver el problema de la variable omitida al remplazar la inobservable con una variable proxy.

Otro método deja a la variable inobservable en el término de error, pero en lugar de estimar el modelo mediante MCO, utiliza un método de estimación que reconoce la presencia de la variable omitida. Esto es lo que hace el método de variables instrumentales.

Como ejemplo, considere el problema de la capacidad inobservable en una ecuación de salario para adultos trabajadores. Un modelo simple es

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e,$$

donde e es el término de error. En el capítulo 9 se mostró cómo, bajo ciertos supuestos, una variable proxy como IQ puede sustituir a la capacidad y, por tanto, se tendrá un estimador consistente de β_1 de la regresión de

$$\log(wage) \text{ sobre } educ, IQ.$$

Sin embargo, suponga que no se cuenta con una variable proxy (o que no dispone de las propiedades necesarias para producir un estimador consistente de β_1). Entonces, se escribe *abil* (capacidad) en el término de error y se tendrá sólo el modelo de regresión simple

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{15.1}$$

donde u contiene a *abil*. Por supuesto, si la ecuación (15.1) se estimó por MCO, se obtendrá un estimador sesgado e inconsistente de β_1 si *educ* y *abil* están correlacionadas.

Resulta que aún se puede utilizar la ecuación (15.1) como base para la estimación, siempre y cuando se pueda encontrar una variable instrumental para *educ*. Para describir este método, el modelo de regresión simple se escribe como

$$y = \beta_0 + \beta_1 x + u, \tag{15.2}$$

donde se supone que x y u están correlacionadas:

$$\mathrm{Cov}(x,u) \neq 0. \tag{15.3}$$

El método de variables instrumentales funciona estén o no x y u correlacionadas, pero por las razones que más tarde se analizarán, MCO se deben utilizar si x no está correlacionada con u.

Con el fin de obtener estimadores consistentes de β_0 y β_1 cuando x y u están correlacionadas, se necesitará alguna información adicional, la cual la ofrece una nueva variable que satisface ciertas propiedades. Suponga que se tiene una variable observable z que satisface los siguientes dos supuestos: 1) z no está correlacionada con u, es decir,

$$\mathrm{Cov}(z,u) = 0; \tag{15.4}$$

2) z está correlacionada con x, es decir

$$\mathrm{Cov}(z,x) \neq 0. \tag{15.5}$$

Entonces, z se denomina **variable instrumental** para x, o algunas veces simplemente un **instrumento** para x.

El requisito de que el instrumento z satisfaga (15.4) se resume diciendo que "z es exógena en la ecuación (15.2)", y por tanto, (15.4) se denomina **exogeneidad del instrumento**. En el contexto de las variables omitidas, la exogeneidad del instrumento significa que z no debe tener ningún efecto parcial sobre y (después de que x y las variables omitidas se han controlado), y z no debe estar correlacionada con las variables omitidas. La ecuación (15.5) significa que z debe estar relacionada, positiva o negativamente, con la variable explicativa endógena x. Esta condición se conoce como **relevancia del instrumento** (como en "z es relevante para explicar la variación en x").

Existe una importante diferencia entre los dos requisitos de una variable instrumental. Dado que (15.4) implica la covarianza entre z y el error inobservable u, generalmente no se espera demostrar este supuesto: en la mayoría de los casos, se debe mantener $\mathrm{Cov}(z,u) = 0$ recurriendo al comportamiento económico o introspección. (En casos extraordinarios, se puede tener una variable proxy observable para algún factor contenido en u, en cuyo caso se podrá revisar si z y la variable proxy no están correlacionadas. Por supuesto, si se tiene una buena proxy para un elemento importante de u, se debe agregar la proxy como una variable explicativa y estimar la ecuación expandida mediante mínimos cuadrados ordinarios. Vea la sección 9.2.)

Por el contrario, sí se puede probar la condición de que z esté correlacionada con x (en la población), dada una muestra aleatoria de la población. La forma más fácil de hacerlo es estimar la regresión simple entre x y z. En la población, se tiene

$$x = \pi_0 + \pi_1 z + v. \tag{15.6}$$

Entonces, debido a que $\pi_1 = \mathrm{Cov}(z,x)/\mathrm{Var}(z)$, el supuesto (15.5) se mantiene si, y sólo si, $\pi_1 \neq 0$. Por tanto, se debe *rechazar* la hipótesis nula

$$\mathrm{H}_0\!: \pi_1 = 0 \tag{15.7}$$

contra la alternativa de dos colas H_0: $\pi_1 \neq 0$, a un nivel de significancia suficientemente pequeño (por ejemplo, 5 o 1%). Si este es el caso, se puede tener la confianza de que (15.5) es válida.

Para la ecuación de log(*wage*) en (15.1), una variable instrumental z para *educ* debe 1) no estar correlacionada con la capacidad (ni ningún otro factor inobservable que afecte al salario) y 2) estar correlacionada con la educación. Algo como el último dígito del número de seguro social de un individuo casi seguramente satisfará este primer requisito: no está correlacionada con la capacidad debido a que está determinado de forma aleatoria. No obstante, esta variable no está correlacionada con la educación, lo cual la convierte en una variable instrumental pobre para *educ*.

Por la razón contraria, lo que se conoce como *variable proxy* para la variable omitida hace de VI una mala aproximación. Por ejemplo, en el caso de log(*wage*) con capacidad omitida, una variable proxy para *abil* debe estar altamente correlacionada con *abil* tanto como sea posible. Una variable instrumental no debe estar *correlacionada* con *abil*. Por tanto, si bien *IQ* es un buen candidato como variable proxy de *abil*, no es una buena variable instrumental para *educ*.

Lo que resulta menos claro es si los demás posibles candidatos para las variables instrumentales pueden satisfacer los requisitos de exogeneidad en (15.4). En las ecuaciones de salario, los economistas laborales han utilizado variables de antecedentes familiares como VI para la educación. Por ejemplo, la educación de la madre (*motheduc*) está positivamente correlacionada con la educación de sus hijos, como se puede apreciar al recabar una muestra de datos sobre las personas trabajadoras y realizar una regresión simple de *educ* sobre *motheduc*. Por tanto, *motheduc* satisface la ecuación (15.5). El problema es que la educación de la madre podría estar correlacionada con la capacidad del niño (a través de la capacidad de la madre y quizá, de la calidad de los cuidados a edades tempranas), en cuyo caso (15.4) falla.

Otra elección de VI para *educ* en (15.1) es el número de hermanos que se tiene durante el crecimiento (*sibs*). Por lo general, tener más hermanos está asociado con niveles promedio inferiores de educación. Por tanto, si el número de hermanos no está correlacionado con la capacidad, puede actuar como una variable instrumental de *educ*.

Como segundo ejemplo, considere el problema de estimar el efecto causal de no asistir a clases sobre la calificación del examen final. En un marco de regresión simple, se tiene

$$score = \beta_0 + \beta_1 skipped + u, \tag{15.8}$$

donde *score* es la puntuación del examen final y *skipped* es el número total de faltas a clase durante el semestre. Por supuesto, sería motivo de preocupación que *skipped* estuviera correlacionada con otros factores en u: estudiantes más capaces y altamente motivados podrían tener menos ausencias. Por tanto, una regresión simple de *score* sobre *skipped* quizá no produzca una buena estimación del efecto causal de faltar a clases.

¿Cuál podría ser una buena VI para *skipped*? Se necesita algo que no tenga efecto directo en *score* y no se correlacione con la capacidad y motivación del estudiante. Asimismo, la VI debe estar correlacionada con *skipped*. Una opción es utilizar la distancia entre los dormitorios y el campus. Algunos estudiantes en una universidad grande tendrán que desplazarse todos los días hasta sus escuelas, lo cual podría aumentar la probabilidad de faltar a clases (debido al mal tiempo, a quedarse dormido, etc.). Por tanto, *skipped* puede estar correlacionada de forma positiva con *distance* (distancia); esto se puede comprobar al realizar una regresión de *skipped* sobre *distance* y al efectuar una prueba t, como se describió antes.

¿No está correlacionada *distance* con u? En el modelo de regresión simple (15.8) algunos factores en u pueden estar correlacionados con *distance*. Por ejemplo, los estudiantes de las familias de bajos ingresos pueden vivir fuera del campus; si el ingreso afecta el desempeño del estudiante, esto podría ocasionar que *distance* estuviera correlacionada con u. La sección 15.2 muestra cómo utilizar VI en el contexto de la regresión múltiple, de manera que todos los demás factores que afectan *score* se puedan incluir directamente en el modelo. Entonces, *distance* podría ser una buena VI para *skipped*. Un método de VI podría no ser necesario si existe una buena proxy para la capacidad del estudiante, como por ejemplo, un promedio acumulativo de calificaciones antes del semestre.

Ahora se mostrará que la disponibilidad de una variable instrumental se puede emplear para estimar de manera consistente los parámetros en la ecuación (15.2). En particular, se mostrará que los supuestos (15.4) y (15.5) sirven para *identificar* el parámetro β_1. La **identificación** de un parámetro en este contexto significa que se puede escribir β_1 en términos de los momentos poblacionales que se pueden estimar mediante una muestra de datos. Para escribir β_1 en términos de las covarianzas poblacionales, se utiliza la ecuación (15.2): la covarianza entre z y y es

$$\mathrm{Cov}(z,y) = \beta_1 \mathrm{Cov}(z,x) + \mathrm{Cov}(z,u).$$

Ahora, de acuerdo con el supuesto (15.4), $\mathrm{Cov}(z,u) = 0$, y bajo el supuesto (15.5), $\mathrm{Cov}(z,x) \neq 0$. Por tanto, se puede resolver para β_1 como

$$\beta_1 = \frac{\mathrm{Cov}(z,y)}{\mathrm{Cov}(z,x)}. \tag{15.9}$$

[Observe cómo estos fáciles cálculos de álgebra son inútiles si z y x no están correlacionadas, es decir, si $\mathrm{Cov}(z, x) = 0$.] La ecuación (15.9) muestra que β_1 es la covarianza poblacional entre z y y dividida entre la covarianza poblacional entre z y x, lo cual muestra que se puede identificar β_1. Dada una muestra aleatoria se estiman las cantidades poblacionales mediante los análogos muestrales. Después de cancelar los tamaños muestrales en el numerador y el denominador, se obtiene el **estimador** de **variables instrumentales (VI)** de β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}. \tag{15.10}$$

Dada una muestra de datos en x, y y z, es simple obtener el estimador de VI en (15.10). El estimador de VI de β_0 es simplemente $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, lo cual se asemeja mucho al estimador del intercepto de MCO, salvo que el estimador de la pendiente, $\hat{\beta}_1$, ahora es el estimador de VI.

No es fortuito que cuando $z = x$ se obtenga el estimador de MCO de β_1. En otras palabras, cuando x es exógena, se puede utilizar como su propia VI, y el estimador de VI entonces es idéntico al estimador de MCO.

Una aplicación simple de la ley de los grandes números muestra que el estimador de VI es consistente para β_1: $\mathrm{plim}(\hat{\beta}_1) = \beta_1$, siempre y cuando se satisfagan los supuestos (15.4) y (15.5). Si cualquier supuesto falla, los estimadores de VI no serán consistentes (se abundará en este punto más adelante). Una característica del estimador de VI es que cuando x y u están correlacionadas, de manera que la estimación de las variables instrumentales en realidad es necesaria, el estimador nunca es insesgado. Esto significa que, en pequeñas muestras, el estimador de VI puede tener un sesgo sustancial, lo cual es una razón del porqué se prefieren las grandes muestras.

Inferencia estadística con el estimador de VI

Dada la estructura similar de los estimadores de VI y MCO, no es de sorprender que el estimador de VI tenga una distribución normal aproximada en tamaños muestrales grandes. Para realizar una inferencia sobre β_1, se necesita un error estándar que se pueda emplear para calcular los estadísticos t y los intervalos de confianza. El método común es imponer un supuesto de homocedasticidad, tal como en el caso de MCO. Ahora, el supuesto de homocedasticidad se expresa de manera condicional en la variable instrumental, z, no en la variable explicativa endógena, x. Junto con los supuestos previos sobre u, x y z, se agrega

$$\mathrm{E}(u^2|z) = \sigma^2 = \mathrm{Var}(u). \tag{15.11}$$

Se puede mostrar que, de acuerdo con (15.4), (15.5) y (15.11), la varianza asintótica de $\hat{\beta}_1$ es

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}, \tag{15.12}$$

donde σ_x^2 es la varianza poblacional de x, σ^2 es la varianza poblacional de u, y $\rho_{x,z}^2$ es el cuadrado de la correlación poblacional entre x y z. Esto mide qué tanta correlación hay entre x y z en la población. Como con el estimador de MCO, la varianza asintótica del estimador de VI disminuye a cero a la tasa $1/n$, donde n es el tamaño de muestra.

La ecuación (15.12) es interesante por dos razones. Primero, ofrece una forma de obtener un error estándar para el estimador de VI. Todas las cantidades en (15.12) pueden estimarse de forma consistente dada una muestra aleatoria. Para estimar σ_x^2, simplemente se calcula la varianza muestral de x_i; para estimar $\rho_{x,z}^2$, se puede realizar la regresión de x_i sobre z_i para obtener la R cuadrada, o sea, $R_{x,z}^2$. Por último, para estimar σ^2, se pueden utilizar los residuales de VI,

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n,$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son las estimaciones de VI. Un estimador consistente de σ^2 se parece mucho al estimador de σ^2 de una regresión simple de MCO:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2,$$

donde es adecuado aplicar la corrección de los grados de libertad (a pesar de su efecto limitado a medida que el tamaño de la muestra crece).

El error estándar asintótico de $\hat{\beta}_1$ es la raíz cuadrada de la varianza asintótica estimada, la última de las cuales está dada por

$$\frac{\hat{\sigma}^2}{\text{STC}_x \cdot R_{x,z}^2}, \tag{15.13}$$

donde STC_x es la suma total de cuadrados de x_i. [Recuerde que la varianza muestral de x_i es STC_x/n y de este modo se cancelan los tamaños muestrales para resultar (15.13).] El error estándar resultante se puede utilizar para construir estadísticos t para probar las hipótesis que involucran β_1 o para intervalos de confianza para β_1. Por su parte, $\hat{\beta}_0$ también cuenta con un error estándar que no se presenta aquí. Cualquier paquete moderno de econometría puede calcular el error estándar después de cualquier estimación con VI.

Una segunda razón por la que (15.12) es interesante es porque permite comparar las varianzas asintóticas de los estimadores de VI y MCO (cuando x y u no se correlacionan). Mediante los supuestos de Gauss-Markov, la varianza del estimador de MCO es σ^2/STC_x, mientras que la fórmula comparable para el estimador VI es $\sigma^2/(\text{STC}_x \cdot R_{x,z}^2)$; difieren sólo en que $R_{x,z}^2$ aparece en el denominador de la varianza de VI. Debido a que una R cuadrada siempre es menor que uno, la varianza de VI siempre es mayor que la varianza de MCO (cuando MCO es válido). Si $R_{x,z}^2$ es pequeño, entonces la varianza de VI puede ser mucho mayor que la varianza de MCO. Recuerde que $R_{x,z}^2$ mide la fortaleza de la relación lineal entre x y z en la muestra. Si x y z sólo tienen una correlación débil, $R_{x,z}^2$ puede ser pequeña y esto se traduce en una varianza de muestreo muy grande para el estimador de VI. Cuanto mayor sea la correlación entre z y x, más cerca estará $R_{x,z}^2$ de uno, y menor será la varianza del estimador de VI. En el caso de que $z = x$, $R_{x,z}^2 = 1$, y se obtiene la varianza de MCO, como es de esperar.

El análisis previo enfatiza un importante costo de realizar la estimación de VI cuando x y u no se correlacionan: la varianza asintótica del estimador de VI es siempre mayor, y algunas veces mucho mayor, que la varianza asintótica del estimador de MCO.

Ejemplo 15.1

[Estimación del rendimiento de la educación para mujeres casadas]

Se utilizan los datos sobre mujeres trabajadoras casadas en MROZ.RAW para estimar el rendimiento de la educación en el modelo de regresión simple

$$\log(wage) = \beta_0 + \beta_1 educ + u. \tag{15.14}$$

Para fines comparativos, se obtienen primero las estimaciones de MCO:

$$\widehat{\log(wage)} = -.185 + .109\ educ$$
$$(.185)\quad (.014) \tag{15.15}$$
$$n = 428,\ R^2 = .118.$$

La estimación para β_1 implica un rendimiento de casi 11% para un año de educación adicional.

Luego, se utiliza la educación del padre (*fatheduc*) como variable instrumental para *educ*. Se tiene que sostener que *fatheduc* no está correlacionada con *u*. El segundo requisito es que *educ* y *fatheduc* se correlacionen. Se puede verificar esto muy fácilmente mediante una regresión simple de *educ* sobre *fatheduc* (usando sólo a las mujeres trabajadoras en la muestra):

$$\widehat{educ} = 10.24 + .269\ fatheduc$$
$$(.28)\quad (.029) \tag{15.16}$$
$$n = 428,\ R^2 = .173.$$

El estadístico *t* de *fatheduc* es 9.28, lo cual indica que *educ* y *fatheduc* tienen una correlación positiva significativa en términos estadísticos. (De hecho, *fatheduc* explica aproximadamente 17% de la variación en *educ* en la muestra.) Si se usa *fatheduc* como una VI para *educ* se tiene

$$\widehat{\log(wage)} = .441 + .059\ educ$$
$$(.446)\quad (.035) \tag{15.17}$$
$$n = 428,\ R^2 = .093.$$

La estimación de VI de los rendimientos de la educación es 5.9%, lo cual es apenas mayor que la mitad del estimador de MCO. Esto *sugiere* que el estimador de MCO es demasiado alto y es consistente con el sesgo provocado por la capacidad innata omitida. Pero debe recordarse que éstos son estimadores de sólo una muestra: nunca se puede saber si .109 está por encima del verdadero rendimiento de la educación, o si .059 está más cerca del verdadero rendimiento de la educación. Además, el error estándar del estimador de VI es dos y media veces el tamaño del error estándar de MCO (lo que era de esperarse por las razones que se dieron antes). El intervalo de confianza a 95% para β_1 mediante MCO es mucho más ajustado que el de VI; en realidad, el intervalo de confianza de VI contiene al de MCO. Por tanto, aunque las diferencias entre (15.15) y (15.17) son grandes en términos prácticos, no se puede decir si la diferencia es *estadísticamente* significativa. Se mostrará cómo probar esta cuestión en la sección 15.5.

En el ejemplo anterior, el rendimiento estimado de la educación con VI fue menor que si se hubiera utilizado MCO, lo cual corresponde a nuestras expectativas. Pero esto no necesariamente es el caso, como lo demuestra el siguiente ejemplo.

Ejemplo 15.2

[Estimación del rendimiento de la educación en los hombres]

Ahora se utilizará WAGE2.RAW para estimar el rendimiento de la educación para los hombres. Se utiliza la variable *sibs* (número de hermanos) como instrumento para *educ*. Éstas están correlacionadas de forma negativa, como se puede constatar a partir de la regresión simple:

$$\widehat{educ} = \underset{(.11)}{14.14} - \underset{(.030)}{.228}\, sibs$$
$$n = 935, R^2 = .057.$$

Esta ecuación implica que cada hermano se asocia, en promedio, con aproximadamente .23 años menos de educación. Si se supone que *sibs* no está correlacionada con el término de error en (15.14), entonces el estimador de VI es consistente. Al estimar la ecuación (15.14) mediante *sibs* como una VI para *educ* se obtiene

$$\widehat{\log(wage)} = \underset{(.36)}{5.13} + \underset{(.026)}{.122}\, educ$$
$$n = 935.$$

(La *R* cuadrada resulta ser negativa, así que no se reporta. A continuación se presenta un análisis de la *R* cuadrada en el contexto de la estimación de VI.) Para comparar, la estimación de MCO de β_1 es .059 con un error estándar de .006. A diferencia del ejemplo anterior, la estimación de VI ahora es mucho mayor que la estimación de MCO. Si bien no se sabe si la diferencia es estadísticamente significativa, esto no concuerda con el sesgo de MCO originado por la capacidad innata omitida. Podría ser que *sibs* también esté correlacionada con la capacidad: más hermanos significan, en promedio, menos atención paterna, lo cual redundaría en una menor capacidad. Otra interpretación es que el estimador de MCO sea sesgado hacia cero debido a un error de medición en *educ*. Esto no es del todo convincente porque, como se analizó en la sección 9.3, no es probable que *educ* satisfaga el modelo de errores clásicos en las variables.

En los ejemplos anteriores, la variable explicativa endógena (*educ*) y las variables instrumentales (*fatheduc*, *sibs*) tenían un significado cuantitativo. Pero nada impide que la variable explicativa o la VI sean variables binarias. Angrist y Krueger (1991), en sus más sencillos análisis, obtuvieron una variable instrumental binaria para *educ*, usando los datos del censo para los hombres estadounidenses. Sea *frstqrt* igual a uno si el hombre nació en el primer trimestre del año, y cero de lo contrario. Parece que el término de error en (15.14) y, en particular, la capacidad, debería no estar relacionado con el trimestre de nacimiento. Pero *frstqrt* también debe estar correlacionada con *educ*. Resulta que los años *sí difieren* sistemáticamente en la población con base en el trimestre de nacimiento. Angrist y Krueger argumentaron de manera convincente que esto se debe a las leyes de asistencia escolar obligatoria vigentes en todos los estados. En resumen, los estudiantes nacidos a principio de año, por lo general, comienzan la escuela a una edad mayor. Por tanto, cumplen la edad obligatoria para asistir a la escuela (16 en la mayoría de los estados) con menor educación que los estudiantes que comenzaron la escuela a edades más tempranas. Para los estudiantes que sí terminan el bachillerato, Angrist y Krueger verificaron que no existía relación entre los años de educación y el trimestre de nacimiento.

Debido a que los años de educación varían ligeramente dentro de un mismo trimestre de nacimiento, lo cual significa que $R^2_{x,z}$ en (15.13) es muy pequeña, Angrist y Krueger necesitaron un tamaño muestral muy grande para llegar a una estimación de VI razonablemente precisa. Utilizando datos de 247,199 hombres nacidos entre 1920 y 1929, la estimación de MCO de los rendimientos de la educación fue de .0801 (con un error estándar de .0004), y la estimación de

VI fue de .0715 (.0219); estos datos se muestran en la tabla III del trabajo de Angrist y Krueger. Observe el tamaño del estadístico t de la estimación de MCO (aproximadamente 200), mientras que el de la estimación de VI es sólo de 3.26. Por tanto, el estimador de VI es estadísticamente diferente de cero, pero su intervalo de confianza es mucho más amplio que el basado en la estimación de MCO.

Un hallazgo interesante de Angrist y Krueger es que la estimación de VI no difiere mucho de aquella de MCO. En realidad, con hombres nacidos en la siguiente década, la estimación de VI es un poco mayor que la de MCO. Esto se podría interpretar como una demostración de que no hay sesgo por la capacidad omitida cuando las ecuaciones del salario se estiman mediante MCO. No obstante, el trabajo de Angrist y Krueger ha sido muy criticado con fundamentos económicos. Como lo analizaron Bound, Jaeger y Baker (1995), no es evidente que la temporada de nacimiento no se relacione con los factores inobservables que afectan al salario. Como se explicará en la siguiente subsección, incluso una pequeña correlación entre z y u podría ocasionar serios problemas para el estimador de VI.

Para el análisis de política, la variable explicativa endógena suele ser una variable binaria. Por ejemplo, Angrist (1990) estudió el efecto que tenía ser un veterano de la Guerra de Vietnam sobre los ingresos de por vida. Un modelo simple es

$$\log(earns) = \beta_0 + \beta_1 veteran + u, \tag{15.18}$$

donde *veteran* (veterano) es una variable binaria. El problema con estimar esta ecuación mediante MCO es que puede haber un problema de *autoselección*, como se mencionó en el capítulo 7: quizá las personas que obtienen las mayores ganancias de la milicia son quienes opten por enlistarse, o esta decisión está correlacionada con otras características que afectan al ingreso. Esto ocasionará que *veteran* y u se correlacionen.

Angrist señaló que el sorteo de reclutamiento a Vietnam ofrecía un **experimento natural** (vea también el capítulo 13) que creaba una variable instrumental para *veteran*. Los jóvenes recibían números de lotería que determinaban si serían llamados a servicio en Vietnam. Debido a que los números que se distribuían se asignaron (finalmente) de forma aleatoria, parece posible que el número de la lotería para el reclutamiento no está correlacionado con el término de error u. Pero aquellas personas con un número suficientemente bajo tenían que servir en Vietnam, de manera que la probabilidad de ser veterano estaba correlacionada con el número de lotería. Si ambas aseveraciones fueran ciertas, el número de la lotería para el reclutamiento es un buen candidato de VI para *veteran*.

Pregunta 15.1

Si algunos de los hombres a los que se les asignaron números bajos en la lotería para reclutamiento, obtuvieran escolaridad adicional para reducir la probabilidad de ser reclutados, ¿el número de la lotería sería un buen instrumento para *veteran* en (15.18)?

También es posible tener una variable explicativa endógena binaria y una variable instrumental binaria. Vea el problema 15.1 para un ejemplo.

Propiedades de VI con una variable instrumental deficiente

Se ha visto que, aunque VI es consistente cuando z y u no están correlacionadas y cuando z y x tienen cualquier correlación positiva o negativa, las estimaciones de VI pueden tener errores estándar mayores, en especial si z y x tienen sólo una correlación débil. La correlación débil entre z y x puede tener consecuencias aun más serias: el estimador de VI puede tener un sesgo asintótico grande aun si z y u estuvieran moderadamente correlacionadas.

Esto se puede observar al estudiar el límite de probabilidad del estimador de VI cuando z y u tienen una posible correlación. Siendo $\hat{\beta}_{1,\mathrm{VI}}$ el estimador de VI, entonces se puede escribir

$$\operatorname{plim} \hat{\beta}_{1,\mathrm{VI}} = \beta_1 + \frac{\mathrm{Corr}(z,u)}{\mathrm{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}, \tag{15.19}$$

donde σ_u y σ_x son las desviaciones estándar de u y x en la población, respectivamente. La parte interesante de esta ecuación implica a los términos de correlación. Muestra que, aunque la Corr(z,u) sea pequeña, la inconsistencia en el estimador de VI puede ser muy grande si la Corr(z,x) también es pequeña. Por tanto, aun si el enfoque sólo estuviera puesto en la consistencia, no es necesariamente mejor utilizar VI que MCO si la correlación entre z y u es menor que entre x y u. Mediante el hecho de que Corr(x,u) = Cov(x,u)/($\sigma_x\sigma_u$) junto con la ecuación (5.3), se puede escribir el plim del estimador de MCO, llámese $\hat{\beta}_{1,\text{MCO}}$, como

$$\text{plim } \hat{\beta}_{1,\text{MCO}} = \beta_1 + \text{Corr}(x,u) \cdot \frac{\sigma_u}{\sigma_x}. \tag{15.20}$$

La comparación de estas fórmulas muestra que es posible que las direcciones de los sesgos asintóticos sean diferentes para VI y MCO. Por ejemplo, suponga que Corr(x,u) > 0, que Corr(z,x) > 0, y que Corr(z,u) < 0. Entonces el estimador de VI tiene un sesgo hacia abajo, mientras que el estimador de MCO tiene un sesgo hacia arriba (asintóticamente). En la práctica, esta situación quizá sea poco común. Lo más problemático es cuando la dirección del sesgo es la misma y la correlación entre z y x es pequeña. Para fines de concreción, suponga que x y z están positivamente correlacionadas con u y que Corr(z,x) > 0. Entonces, el sesgo asintótico en el estimador de VI es menor que el de MCO sólo si Corr(z,u)/Corr(z,x) $<$ Corr(x,u). Si Corr(z,x) es pequeña, entonces una correlación en apariencia pequeña entre z y u puede aumentarse para hacer que VI sea peor que MCO, aun si la atención se fija sólo en el sesgo. Por ejemplo, si Corr(z,x) = .2, la Corr(z,u) debe ser menor que una quinta parte de la Corr(x,u) para que VI tenga un sesgo asintótico menor que MCO. En varias aplicaciones, la correlación entre el instrumento y x es menor que .2. Por desgracia, debido a que rara vez se tiene idea de las magnitudes relativas de Corr(z,u) y Corr(x,u), nunca se tiene la certeza de qué estimador tiene el sesgo asintótico mayor [a menos, desde luego, que se suponga que la Corr(z,u) = 0].

En el ejemplo de Angrist y Krueger (1991) que se mencionó antes, donde x son los años de escolaridad y z es una variable binaria que indica el trimestre de nacimiento, la correlación entre z y x es muy pequeña. Bound, Jaeger y Baker (1995) analizaron las razones de por qué el trimestre de nacimiento y u podrían tener algún tipo de correlación. De la ecuación (15.19) se aprecia que esto puede producir un sesgo sustancial en el estimador de VI.

Cuando z y x en absoluto están correlacionadas, las cosas resultan especialmente mal, sin importar que z se correlacione o no con u. El siguiente ejemplo ilustra por qué siempre se debe verificar que la variable explicativa endógena esté correlacionada con la candidata para ser VI.

Ejemplo 15.3

[Estimación del efecto del tabaquismo sobre el peso al nacer]

En el capítulo 6 se estimó el efecto del tabaquismo sobre el peso al nacer (*bwght*). Sin otras variables explicativas, el modelo es

$$\log(bwght) = \beta_0 + \beta_1 packs + u, \tag{15.21}$$

donde *packs* es el número de cajetillas que una madre fuma al día. Tal vez nos preocupe que *packs* se correlacione con otros factores de salud o con la disponibilidad de un buen cuidado prenatal, de manera que *packs* y u quizá sí se correlacionen. Una variable instrumental posible para *packs* es el precio promedio de cigarrillos en el estado de residencia, *cigprice*. Se supondrá que *cigprice* y u no están correlacionadas (aun cuando el apoyo del estado al cuidado de la salud pudiera estar correlacionado con los impuestos al tabaco).

Si los cigarrillos suelen ser un bien de consumo normal, la teoría económica básica sugiere que *packs* y *cigprice* están correlacionados de forma negativa, de manera que *cigprice* se puede utilizar como una VI para *packs*. Para comprobar esto, se realiza una regresión de *packs* sobre *cigprice*, con los datos en BWGHT.RAW:

$$\begin{aligned}\widehat{packs} &= \underset{(.103)}{.067} + \underset{(.0008)}{.0003}\, cigprice \\ n &= 1{,}388,\ R^2 = .0000,\ \bar{R}^2 = -.0006.\end{aligned}$$

Esto indica que no existe relación entre fumar durante el embarazo y el precio de los cigarrillos, lo cual quizá no resulte tan sorprendente dada la naturaleza adictiva del tabaquismo.

Debido a que *packs* y *cigprice* no están correlacionados, no se debe utilizar *cigprice* como una VI para *packs* en (15.21). Pero, ¿qué sucede si se emplea? Los resultados de VI serían

$$\begin{aligned}\widehat{\log(bwght)} &= \underset{(.91)}{4.45} + \underset{(8.70)}{2.99}\, packs \\ n &= 1{,}388\end{aligned}$$

(la R cuadrada obtenida es negativa). El coeficiente sobre *packs* es enorme y de un signo inesperado. El error estándar es también muy grande, así que *packs* no es significativo. Pero las estimaciones carecen de importancia dado que *cigprice* no cumple con uno de los requisitos de una VI que siempre se puede poner a prueba: el supuesto (15.5).

El ejemplo anterior muestra que la estimación de VI puede producir resultados extraños cuando la condición de relevancia del instrumento, $\text{Corr}(z,x) \neq 0$, falla. De un interés práctico mayor resulta el llamado problema de los **instrumentos débiles**, que se define, a grandes rasgos, como el problema de correlación "baja" (pero diferente de cero) entre z y x. En una aplicación determinada, es difícil definir qué tan bajo es demasiado bajo, pero las recientes investigaciones teóricas, complementadas por estudios de simulación, han aclarado considerablemente la cuestión. Staiger y Stock (1997) formalizaron el problema de los instrumentos débiles al crear un modelo de la correlación entre z y x en función del tamaño de muestra; en particular, se supone que la correlación se reduce a cero a una tasa de $1/\sqrt{n}$. No es de sorprender que la distribución asintótica del estimador de variables instrumentales sea diferente comparada con las distribuciones asintóticas comunes, donde se supone que la correlación es fija y diferente de cero. Una de las aplicaciones del trabajo de Stock-Staiger es que la inferencia estadística usual, basada en estadísticos t y la distribución normal estándar, puede ser seriamente engañosa. [Vea Imbens y Wooldridge (2007) para un análisis más profundo.]

Cálculo de la *R* cuadrada después de la estimación de VI

La mayoría de los paquetes de regresión calcula la R cuadrada después de la estimación de VI, mediante la fórmula estándar $R^2 = 1 - \text{SRC/STC}$, donde SRC es la suma de los *residuales cuadrados de VI* y STC es la suma total de los cuadrados de y. A diferencia del caso de MCO, la R cuadrada en una estimación de VI puede ser negativa debido a que la SRC para VI en realidad puede ser mayor que la STC. Aunque no se pierde nada con reportar la R cuadrada para la estimación de VI, tampoco es útil. Cuando x y u están correlacionadas, no se puede descomponer la varianza de y en $\beta_1^2\text{Var}(x) + \text{Var}(u)$, y por tanto R cuadrada no tiene una interpretación natural.

Además, como se analizará en la sección 15.3, estas R cuadradas *no pueden* utilizarse en la forma acostumbrada para calcular las pruebas F de restricciones conjuntas.

Si la meta fuera producir la R cuadrada mayor, siempre se utilizaría MCO. Los métodos de VI tienen la intención de ofrecer mejores estimaciones del efecto *ceteris paribus* de x sobre y cuando x y u están correlacionadas; la bondad de ajuste no tiene importancia aquí. Una R cuadrada alta resultante de MCO es de poca importancia si no se estima β_1 de manera consistente

15.2 Estimación de VI del modelo de regresión múltiple

El estimador de VI del modelo de regresión múltiple se puede extender con facilidad al caso de regresión múltiple. Se comenzará con el caso donde sólo una de las variables explicativas está correlacionada con el error. De hecho, considere un modelo lineal estándar con dos variables explicativas:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1. \tag{15.22}$$

Esto recibe el nombre de **ecuación estructural** para enfatizar el hecho de que lo que interesa son las β_j, lo cual significa que se supone que la ecuación mide una relación causal. Aquí se utiliza una nueva notación para distinguir las variables endógenas de las **exógenas**. La variable dependiente y_1 es claramente endógena, pues está correlacionada con u_1. Las variables y_2 y z_1 son las variables explicativas, y u_1 es el error. Como es lo usual, se supone que el valor esperado de u_1 es cero: $E(u_1) = 0$. Se utiliza z_1 para indicar que esta variable es exógena en (15.22) (z_1 no está correlacionada con u_1). Se usa y_2 para indicar que se sospecha que esta variable está correlacionada con u_1. No se especifica por qué y_2 y u_1 están correlacionadas, pero por ahora es mejor pensar en u_1 como si contuviera una variable omitida que está correlacionada con y_2. La notación en la ecuación (15.22) se origina de los modelos de ecuaciones simultáneas (que se analizan en el capítulo 16), pero se utiliza en forma más general para distinguir con más facilidad las variables explicativas endógenas de las exógenas en un modelo de regresión múltiple.

Un ejemplo de (15.22) es

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1, \tag{15.23}$$

donde $y_1 = \log(wage)$, $y_2 = educ$ y $z_1 = exper$. En otras palabras, se supone que *exper* es exógena en (15.23), pero se permite que (por las razones acostumbradas) *educ* se correlacione con u_1.

Se sabe que si (15.22) se estima por MCO, todos los estimadores serán sesgados e inconsistentes. Por tanto, se sigue la estrategia sugerida en la sección previa y se busca una variable instrumental para y_2. Dado que se supone que z_1 no está correlacionada con u_1, ¿se podrá emplear z_1 como instrumento para y_2, si se supone que y_2 y z_1 están correlacionadas? La respuesta es no. Dado que z_1 por sí misma aparece como variable explicativa en (15.22), no sirve como variable instrumental para y_2. Se necesita otra variable exógena, que se denominará z_2, que *no* aparece en (15.22). Por tanto, los supuestos clave son que z_1 y z_2 no están correlacionadas con u_1; también se supone que u_1 tiene un valor esperado de cero, lo que sucede, sin pérdida de generalidad, cuando la ecuación contiene un intercepto:

$$E(u_1) = 0, \text{Cov}(z_1,u_1) = 0 \text{ y } \text{Cov}(z_2,u_1) = 0. \tag{15.24}$$

Dado el supuesto de que la media es igual a cero, los últimos dos supuestos son equivalente a $E(z_1u_1) = E(z_2u_1) = 0$, de manera que el enfoque del método de momentos sugiere que los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ se pueden obtener al resolver las contrapartes muestrales de (15.24):

$$\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$
$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \qquad \textbf{15.25}$$
$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0.$$

Este es un conjunto de tres ecuaciones lineales con tres incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$, y se resuelve fácilmente dados los datos para y_1, y_2, z_1 y z_2. Los estimadores reciben el nombre de *estimadores de variables instrumentales*. Si se piensa que y_2 es exógena y se eligen $z_2 = y_2$, las ecuaciones (15.25) son exactamente las condiciones de primer orden para los estimadores de MCO; vea las ecuaciones (3.13)

Aún se necesita que la variable instrumental z_2 esté correlacionada con y_2, pero el sentido en que deben estar correlacionadas estas dos variables se complica por la presencia de z_1 en la ecuación (15.22). Ahora se debe expresar el supuesto en términos de la correlación *parcial*. La forma más sencilla de expresar la condición es escribir la variable explicativa endógena como una función lineal de las variables exógenas y un término de error:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2, \qquad \textbf{15.26}$$

donde, por definición, $E(v_2) = 0$, $\mathrm{Cov}(z_1,v_2) = 0$ y $\mathrm{Cov}(z_2,v_2) = 0$, y las π_j son parámetros desconocidos. La condición clave de identificación [junto con (15.24)] es que

$$\pi_2 \neq 0. \qquad \textbf{15.27}$$

En otras palabras, después de descontar el efecto parcial de z_1, sucede que y_2 y z_2 aún están correlacionadas. Esta correlación puede ser positiva o negativa, pero no puede ser igual a cero. Probar (15.27) es sencillo: se estima (15.26) mediante MCO y se utiliza una prueba t (posiblemente haciéndola robusta a la heterocedasticidad). Siempre se debe probar este supuesto. Por desgracia, no se puede probar que z_1 y z_2 no se correlacionen con u_1; con suerte, esto se puede justificar con base en el razonamiento económico o introspección.

La ecuación (15.26) es un ejemplo de una **ecuación en la forma reducida**, lo cual significa que se ha escrito una variable endógena en términos de variables exógenas. Este nombre proviene de los modelos de ecuaciones simultáneas, que se estudiarán en el siguiente capítulo, pero es un concepto útil, siempre que se tenga una variable explicativa endógena. El nombre ayuda a distinguirla de la ecuación estructural (15.22).

Pregunta 15.2

Suponga que se desea estimar el efecto del consumo de la marihuana en el promedio de las calificaciones universitarias (*colGPA*). Para la población de estudiantes universitarios de últimos grados, *daysused* denota el número de días en el mes pasado en que un estudiante fumó marihuana y considere la ecuación estructural

$$colGPA = \beta_0 + \beta_1 daysused + \beta_2 SAT + u.$$

i) Sea *percHS* el porcentaje de estudiantes graduados de una clase de bachillerato que reportaron consumir marihuana de manera regular. Si esta es un candidato para *daysused*, escriba la ecuación de la forma reducida para *daysused*. ¿Considera que (15.27) tenga probabilidad de ser cierta?

ii) ¿Considera que *percHS* es verdaderamente exógena en la ecuación estructural? ¿Qué problemas podrían existir?

Agregar más **variables explicativas exógenas** al modelo es sencillo. Escriba el modelo estructural de la siguiente manera

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1, \tag{15.28}$$

donde y_2 está correlacionada con u_1. Sea z_k una variable que no está en (15.28) que también es exógena. Por tanto, se supone que

$$\mathrm{E}(u_1) = 0, \mathrm{Cov}(z_j,u_1) = 0, \quad j = 1, \dots, k. \tag{15.29}$$

De acuerdo con (15.29), z_1, ..., z_{k-1} son las variables exógenas que aparecen en (15.28). En realidad, éstas actúan como sus propias variables instrumentales cuando se estiman las β_j en (15.28). El caso especial de $k = 2$ se da en las ecuaciones en (15.25); junto con z_2, z_1 aparece en el conjunto de las condiciones de momentos que se utilizaron para obtener las estimaciones de VI. En términos más generales, z_1, ..., z_{k-1} se emplean en las condiciones de momentos junto con la variable instrumental para y_2, z_k.

La forma reducida para y_2 es

$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2, \tag{15.30}$$

y se necesita alguna correlación parcial entre z_k y y_2:

$$\pi_k \neq 0. \tag{15.31}$$

De acuerdo con (15.29) y (15.31), z_k es una VI válida para y_2. [No importan las π_j restantes en (15.30); algunas o todas podrían ser cero.] Un supuesto adicional menor es que no existen relaciones lineales perfectas entre las variables exógenas; esto es análogo al supuesto de ninguna colinealidad perfecta en el contexto de MCO.

Para la inferencia estadística estándar, es necesario suponer la homocedasticidad de u_1. Se da un planteamiento cuidadoso de estos supuestos en un escenario más general en la sección 15.3.

Ejemplo 15.4

[La proximidad de la universidad como una VI para la educación]

Card (1995) empleó los datos del salario y la educación para una muestra de hombres en 1976 para estimar el rendimiento de la educación. Utilizó una variable binaria para el caso de cualquiera que hubiera crecido cerca de una universidad con carreras de 4 años (*nearc4*) como una variable instrumental para la educación. En una ecuación de log(*wage*) incluyó otros controles estándar: experiencia (*exper*), una variable binaria para raza negra (*black*), una variable binaria por vivir en un área metropolitana (*smsa*) y una binaria por vivir en el sur (*south*), así como un conjunto completo de variables binarias regionales y una variable binaria de área metropolitana para el lugar donde el hombre vivía en 1966. Con el fin de que *nearc4* sea un instrumento válido, no debe estar correlacionada con el término de error en la ecuación de salario (se supone esto) y debe estar correlacionada parcialmente con *educ*. Para verificar este último requisito, se realiza una regresión de *educ* sobre *nearc4* y todas las variables exógenas que aparecen en la ecuación. (Es decir, se estima la forma reducida para *educ*.) Mediante los datos en CARD.RAW, se obtiene, en forma condensada,

$$\begin{aligned} \widehat{educ} &= \underset{(.24)}{16.64} + \underset{(.088)}{.320}\, nearc4 - \underset{(.034)}{.413}\, exper + \dots \\ n &= 3{,}010,\ R^2 = .477. \end{aligned} \tag{15.32}$$

TABLA 15.1

Variable dependiente: log(*wage*)

Variables explicativas	MCO	VI
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observaciones *R*-cuadrada	3,010 .300	3,010 .238
Otros controles: *smsa66*, *reg662*, …, *reg669*		

Se busca el coeficiente y el estadístico *t* de *nearc4*. El coeficiente implica que en 1976, todo lo demás constante (experiencia, raza, región, etc.), las personas que vivían cerca de una universidad en 1966 tenían, en promedio, cerca de un tercio de año más de educación que aquellos que no crecieron cerca de una universidad. El estadístico *t* de *nearc4* es 3.64, lo cual da un valor-*p* que es de cero en los primeros tres decimales. Por tanto, si *nearc4* no está correlacionada con los factores inobservables en el término de error, se puede utilizar *nearc4* como una VI para *educ*.

Las estimaciones de MCO y de VI se dan en la tabla 15.1. Es interesante que la estimación de VI del rendimiento de la educación sea casi el doble que la de MCO, pero el error estándar de la estimación de VI es más de 18 veces mayor que el error estándar de MCO. El intervalo de confianza a 95% para la estimación de VI está entre .024 y .239, lo cual es un rango muy amplio. La presencia de intervalos de confianza mayores es un precio que se debe pagar para obtener un estimador consistente del rendimiento de la educación cuando se piensa que *educ* es endógena.

Como se analizó antes, no se debe hacer nada con la *R* cuadrada menor en la estimación de VI: por definición, la *R*-cuadrada de MCO siempre será más grande porque MCO minimiza la suma de los residuales cuadrados.

15.3 Mínimos cuadrados en dos etapas

En la sección anterior se supuso que se tenía una sola variable explicativa endógena (y_2), junto con una variable instrumental para y_2. Suele ocurrir que se tiene más de una variable exógena que se excluye del modelo estructural y que puede estar correlacionada con y_2, lo cual significa que son VI válidas para y_2. En esta sección se analizará cómo emplear variables instrumentales múltiples.

Una sola variable explicativa endógena

Considere nuevamente el modelo estructural (15.22), el cual tiene una variable explicativa endógena y una exógena. Ahora suponga que se tienen *dos* variables exógenas excluidas de (15.22): z_2 y z_3. Los supuestos de que z_2 y z_3 no aparecen en (15.22) y que no están correlacionadas con el error u_1 se conocen como **restricciones de exclusión**.

Si z_2 y z_3 están correlacionadas con y_2, se podría emplear cada una como una VI, al igual que en la sección anterior. Pero entonces, se tendrían dos estimadores de VI y, en general, ninguno de éstos sería eficiente. Ya que cada una de z_1, z_2 y z_3 no están correlacionadas con u_1, cualquier combinación lineal tampoco está correlacionada con u_1, y por tanto, cualquier combinación lineal de variables exógenas es una VI válida. Para encontrar la mejor VI, se elige la combinación lineal que tenga la correlación más alta con y_2. Esto está dado por la ecuación en la forma reducida para y_2. Esta se escribe

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2, \quad \textbf{15.33}$$

donde

$$\mathrm{E}(v_2) = 0,\ \mathrm{Cov}(z_1,v_2) = 0,\ \mathrm{Cov}(z_2,v_2) = 0 \text{ y } \mathrm{Cov}(z_3,v_2) = 0.$$

Entonces, la mejor VI para y_2 (de acuerdo con los supuestos que se dan en el apéndice del capítulo) es la combinación lineal de las z_j en (15.33), que se denomina y_2^*:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3. \quad \textbf{15.34}$$

Para que esta VI no se correlacione de manera perfecta con z_1 se necesita al menos que una de π_2 o π_3 sea diferente de cero:

$$\pi_2 \neq 0 \text{ o } \pi_3 \neq 0. \quad \textbf{15.35}$$

Este es el supuesto clave de identificación, una vez que se supone que todas las z_j son exógenas. (El valor de π_1 es irrelevante.) La ecuación estructural (15.22) no se identifica si $\pi_2 = 0$ y $\pi_3 = 0$. Se puede probar H_0: $\pi_2 = 0$ y $\pi_3 = 0$ contra (15.35) mediante un estadístico F.

Una forma útil de pensar en (15.33) es que se descompone y_2 en dos partes. La primera es y_2^*; esta es la parte de y_2 que no está correlacionada con el término de error, u_1. La segunda es v_2, la cual está correlacionada quizá con u_1, razón por la cual y_2 sea posiblemente endógena.

Dados los datos sobre las z_j, se puede calcular y_2^* para cada observación, siempre y cuando se conozcan los parámetros poblacionales π_j, lo cual nunca sucede en la práctica. No obstante, como se vio en la sección anterior, siempre se puede estimar la forma reducida mediante MCO. Por tanto, empleando la muestra, se realiza la regresión de y_2 sobre z_1, z_2 y z_3 y se obtienen los valores ajustados:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \quad \textbf{15.36}$$

(es decir, se tiene $\hat{y}_{i2}$ para cada i). En este punto, se debe verificar que z_2 y z_3 sean conjuntamente significativas en (15.33) a un nivel de significancia razonablemente pequeño (no mayor que 5%). Si z_2 y z_3 no son conjuntamente significativas en (15.33), entonces se está perdiendo el tiempo con la estimación de VI.

Una vez que se tiene $\hat{y}_2$, se puede utilizar como la VI para y_2. Las tres ecuaciones para estimar β_0, β_1 y β_2 son las primeras dos ecuaciones de (15.25), con la tercera remplazada por

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0. \tag{15.37}$$

Al resolver las tres incógnitas de las tres ecuaciones se obtienen los estimadores de VI.

Con instrumentos múltiples, el estimador de VI que usa $\hat{y}_{i2}$ como el instrumento también recibe el nombre de **estimador de mínimos cuadrados en dos etapas (MC2E)**. La razón es simple. Al usar el álgebra de MCO, se puede mostrar que cuando se emplea $\hat{y}_2$ como la VI para y_2, las estimaciones de VI $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ son *idénticas* a las estimaciones de MCO a partir de la regresión de

$$y_1 \text{ sobre } \hat{y}_2 \text{ y } z_1. \tag{15.38}$$

En otras palabras, se puede obtener el estimador de MC2E en dos etapas. La primera consiste en realizar la regresión en (15.36), donde se obtienen los valores ajustados de $\hat{y}_2$. La segunda etapa es la regresión de MCO (15.38). Debido a que se utiliza $\hat{y}_2$ en lugar de y_2, las estimaciones de MC2E pueden diferir sustancialmente de las de MCO.

Algunos economistas gustan de interpretar la regresión en (15.38) de la manera siguiente. El valor ajustado, $\hat{y}_2$, es la versión estimada de y_2^*, y y_2^* no está correlacionada con u_1. Por tanto, MC2E primero "purga" a y_2 de su correlación con u_1 antes de efectuar la regresión por MCO en (15.38). Esto se puede demostrar al insertar $y_2 = y_2^* + v_2$ en (15.22):

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2. \tag{15.39}$$

Ahora, el error compuesto $u_1 + \beta_1 v_2$ tiene una media de cero y no está correlacionada con y_2^* ni con z_1, razón por la cual funciona la regresión de MCO en (15.38).

La mayoría de los paquetes econométricos tiene comandos especiales para MC2E, así que no hay necesidad de desarrollar de manera explícita las dos etapas. De hecho, en la mayoría de los casos se debe evitar realizar la segunda etapa de forma manual, pues los errores estándar y los estadísticos de prueba que se obtienen de esta forma *no son* válidos. [La razón es que el término de error en (15.39) incluye a v_2, pero los errores estándar implican sólo la varianza de u_1.] Cualquier software de regresión que soporte MC2E pedirá la variable dependiente, la lista de variables explicativas (tanto exógenas como endógenas) y la lista completa de variables instrumentales (es decir, todas las variables exógenas). El resultado suele ser muy similar al de MCO.

En el modelo (15.28) con una sola VI para y_2, el estimador de VI de la sección 15.2 es idéntico al estimador de MC2E. Por tanto, cuando se tiene una VI para cada variable explicativa endógena, se puede llamar al método de estimación de VI o MC2E.

Agregar más variables exógenas tiene repercusiones mínimas. Por ejemplo, suponga que la ecuación de salario es

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1, \tag{15.40}$$

donde u_1 no está correlacionada con *exper* ni con $exper^2$. Suponga también que se piensa que la educación del padre y de la madre no están correlacionadas con u_1. Entonces, se pueden emplear ambas como VI para *educ*. La ecuación en su forma reducida para *educ* es

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2, \tag{15.41}$$

y la identificación requiere que $\pi_3 \neq 0$ o $\pi_4 \neq 0$ (o ambas, por supuesto).

Ejemplo 15.5

[Rendimientos de la educación para la mujer trabajadora]

Se estima la ecuación (15.40) mediante los datos en MROZ.RAW. Primero, se prueba H_0: $\pi_3 = 0$, $\pi_4 = 0$ en (15.41) usando una prueba F. El resultado es $F = 55.40$, y valor-$p = .0000$. Como se esperaba, *educ* está correlacionada (parcialmente) con la educación de los padres.

Cuando se estima (15.40) mediante MC2E, se obtiene, en forma de ecuación,

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(.400)}{.048} + \underset{(.031)}{.061}\, educ + \underset{(.013)}{.044}\, exper - \underset{(.0004)}{.0009}\, exper^2 \\ n &= 428,\ R^2 = .136. \end{aligned}$$

El rendimiento estimado de la educación es de aproximadamente 6.1%, comparado con una estimación de MCO de cerca de 10.8%. Debido a su error estándar relativamente grande, la estimación de MC2E es apenas estadísticamente significativa al nivel de 5% contra una alternativa de dos colas.

Los supuestos necesarios para que MC2E tenga las propiedades deseadas de muestras grandes se dan en el apéndice del capítulo, pero sería útil resumirlas brevemente aquí. Si se escribe la ecuación estructural como en (15.28),

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1, \tag{15.42}$$

entonces, se supone que cada z_j no está correlacionada con u_1. Además, se necesita que al menos una variable exógena que *no* esté en (15.42) esté parcialmente correlacionada con y_2. Esto asegura consistencia. Para que los errores estándar habituales de MC2E y los estadísticos t sean asintóticamente válidos, también se necesita un supuesto de homocedasticidad: la varianza del error estructural, u_1, no puede depender de ninguna de las variables exógenas. Para las aplicaciones de series de tiempo, se necesitan más supuestos, como se verá en la sección 15.7.

Multicolinealidad y MC2E

En el capítulo 3 se presentó el problema de la multicolinealidad y se mostró cómo la correlación entre regresores puede ocasionar grandes errores estándar para las estimaciones de MCO. La multicolinealidad puede ser incluso más seria con MC2E. Para saber por qué, la varianza (asintótica) del estimador de MC2E de β_1 puede aproximarse como

$$\sigma^2/[\widehat{\mathrm{STC}}_2(1 - \hat{R}_2^2)], \tag{15.43}$$

donde $\sigma^2 = \mathrm{Var}(u_1)$, $\widehat{\mathrm{STC}}_2$ es la variación total en $\hat{y}_2$, y $\hat{R}_2^2$ es la R cuadrada de una regresión de $\hat{y}_2$ sobre todas las demás variables exógenas que aparecen en la ecuación estructural. Existen dos razones por las que la varianza del estimador de MC2E sea mayor que la de MCO. Primero,

$\hat{y}_2$, por definición, tiene una variación menor que y_2. (Recuerde que: suma total de cuadrados = suma explicada de cuadrados + suma residual de cuadrados; la variación en y_2 es la suma total de cuadrados, mientras que la variación en $\hat{y}_2$ es la suma explicada de cuadrados a partir de la regresión de la primera etapa.) Segundo, la correlación entre $\hat{y}_2$ y las variables exógenas en (15.42) es mucho más alta que la correlación entre y_2 y esas variables. Esto esencialmente define el problema de la multicolinealidad en MC2E.

Como ilustración, considere el ejemplo 15.4. Cuando se realiza la regresión de *educ* sobre las variables exógenas en la tabla 15.1 (sin incluir *nearc4*), R-cuadrada = .475; este es un grado moderado de multicolinealidad, pero lo importante es que el error estándar de MCO para $\hat{\beta}_{educ}$ es muy pequeño. Cuando se obtienen los valores ajustados de la primera etapa, $\widehat{educ}$, y se realiza la regresión de ellos sobre las variables exógenas en la tabla 15.1, R-cuadrada = .995, lo cual indica un grado muy alto de multicolinealidad entre $\widehat{educ}$ y las variables exógenas restantes en la tabla. (No es de sorprender esta R-cuadrada alta, ya que $\widehat{educ}$ es una función de todas las variables exógenas en la tabla 15.1, más *nearc4*.) La ecuación (15.43) muestra que una $\hat{R}_2^2$ cercana a uno puede producir un error estándar muy grande para el estimador de MC2E. Pero al igual que en MCO, un tamaño muestral grande puede ayudar a compensar una $\hat{R}_2^2$ grande.

Múltiples variables explicativas endógenas

Los mínimos cuadrados en dos etapas también se pueden utilizar en modelos con más de una variable explicativa endógena. Por ejemplo, considere el modelo

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1, \tag{15.44}$$

donde $E(u_1) = 0$ y u_1 no está correlacionada con z_1, z_2 ni z_3. Las variables y_2 y y_3 son variables explicativas endógenas: cada una puede estar correlacionada con u_1.

Para estimar (15.44) mediante MC2E, se necesitan *al menos dos* variables exógenas que no aparezcan en (15.44), pero que estén correlacionadas con y_2 y y_3. Suponga que se tienen dos variables exógenas excluidas, sean z_4 y z_5. Entonces, a partir del análisis de una variable explicativa, se necesita que z_4 o z_5 aparezcan en cada forma reducida para y_2 y y_3. (Como antes, se puede emplear el estadístico F para comprobar esto.) Aunque esto es necesario para la identificación, por desgracia, no basta. Suponga que z_4 aparece en cada forma reducida, pero z_5 no aparece en ninguna. Entonces, en realidad no se tienen dos variables exógenas parcialmente correlacionadas con y_2 y y_3. Los mínimos cuadrados en dos etapas no producirán estimadores consistentes de las β_j.

En general, cuando se tiene más de una variable explicativa endógena en un modelo de regresión, la identificación puede fallar por varios motivos complejos. Pero se puede establecer con facilidad una condición necesaria para la identificación, que recibe el nombre de **condición de orden**.

Condición de orden para la identificación de una ecuación. Se necesitan al menos tantas variables exógenas excluidas como variables explicativas endógenas incluidas en la ecuación estructural. La condición de orden es fácil de revisar, pues sólo exige contar las variables endógenas y exógenas. La condición suficiente

Pregunta 15.3

El siguiente modelo explica las tasas de delitos violetos (*violent*) al nivel de ciudad, en términos de una variable binaria para la existencia de leyes de control de armas (*guncontrol*) y otros controles:

$$violent = \beta_0 + \beta_1 guncontrol + \beta_2 unem + \beta_3 popul + \beta_4 percblck + \beta_5 age18\_21 + \dots.$$

Algunos investigadores han estimado ecuaciones similares con variables como el número de miembros en la Asociación Nacional de Poseedores de Rifle en la ciudad y el número de suscriptores a las revistas de armas como variables instrumentales para *guncontrol* [vea, por ejemplo, Kleck y Patterson (1993)]. ¿Son convincentes estos instrumentos?

para la identificación se llama **condición de rango**. Antes se han observado casos especiales de condición de rango, por ejemplo en el análisis sobre la ecuación (15.35). Un planteamiento general de la condición de rango requiere álgebra matricial y está más allá del alcance de este libro. [Vea Wooldridge (2002, capítulo 5).]

Pruebas de hipótesis múltiples después de la estimación de MC2E

Es necesario ser muy cuidadoso cuando se prueban hipótesis múltiples en un modelo estimado por MC2E. Resulta tentador usar la forma de la suma de residuales cuadrados o la forma de la R cuadrada del estadístico F, como se observó con MCO en el capítulo 4. El hecho de que la R cuadrada en MC2E pueda ser negativa sugiere que la forma habitual de calcular los estadísticos F quizá no sea la adecuada; este es el caso. De hecho, si se utilizan los residuales de MC2E para calcular las SRC tanto para el modelo restringido como para el no restringido, no hay garantía de que $SRC_r \geq SRC_{nr}$; si lo contrario es cierto, el estadístico F sería negativo.

Es posible combinar la suma de residuales cuadrados de la regresión en la segunda etapa [como (15.38)] con la SRC_{nr} para obtener un estadístico con una distribución F aproximada en muestras grandes. Debido a que muchos paquetes de econometría tienen comandos sencillos de utilizar para probar hipótesis múltiples después de la estimación de MC2E, se omiten los detalles. Davidson y MacKinnon (1993) y Wooldridge (2002, capítulo 5) contienen el análisis de cómo calcular estadísticos del tipo F para MC2E.

15.4 Soluciones de VI a los problemas de errores en las variables

En las secciones anteriores se presentó el uso de las variables instrumentales como una forma de resolver el problema de las variables omitidas, pero también se pueden utilizar para manejar el problema de error de medición. Como ejemplo, considere el siguiente modelo

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u, \tag{15.45}$$

donde y y x_2 se observan pero x_1^* no. Sea x_1 una medida observada de x_1^*: $x_1 = x_1^* + e_1$, donde e_1 es el error de medición. En el capítulo 9 se mostró que la correlación entre x_1 y e_1 ocasiona que MCO, en los que x_1 se emplea en lugar de x_1^*, sean sesgados e inconsistentes. Esto se puede observar al escribir

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1). \tag{15.46}$$

Si los supuestos de los errores clásicos en las variables (ECV) son válidos, el sesgo en el estimador de MCO de β_1 es hacia cero. Sin supuestos adicionales, no se puede hacer nada al respecto.

En algunos casos se puede utilizar un procedimiento de VI para resolver el problema de error de medición. En (15.45), se supone que u no está correlacionada con x_1^*, x_1 ni x_2; en el caso de los ECV, se supone que e_1 no está correlacionada con x_1^* ni x_2. Esto implica que x_2 es exógena en (15.46), pero que x_1 está correlacionada con e_1. Lo que se necesita es una VI para x_1. Tal VI debe estar correlacionada con x_1, pero no con u, de manera que pueda excluirse de (15.45), y no estar correlacionada con el error de medición, e_1.

Una posibilidad es obtener una segunda medida de x_1^*, por ejemplo, z_1. Debido a que es x_1^* lo que influye a y, es natural suponer que z_1 no esté correlacionada con u. Si se escribe $z_1 = x_1^* + a_1$, donde a_1 es el error de medición en z_1, entonces se debe suponer que a_1 y e_1 no están correlacionadas. En otras palabras, x_1 y z_1 son medidas equivocadas de x_1^*, pero sus errores de medición no están correlacionados. Desde luego, x_1 y z_1 están correlacionados a través de su dependencia de x_1^*, de manera que se puede usar z_1 como una VI para x_1.

¿Dónde se podrán obtener dos medidas sobre una variable? A veces, cuando se le pregunta a un grupo de trabajadores sobre su salario anual, sus empleadores pueden ofrecer una segunda medida. Para parejas casadas, cada cónyuge puede informar de forma independiente el nivel de ahorros o el nivel de ingreso familiar. En el estudio de Ashenfelter y Krueger (1994) citado en la sección 14.3, a cada gemelo se le preguntó acerca de los años de educación de sus hermanos; esto da una segunda medida que se puede utilizar como una VI para la educación reportada en una ecuación de salario. (Ashenfelter y Krueger combinaron las primeras diferencias y las VI para dar cuenta también del problema de la capacidad innata omitida; se abundará más al respecto en la sección 15.8.) Por lo general, sin embargo, tener dos medidas de una variable explicativa es raro.

Una alternativa es utilizar otras variables exógenas como VI para una variable potencialmente mal medida. Por ejemplo, el uso de *motheduc* y *fatheduc* como VI de *educ* en el ejemplo 15.5 sirve a este propósito. Si se piensa que $educ = educ^* + e_1$, entonces las estimaciones de VI en el ejemplo 15.5 no adolecen del error de medición si *motheduc* y *fatheduc* no están correlacionados con el error de medición, e_1. Esto probablemente es más razonable que suponer que *motheduc* y *fatheduc* no están correlacionados con la capacidad, la cual está contenida en u en (15.45).

Los métodos de VI también se pueden adoptar cuando se utilizan cuestiones como las calificaciones de exámenes para controlar las características inobservables. En la sección 9.2 se mostró que, con base en ciertos supuestos, se pueden utilizar variables proxy para resolver el problema de variables omitidas. En el ejemplo 9.3 se utilizó IQ como una variable proxy para la capacidad inobservable. Esto simplemente implica sumar IQ al modelo y realizar una regresión de MCO. Pero existe una alternativa que funciona cuando IQ no satisface completamente los supuestos de la variable proxy. A manera de ilustración, escriba una ecuación de salario como

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u, \quad \boxed{\mathbf{15.47}}$$

donde nuevamente se tiene el problema de la capacidad omitida. Pero se tienen dos calificaciones de exámenes ($test_1$ y $test_2$) que son *indicadores* de la capacidad (*abil*). Se supone que las calificaciones se pueden escribir como

$$test_1 = \gamma_1 abil + e_1$$

y

$$test_2 = \delta_1 abil + e_2,$$

donde $\gamma_1 > 0$, $\delta_1 > 0$. Dado que la capacidad es la que afecta al salario, se puede suponer que $test_1$ y $test_2$ no están correlacionados con u. Si se escribe *abil* en términos de la calificación del primer examen y se inserta el resultado en (15.47), se obtiene

$$\begin{aligned} \log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 \\ + \alpha_1 test_1 + (u - \alpha_1 e_1), \end{aligned} \quad \boxed{\mathbf{15.48}}$$

donde $\alpha_1 = 1/\gamma_1$. Ahora, si se supone que e_1 no está correlacionada con ninguna de las variables explicativas en (15.47), incluida *abil*, entonces e_1 y $test_1$ *deben* estar correlacionadas. [Observe que *educ no* es endógena en (15.48); no obstante, $test_1$ sí lo es.] Esto significa que estimar (15.48) mediante MCO producirá estimadores inconsistentes de las β_j (y α_1). De acuerdo con los supuestos que se han formulado, $test_1$ no satisface los supuestos de la variable proxy.

Si se supone que e_2 tampoco está correlacionada con ninguna de las variables explicativas en (15.47) *además de* que e_1 y e_2 no están correlacionadas, entonces e_1 no está correlacionada con la segunda calificación del examen, $test_2$. Por tanto, $test_2$ puede utilizarse como una VI para $test_1$.

Ejemplo 15.6

[Uso de dos calificaciones de un examen como indicadores de la capacidad]

Se emplean los datos en WAGE2.RAW para implementar el procedimiento anterior, donde *IQ* desempeña la función de la primera calificación de examen y *KWW* (conocimiento del mundo laboral) es la segunda calificación de examen. Las variables explicativas son las mismas que en el ejemplo 9.3: *educ*, *exper*, *tenure*, *married*, *south*, *urban* y *black*. En lugar de agregar *IQ* y hacer MCO, como en la columna (2) de la tabla 9.2, se agrega *IQ* y se usa *KWW* como su instrumento. El coeficiente de *educ* es .025 (ee = .017). Esta es una estimación baja y no es estadísticamente diferente de cero. Esto es un hallazgo desconcertante, y sugiere que uno de nuestros supuestos no es válido; quizá e_1 y e_2 están correlacionados.

15.5 Pruebas de endogeneidad y pruebas de restricciones de sobreidentificación

En esta sección se describen dos pruebas importantes en el contexto de la estimación de variables instrumentales.

Prueba de endogeneidad

El estimador de MC2E es menos eficiente que el de MCO cuando las variables explicativas son exógenas; como se ha observado, las estimaciones de MC2E pueden tener errores estándar muy grandes. Por tanto, es útil tener una prueba para la endogeneidad de una variable explicativa que muestre si los MC2E aún se necesitan. Obtener una prueba de esta naturaleza es muy sencillo.

Para ejemplificar, suponga que se tiene una sola supuesta variable endógena,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1, \quad \boxed{15.49}$$

donde z_1 y z_2 son exógenas. Se tienen dos variables exógenas adicionales, z_3 y z_4, lo cual no aparece en (15.49). Si y_2 no está correlacionada con u_1, se debe estimar (15.49) mediante MCO. ¿Cómo se puede probar esto? Hausman (1978) sugirió que se compararan directamente las estimaciones de MCO y MC2E y se determinase si las diferencias eran estadísticamente significativas. Después de todo, MCO y MC2E son consistentes si todas las variables son exógenas. Si MC2E y MCO difieren de forma significativa, se concluye que y_2 debe ser endógena (sosteniendo que las z_j son exógenas).

Sería una buena idea calcular MCO y MC2E para saber si las estimaciones son diferentes en términos prácticos. Para determinar si las diferencias son estadísticamente significativas, es más sencillo emplear una prueba de regresión. Ésta se basa en la estimación de la forma reducida de y_2, que en este caso es

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2. \quad \boxed{15.50}$$

Ahora, dado que cada z_j no está correlacionada con u_1, y_2 no está correlacionada con u_1 si, y sólo si, v_2 no está correlacionada con u_1; esto es lo que se desea probar. Escriba $u_1 = \delta_1 v_2 + e_1$, donde

e_1 no está correlacionada con v_2 y tiene media cero. Entonces, u_1 y v_2 no están correlacionadas si, y sólo si, $\delta_1 = 0$. La forma más fácil de probar esto es incluir v_2 como un regresor adicional en (15.49) y hacer una prueba t. Existe sólo un problema con implementar esto: v_2 no se observó, porque es el término de error en (15.50). Dado que se puede estimar la forma reducida de y_2 mediante MCO, se pueden obtener los residuales de la misma, $\hat{v}_2$. Por tanto, se estima

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error \qquad \boxed{15.51}$$

por MCO y se prueba H_0: $\delta_1 = 0$ mediante un estadístico t. Si se rechaza H_0 a un nivel de significancia pequeño, se concluye que y_2 es endógeno debido a que v_2 y u_1 están correlacionadas.

Prueba de endogeneidad de una sola variable explicativa:

i) Se estima la forma reducida de y_2 mediante su regresión sobre *todas* las variables exógenas (incluidas las de la ecuación estructural y las VI adicionales). Se obtienen los residuales, $\hat{v}_2$.

ii) Se agrega $\hat{v}_2$ a la ecuación estructural (que incluye a y_2) y se ejecuta una prueba de la significancia de $\hat{v}_2$ mediante una regresión de MCO. Si el coeficiente de $\hat{v}_2$ es estadísticamente diferente de cero, se concluye que y_2 es en realidad endógena. Quizá se desee utilizar la prueba t robusta a la heterocedasticidad.

Ejemplo 15.7

[Rendimiento de la educación para las mujeres trabajadoras]

Se puede probar la endogeneidad de *educ* en (15.40) al obtener los residuales $\hat{v}_2$ de estimar la forma reducida (15.41), sólo con datos de mujeres trabajadoras, e incluirlos en (15.40). Cuando se hace esto, el coeficiente de $\hat{v}_2$ es $\hat{\delta}_1 = .058$, y $t = 1.67$. Esta es evidencia moderada de correlación positiva entre u_1 y v_2. Quizá sea buena idea reportar ambas estimaciones, porque la estimación de MC2E del rendimiento de la educación (6.1%) está muy por debajo de la estimación de MCO (10.8%).

Una característica interesante de la regresión del paso ii) de la prueba de endogeneidad es que los coeficientes estimados de todas las variables explicativas (salvo, por supuesto, $\hat{v}_2$) son idénticas a las estimaciones de MC2E. Por ejemplo, estimar (15.51) por MCO produce las mismas $\hat{\beta}_j$ que estimar (15.49) por MC2E. Un beneficio de esta equivalencia es que ofrece una forma fácil de verificar si se ha realizado la regresión apropiada para probar la endogeneidad. Pero esto también da una interpretación útil y diferente de MC2E: agregar $\hat{v}_2$ a la ecuación original como una variable explicativa y aplicar MCO resuelve la endogeneidad de y_2. Así, cuando se comienza por estimar (15.49) mediante MCO, se puede cuantificar la importancia de permitir que y_2 sea endógena al ver cuánto cambia $\hat{\beta}_1$ cuando $\hat{v}_2$ se agrega a la ecuación. Sin importar el resultado de las pruebas estadísticas, se puede apreciar si el cambio en $\hat{\beta}_1$ es el esperado y si es prácticamente significativo.

También se puede probar la endogeneidad para muchas variables explicativas. Para cada supuesta variable endógena, se obtienen los residuales de la forma reducida como en la parte i). Después, se prueba la significancia conjunta de los residuales en la ecuación estructural, mediante una prueba F La significancia conjunta indica que al menos una supuesta variable explicativa es endógena. El número de las restricciones de exclusión probadas es el número de variables explicativas endógenas supuestas.

Prueba de restricciones de sobreidentificación

Cuando se presentó el estimador simple de variables instrumentales en la sección 15.1, se enfatizó que el instrumento tenía que satisfacer dos requisitos: no estar correlacionado con el error (exogeneidad) y estar correlacionado con la variable explicativa endógena (relevancia). Ahora se ve que, incluso en modelos con variables explicativas adicionales, el segundo requerimiento se puede probar mediante una prueba t (con sólo un instrumento) o una prueba F (cuando existen múltiples instrumentos). En el contexto del estimador simple de VI, se observa que el requisito de exogeneidad no se puede probar. No obstante, si se tienen más instrumentos de los necesarios, se puede probar efectivamente si algunas de ellas no están correlacionadas con el error estructural.

Como ejemplo específico, considere nuevamente la ecuación (15.49) con dos variables instrumentales para y_2, z_3 y z_4. Recuerde que z_1 y z_2 actúan esencialmente como sus propios instrumentos. Dado que se tienen dos instrumentos para y_2, se puede estimar (15.49) sólo mediante z_3 como VI para y_2; sea $\check{\beta}_1$ el estimador de VI resultante de β_1. Después, se estima (15.49) sólo con z_4 como una VI para y_2; sea éste el estimador $\tilde{\beta}_1$ de VI. Si todas las z_j son exógenas, y si z_3 y z_4 están correlacionadas cada una de manera parcial con y_2, entonces tanto $\check{\beta}_1$ como $\tilde{\beta}_1$ son consistentes para β_1. Por tanto, si la lógica con que se eligieron los instrumentos es correcta, $\check{\beta}_1$ y $\tilde{\beta}_1$ deben diferir sólo por el error de muestreo. Hausman (1978) propuso basar una prueba de si tanto z_3 como z_4 eran exógenas en la diferencia $\check{\beta}_1 - \tilde{\beta}_1$. Más adelante se proporcionará una forma más sencilla de obtener una prueba válida, pero antes de hacerlo, es necesario comprender cómo interpretar el resultado de la prueba.

Si se concluye que $\check{\beta}_1$ y $\tilde{\beta}_1$ son estadísticamente diferentes entre sí, entonces no se tendrá otra opción más que concluir que z_3, z_4, o ambas no cumplieron con el requisito de exogeneidad. Por desgracia, no se puede saber cuál es el caso (a menos que simplemente se afirme desde el principio que, por decir, z_3 es exógena). Por ejemplo, si y_2 denota años de escolaridad en una ecuación para el logaritmo del salario, z_3 es la educación de la madre y z_4 es la educación del padre, una diferencia estadísticamente significativa en los dos estimadores de VI implica que una o ambas variables de la educación de los padres están correlacionadas con u_1 en (15.54).

Por supuesto, rechazar que los instrumentos sean exógenos es un asunto serio y requiere un nuevo enfoque. Pero el problema más serio y delicado cuando se comparan las estimaciones de VI es que pueden ser similares aunque ambos instrumentos no cumplan con el requisito de la exogeneidad. En el ejemplo anterior parece probable que si la educación de la madre está positivamente correlacionada con u_1, entonces así sucederá también con la del padre. Por tanto, las dos estimaciones de VI pueden ser similares, aunque cada una de ellas sea inconsistente. En efecto, debido a que las VI en este ejemplo se eligen con un razonamiento similar, su uso individual en los procedimientos de VI puede ocasionar estimaciones similares que, sin embargo, sean ambas inconsistentes. El punto es que no hay por qué estar particularmente tranquilo si los procedimientos de VI pasan la prueba de Hausman.

Otro problema cuando se comparan dos estimaciones de VI es que, con frecuencia, pueden parecer prácticamente diferentes, sin embargo, estadísticamente, no se puede rechazar la hipótesis nula de que sean consistentes para el mismo parámetro poblacional. Por ejemplo, cuando se estima (15.40) mediante VI empleando *motheduc* como el único instrumento, el coeficiente de *educ* es .049 (.037). Si se utiliza sólo *fatheduc* como la VI para *educ*, el coeficiente de *educ* es .070(.034). [Quizá no sea de sorprender que la estimación que usa la educación de ambos padres como variables instrumentales está entre aquellas dos, .061(.031).] Para propósitos de políticas, la diferencia entre 5 y 7% para el rendimiento estimado de un año de escuela es sustancial. Sin embargo, como se mostró en el ejemplo 15.8, la diferencia no es estadísticamente significativa.

El procedimiento de comparar diferentes estimaciones de VI del mismo parámetro es un ejemplo de cómo probar **restricciones de sobreidentificación**. La idea general es que se tienen más instrumentos de los necesarios para estimar consistentemente los parámetros. En el ejemplo anterior se tenía un instrumento más de los que se necesitaban, y esto resultó en una restricción de sobreidentificación que se puede probar. En el caso general, suponga que se tienen q instrumentos más de los que se necesitan. Por ejemplo, con una variable explicativa endógena, y_2, y tres instrumentos propuestos para y_2, se tienen $q = 3 - 1 = 2$ restricciones de sobreidentificación. Cuando q es dos o más, comparar varias estimaciones de VI es engorroso. En lugar de ello, se puede calcular con facilidad un estadístico de prueba con base en los residuales de MC2E. La idea es que, si todos los instrumentos son exógenos, los residuales de MC2E no deben estar correlacionados con los instrumentos, excepto por el error de muestreo. Pero si existen $k + 1$ parámetros y $k + 1 + q$ instrumentos, los residuales de MC2E tienen una media de cero y no están correlacionados con ninguna de las k combinaciones lineales de los instrumentos. (Este hecho algebraico contiene, como caso especial, que los residuales de MCO tienen una media de cero y no están correlacionados con las k variables explicativas.) Por tanto, la prueba verifica si los residuales de MC2E están correlacionados con q funciones lineales de los instrumentos y no se necesita decidir sobre las funciones; la prueba lo hace automáticamente.

La siguiente prueba basada en una regresión es válida cuando el supuesto de homocedasticidad, listada como el supuesto MC2E.5 en el apéndice del capítulo, también lo es.

Prueba de restricciones de sobreidentificación:

i) Se estima la ecuación estructural mediante MC2E y se obtienen los residuales de MC2E, $\hat{u}_1$.

ii) Se realiza la regresión de $\hat{u}_1$ sobre todas las variables *exógenas*. Se obtiene la R-cuadrada, digamos, R_1^2.

iii) Con base en la hipótesis nula de que todas las VI no están correlacionadas con u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$, donde q es el número de variables instrumentales externas al modelo menos el número total de variables explicativas endógenas. Si nR_1^2 excede (por ejemplo) el valor crítico de 5% en la distribución χ_q^2, se rechaza H_0 y se concluye que por lo menos algunas de las VI no son exógenas.

Ejemplo 15.8

[Rendimiento de la educación para las mujeres trabajadoras]

Cuando se utiliza *motheduc* y *fatheduc* como VI para *educ* en (15.40), se tiene una sola restricción de sobreidentificación. Realizar la regresión de los residuales de MC2E $\hat{u}_1$ sobre *exper*, *exper*^{2}, *motheduc*, y *fatheduc* produce $R_1^2 = .0009$. Por tanto, $nR_1^2 = 428(.0009) = .3852$, lo cual es un valor muy pequeño en una distribución χ_1^2 (valor-p = .535). Así, las variables de la educación de los padres pasan la prueba de sobreidentificación. Cuando se agrega la educación del cónyuge a la lista de VI, se obtienen dos restricciones de sobreidentificación, y $nR_1^2 = 1.11$ (valor-p = .574). Sujeto a las precauciones precedentes, parece razonable agregar *huseduc* a la lista de VI, pues esto reduce el error estándar de la estimación de MC2E: la estimación de MC2E para *educ* mediante los tres instrumentos es .080 (ee = .022), así que esto hace que *educ* sea mucho más significativa que cuando *huseduc* no se utilizó como una IV ($\hat{\beta}_{educ} = .061$, ee = .031).

Cuando $q = 1$, una pregunta natural es: ¿cómo se compara la prueba basada en una regresión con una prueba basada en comparar directamente las estimaciones? En realidad, los dos procedimientos son asintóticamente iguales. Como una cuestión práctica, tiene sentido calcular las dos estimaciones de VI para saber en qué difieren. En términos más generales, cuando $q \geq 2$, se pueden comparar las estimaciones de MC2E que usan todas las instrumentales con las estimaciones de VI que usan instrumentos individuales. Al hacerlo, se puede apreciar si las diferentes estimaciones de VI son prácticamente diferentes, sin importar si la prueba de sobreidentificación se rechace o no.

En el ejemplo anterior se aludió a un dato general acerca de MC2E: de acuerdo con los supuestos estándar de MC2E, agregar instrumentos a la lista mejora la eficiencia asintótica de MC2E. Pero esto requiere que cualquier nuevo instrumento sea en realidad exógeno, de lo contrario, MC2E no serán ni siquiera consistentes, y este es sólo un resultado asintótico. Con los tamaños de muestra típicos disponibles, agregar demasiados instrumentos, es decir, aumentar el número de restricciones de sobreidentificación puede ocasionar un severo sesgo en MC2E. Un análisis detallado rebasa el alcance de este libro. Bound, Jaeger y Baker (1995) dan una buena ilustración; ellos argumentan que las estimaciones de MC2E del rendimiento de la educación que obtuvieron Angrist y Krueger (1991), con muchas variables instrumentales, quizás estén seriamente sesgadas (¡incluso con cientos de miles de observaciones!).

La prueba de sobreidentificación se puede utilizar siempre que se tengan más instrumentos de los que se necesitan. Si se tienen los instrumentos suficientes, se dice que el modelo está *exactamente identificado*, y la R cuadrada en la parte ii) de la prueba de sobreidentificación será idéntica a cero. Como se mencionó antes, no se puede probar la exogeneidad de los instrumentos en el caso de estar exactamente identificados.

Se puede hacer una prueba robusta a la heterocedasticidad de la forma arbitraria; para más detalles, vea Wooldridge (2002, capítulo 5).

15.6 MC2E con heterocedasticidad

La heterocedasticidad en el contexto de MC2E plantea esencialmente las mismas cuestiones que con MCO. Pero, lo más importante es que es posible obtener los errores estándar y los estadísticos de prueba que son (asintóticamente) robustos a la heterocedasticidad de la forma arbitraria y desconocida. En realidad, la expresión (8.4) continúa siendo válida si las $\hat{r}_{ij}$ se obtienen como los residuales de la regresión de $\hat{x}_{ij}$ sobre las demás $\hat{x}_{ih}$, donde el símbolo "^" denota los valores ajustados de las regresiones de la primera etapa (para variables explicativas endógenas). Wooldridge (2002, capítulo 5) contiene más detalles. Algunos paquetes de software hacen esto de forma rutinaria.

Se puede probar también la heterocedasticidad, mediante una prueba análoga a la prueba de Breusch-Pagan que se estudió en el capítulo 8. Sean $\hat{u}$ los residuales de MC2E y $z_1, z_2, \dots, z_m$ todas las variables exógenas (incluidas aquellas que se emplearon como VI para las variables explicativas endógenas). Entonces, bajo supuestos razonables [expresados, por ejemplo, en Wooldridge (2002, capítulo 5)], un estadístico asintóticamente válido es el estadístico F para la significancia conjunta en una regresión de $\hat{u}^2$ sobre $z_1, z_2, \dots, z_m$. La hipótesis nula de homocedasticidad se rechaza si las z_j son conjuntamente significativas.

Si se aplica esta prueba al ejemplo 15.8, mediante *motheduc*, *fatheduc* y *huseduc* como instrumentos para *educ*, se obtiene $F_{5,422} = 2.53$, y el valor-p = .029. Esto es evidencia de heterocedasticidad al nivel de 5%. Quizá se desee calcular los errores estándar robustos a la heterocedasticidad para dar cuenta de esto.

Si se sabe cómo depende la varianza del error de las variables exógenas, se puede utilizar un procedimiento ponderado de MC2E, esencialmente el mismo que en la sección 8.4. Después de estimar un modelo para $\mathrm{Var}(u|z_1, z_2, \dots, z_m)$, se dividen la variable dependiente, las variables explicativas y todas las variables instrumentales para la observación i entre $\sqrt{\hat{h}_i}$, donde $\hat{h}_i$ denota la varianza estimada. (La constante, que es al mismo tiempo una variable explicativa y una VI, se divide entre $\sqrt{\hat{h}_i}$; vea la sección 8.4.) Entonces se aplica MC2E a la ecuación transformada empleando los instrumentos transformados.

15.7 Aplicación de MC2E a las ecuaciones de series de tiempo

Cuando se aplica MC2E a datos de series de tiempo, muchas de las consideraciones que surgieron para MCO en los capítulos 10, 11 y 12 son relevantes. Escriba la ecuación estructural para cada periodo de tiempo como

$$y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t, \tag{15.52}$$

donde una o más variables explicativas x_{tj} pueden estar correlacionadas con u_t. Denote el conjunto de variables exógenas mediante $z_{t1}, \ldots, z_{tm}$:

$$\mathrm{E}(u_t) = 0, \mathrm{Cov}(z_{tj}, u_t) = 0, \quad j = 1, \ldots, m.$$

Cualquier variable explicativa exógena es también una z_{tj}. Para fines de identificación es necesario que $m \geq k$ (se tienen tantas variables exógenas como variables explicativas).

Pregunta 15.4

Un modelo para probar el efecto del crecimiento en el gasto gubernamental sobre el crecimiento de la producción es

$$gGDP_t = \beta_0 + \beta_1 gGOV_t + \beta_2 INVRAT_t + \beta_3 gLAB_t + u_t,$$

donde *g* indica el crecimiento, *GDP* es el producto interno bruto real, *GOV* es el gasto real del gobierno, *INVRAT* es la proporción de la inversión interna bruta respecto a *GDP*, y *LAB* es el tamaño de la fuerza laboral. [Vea la ecuación (6) en Ram (1986).] ¿Bajo qué supuestos una variable binaria que indicara si el presidente en el año $t - 1$ es un republicano sería una VI adecuada para $gGOV_t$?

La mecánica de los MC2E es idéntica para los datos de series de tiempo o los de corte transversal, pero para los datos de series de tiempo las propiedades estadísticas de MC2E dependen de la tendencia y las propiedades de correlación de las series subyacentes. En particular, se debe tener cuidado de incluir tendencias si se tienen variables explicativas o dependientes con tendencias. Dado que una tendencia temporal es exógena, siempre sirve como su propia variable instrumental. Lo mismo aplica para las variables binarias estacionales, si se utilizan datos mensuales o trimestrales.

Las series con una fuerte persistencia (que tienen raíces unitarias) se deben emplear con cuidado, tal como con MCO. Con frecuencia, diferenciar la ecuación es indispensable antes de la estimación, y esto aplica a todos los instrumentos por igual.

Con base en los supuestos análogos a los del capítulo 11 para las propiedades asintóticas de MCO, MC2E que utiliza datos de series de tiempo son consistentes y se distribuyen de forma asintóticamente normal. En realidad, si se remplazan las variables explicativas por las variables instrumentales cuando se plantean los supuestos, sólo será necesario agregar los supuestos de identificación para MC2E. Por ejemplo, el supuesto de homocedasticidad se expresa como

$$\mathrm{E}(u_t^2 | z_{t1}, \ldots, z_{tm}) = \sigma^2, \tag{15.53}$$

y el supuesto de no correlación serial se expresa como

$$\mathrm{E}(u_t u_s | \mathbf{z}_t, \mathbf{z}_s) = 0, \quad \text{para toda } t \neq s, \tag{15.54}$$

donde $\mathbf{z}_t$ denota todas las variables exógenas en el tiempo *t*. En el apéndice del capítulo se ofrece un planteamiento completo sobre los supuestos. Se ofrecerán ejemplos de MC2E para problemas de series de tiempo en el capítulo 16; vea también el ejercicio para computadora C15.4.

Como en el caso de MCO, el supuesto de no correlación serial, con frecuencia, puede infringirse con datos de series de tiempo. Por fortuna, es muy fácil probar la correlación serial AR(1). Si se escribe $u_t = \rho u_{t-1} + e_t$ y se inserta esto en la ecuación (15.52), se obtiene

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho u_{t-1} + e_t, t \geq 2. \tag{15.55}$$

Para probar H_0: $\rho_1 = 0$, se debe remplazar u_{t-1} con los residuales de MC2E, $\hat{u}_{t-1}$. Además, si x_{tj} es endógena en (15.52), entonces es endógena en (15.55), así que aún se necesitará emplear una VI. Dado que e_t no está correlacionada con ningún valor pasado de u_t, entonces $\hat{u}_{t-1}$ se puede emplear como su propio instrumento.

Prueba de correlación serial AR(1) después de MC2E:

i) Se estima (15.52) mediante MC2E y se obtienen sus residuales, $\hat{u}_t$.

ii) Se estima

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho\hat{u}_{t-1} + error_t, \quad t = 2, \ldots, n$$

mediante MC2E y los mismos instrumentos de la parte i), además de $\hat{u}_{t-1}$. Se emplea el estadístico t de $\hat{\rho}$ para probar H_0: $\rho = 0$.

Como en la versión de MCO de esta prueba del capítulo 12, el estadístico t sólo tiene una justificación asintótica, pero en la práctica tiende a funcionar bien. Se puede utilizar una versión robusta a la heterocedasticidad para evitar que ésta se presente. Además, los residuales rezagados se pueden agregar a la ecuación para probar formas más elevadas de correlación serial con ayuda de una prueba F conjunta.

¿Qué sucede si se detecta una correlación serial? Algunos paquetes econométricos calcularán errores estándar que son robustos a formas bastante generales de correlación serial y heterocedasticidad. Esta es una forma simple de proceder si los paquetes de econometría hacen esto. Los cálculos son muy similares a los de la sección 12.5 para MCO. [Vea Wooldridge (1995) para las fórmulas y otros métodos computacionales.]

Una alternativa es utilizar el modelo AR(1) y corregir la correlación serial. El procedimiento es similar al de MCO e impone restricciones adicionales a las variables instrumentales. La ecuación cuasi diferenciada es la misma que en la ecuación (12.32):

$$\tilde{y}_t = \beta_0(1 - \rho) + \beta_1\tilde{x}_{t1} + \ldots + \beta_k\tilde{x}_{tk} + e_t, \quad t \geq 2, \qquad \boxed{15.56}$$

donde $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,j}$. (Se puede utilizar la observación $t = 1$ tal como en la sección 12.3, pero se omitió aquí por fines simplificativos.) La pregunta es: ¿Qué se puede emplear como variables instrumentales? Parece natural utilizar los instrumentos cuasi diferenciados, $\tilde{z}_{tj} = z_{tj} - \rho z_{t-1,j}$. Sin embargo, esto sólo funciona si en (15.52) el error original u_t no está correlacionado con los instrumentos en los tiempos t, $t - 1$ y $t + 1$. Es decir, las variables instrumentales deben ser estrictamente exógenas en (15.52). Esto descarta las variables dependientes rezagadas como VI, por ejemplo. También elimina los casos donde los movimientos futuros en las VI reaccionan a cambios actuales y pasados en el error, u_t.

MC2E con errores AR(1):

i) Se estima (15.52) mediante MC2E y se obtienen los residuales de MC2E, $\hat{u}_t$, $t = 1, 2, \ldots, n$.

ii) Se obtiene $\hat{\rho}$ de la regresión de $\hat{u}_t$ sobre $\hat{u}_{t-1}$, $t = 2, \ldots, n$ y se construyen las variables cuasi diferenciadas $\tilde{y}_t = y_t - \hat{\rho}y_{t-1}$, $\tilde{x}_{tj} = x_{tj} - \hat{\rho}x_{t-1,j}$ y $\tilde{z}_{tj} = z_{tj} - \hat{\rho}z_{t-1,j}$ para $t \geq 2$. (Recuerde, en la mayoría de los casos, algunas de las VI serán también variables explicativas.)

iii) Se estima (15.56) (donde ρ se remplaza con $\hat{\rho}$) mediante MC2E y usando las $\tilde{z}_{tj}$ como los instrumentos. Suponiendo que (15.56) satisface los supuestos de MC2E en el apéndice del capítulo, los estadísticos de prueba habituales de MC2E son asintóticamente válidos.

También se puede emplear el primer periodo de tiempo como en la estimación Prais-Winsten del modelo con las variables explicativas exógenas. Las variables transformadas en el primer periodo (la variable dependiente, las variables explicativas y las variables instrumentales) se obtienen simplemente al multiplicar todos los valores del primer periodo por $(1 - \hat{\rho})^{1/2}$. (Vea también la sección 12.3.)

15.8 Aplicación de MC2E a cortes transversales combinados y a datos de panel

Aplicar los métodos de variables instrumentales a la combinación de cortes transversales independientes no plantea nuevas dificultades. Al igual que en los modelos estimados por MCO, con frecuencia, se deben incluir variables binarias de periodo para permitir los efectos temporales agregados. Estas variables binarias son exógenas, debido a que el paso del tiempo también lo es, y también actúan como sus propios instrumentos.

Ejemplo 15.9

[Efecto de la educación sobre la fertilidad]

En el ejemplo 13.1 se utilizó la combinación de cortes transversales en FERTIL1.RAW para estimar el efecto de la educación sobre la fertilidad femenina, controlando otros diferentes factores. Como en Sander (1992), se toma en cuenta la posibilidad de que *educ* sea endógena en la ecuación. Como las variables instrumentales para *educ*, se emplean los niveles de educación del padre y la madre (*meduc*, *feduc*). La estimación de MC2E de β_{educ} es $-.153$ (ee $=$ .039), en comparación con la estimación de MCO $-.128$ (ee $=$ .018). La estimación de MC2E muestra un efecto un poco mayor de la educación sobre la fertilidad, pero el error estándar de MC2E es más de dos veces mayor que el de MCO. (De hecho, el intervalo de confianza a 95% basado en MC2E contiene fácilmente la estimación de MCO.) Las estimaciones de MCO y MC2E de β_{educ} *no son estadísticamente* diferentes, como se puede apreciar al probar la endogeneidad de *educ* como en la sección 15.5: cuando el residual de la forma reducida, $\hat{v}_2$, se incluye con los demás regresores en la tabla 13.1 (incluido *educ*), su estadístico *t* es .702, que no es significativo a ningún nivel razonable. Por tanto, en este caso, se concluye que la diferencia entre MC2E y MCO podría deberse a un error de muestreo.

La estimación de variables instrumentales se puede combinar con métodos de datos de panel, en particular las primeras diferencias, para estimar de forma consistente los parámetros en presencia de los efectos inobservables y la endogeneidad en una o más variables explicativas que varían con el tiempo. El siguiente ejemplo ilustra esta combinación de métodos.

Ejemplo 15.10

[Capacitación laboral y productividad del trabajador]

Suponga que se desea estimar el efecto de una hora de capacitación laboral adicional sobre la productividad del trabajador. Para dos años, 1987 y 1988, considere el modelo simple para datos de panel

$$\log(scrap_{it}) = \beta_0 + \delta_0 d88_t + \beta_1 hrsemp_{it} + a_i + u_{it},\ t = 1, 2,$$

donde $scrap_{it}$ es la tasa de desperdicio en la empresa *i* en el año *t*, y $hrsemp_{it}$ son las horas de capacitación laboral por empleado. Como es costumbre, se permiten diferentes interceptos anuales y un efecto constante e inobservable de la empresa, a_i.

Por las razones analizadas en la sección 13.2, preocuparía que $hrsemp_{it}$ esté correlacionada con a_i, la última de las cuales contiene una capacidad del trabajador no medida. Como antes, se diferencia para eliminar a_i:

$$\Delta\log(scrap_i) = \delta_0 + \beta_1 \Delta hrsemp_i + \Delta u_i. \tag{15.57}$$

Normalmente, se estimaría esta ecuación mediante MCO. Pero, ¿qué sucedería si Δu_i estuviera correlacionada con $\Delta hrsemp_i$? Por ejemplo, una empresa podría contratar más trabajadores capaces, mientras reduce el nivel de capacitación laboral. En este caso, se necesita una variable instrumental para $\Delta hrsemp_i$. Por lo general, tal VI sería difícil de encontrar, pero se puede explotar el hecho de que algunas empresas recibieron subsidios para capacitación laboral en 1988. Si se supone que la designación del subsidio no está correlacionada con Δu_i —algo razonable, pues los subsidios se dieron al principio de 1988—, entonces $\Delta grant_i$ es válida como VI, siempre y cuando $\Delta hrsemp$ y $\Delta grant$ estén correlacionadas. Mediante los datos en JTRAIN.RAW, diferenciados entre 1987 y 1988, la regresión de la primera etapa es

$$\widehat{\Delta hrsemp} = .51 + 27.88\, \Delta grant$$
$$(1.56) \quad (3.13)$$
$$n = 45,\ R^2 = .392.$$

Lo que confirma que el cambio en las horas de capacitación laboral por empleado tiene una fuerte relación positiva con recibir un subsidio en 1988. En realidad, recibir un subsidio para capacitación laboral aumentó la capacitación por empleado por casi 28 horas, y la designación del subsidio dio cuenta de casi 40% de la variación en $\Delta hrsemp$. La estimación de mínimos cuadrados en dos etapas de (15.57) da

$$\widehat{\Delta \log(scrap)} = -.033 - .014\, \Delta hrsemp$$
$$(.127) \quad (.008)$$
$$n = 45,\ R^2 = .016.$$

Esto significa que se estima que 10 horas más de capacitación laboral por trabajador reducirán la tasa de desperdicio 14%. Para las empresas en la muestra, la cantidad promedio de capacitación laboral en 1988 fue de aproximadamente 17 horas por trabajador, con un mínimo de cero y un máximo de 88.

En comparación, la estimación de MCO de (15.57) da $\hat{\beta}_1 = -.0076$ (ee $= .0045$), así que la estimación de MC2E de β_1 es casi el doble en magnitud y es un tanto más significativa en términos estadísticos.

Cuando $T \geq 3$, la ecuación diferenciada puede contener una correlación serial. Se puede utilizar la misma prueba y corrección de correlación serial AR(1) de la sección 15.7, donde todas las regresiones se combinan tanto a lo largo de i como de t. Debido a que no se quiere perder todo un periodo de tiempo, se debe emplear la transformación de Prais-Winsten para el periodo inicial.

Los modelos de efectos inobservables que contienen variables dependientes rezagadas también requieren métodos de VI para una estimación consistente. La razón es que, después de diferenciar, $\Delta y_{i,t-1}$ está correlacionada con Δu_{it} debido a que $y_{i,t-1}$ y $u_{i,t-1}$ están correlacionadas también. Se pueden utilizar dos o más rezagos de y como VI para $\Delta y_{i,t-1}$. [Vea Wooldridge (2002, capítulo 11) para más detalles.]

Después de la diferenciación se pueden emplear variables instrumentales también en muestras apareadas. Ashenfelter y Krueger (1994) diferenciaron la ecuación de salario entre gemelos para eliminar la capacidad inobservable:

$$\log(wage_2) - \log(wage_1) = \delta_0 + \beta_1(educ_{2,2} - educ_{1,1}) + (u_2 - u_1),$$

donde $educ_{1,1}$ son los años de escolaridad para el primer gemelo como lo reportó él mismo, y $educ_{2,2}$ son los años de escolaridad para el segundo gemelo según lo reportó él mismo. Para dar cuenta de un posible error de medición en las medidas de escolaridad que cada quien reportó para sí mismo, Ashenfelter y Krueger emplearon $(educ_{2,1} - educ_{1,2})$ como una VI para $(educ_{2,2} -$

$educ_{1,1}$), donde $educ_{2,1}$ son los años de escolaridad para el segundo gemelo según lo reportado por el primero, y $educ_{1,2}$ son los años de escolaridad para el primer gemelo, según lo reportado por el segundo. La estimación de VI β_1 es .167 (t = 3.88), en comparación con la estimación de MCO sobre las primeras diferencias de .092 (t = 3.83) [vea Ashenfelter y Krueger (1994, tabla 3)].

RESUMEN

En el capítulo 15 se presentó el método de variables instrumentales como una forma de estimar de forma consistente los parámetros de un modelo lineal cuando una o más variables explicativas son endógenas. Una variable instrumental debe tener dos propiedades: 1) debe ser exógena, es decir, no estar correlacionada con el término de error de la ecuación estructural; 2) debe estar correlacionada de forma parcial con la variable explicativa endógena. Hallar una variable con estas dos propiedades suele ser complicado.

El método de mínimos cuadrados en dos etapas, que permite más variables instrumentales que variables explicativas, suele usarse en las ciencias sociales empíricas. Cuando se utiliza como es debido, puede permitir estimar los efectos *ceteris paribus* en presencia de variables explicativas endógenas. Esto es válido en las aplicaciones de datos de corte transversal, series de tiempo y datos de panel. Pero cuando los instrumentos son deficientes, lo cual significa que están correlacionados con el término de error, que sólo está correlacionado débilmente con la variable explicativa endógena, o ambos, entonces MC2E puede ser peor que MCO.

Cuando se tienen variables instrumentales válidas, se puede probar si una variable explicativa es endógena, mediante la prueba en la sección 15.5. Además, aunque nunca se puede probar si todas las VI son exógenas, se puede probar que al menos algunas de ellas lo sean, en el supuesto que se tengan más instrumentos de los necesarios para una estimación consistente (es decir, el modelo está sobreidentificado). Se puede probar y manejar la heterocedasticidad y la correlación serial mediante métodos similares al caso de los modelos con variables explicativas exógenas.

En este capítulo se emplearon las variables omitidas y el error de medición para ilustrar el método de variables instrumentales. Los métodos de VI también son indispensables para los modelos de ecuaciones simultáneas, que se analizarán en el capítulo 16.

TÉRMINOS CLAVE

Condición de orden
Condición de rango
Ecuación en la forma reducida
Ecuación estructural
Errores en variables
Estimador de mínimos cuadrados en dos etapas
Estimador de variables instrumentales (VI)
Exogeneidad de los instrumentos
Experimento natural
Identificación
Instrumentos débiles
Relevancia de los instrumentos
Restricciones de exclusión
Restricciones de sobreidentificación
Variable instrumental
Variables explicativas endógenas
Variables explicativas exógenas
Variables exógenas
Variables omitidas

PROBLEMAS

15.1 Considere un modelo simple para estimar el efecto de la posesión de una computadora personal (PC) en el promedio de calificaciones (GPA) para los alumnos del último año de la carrera en una universidad pública grande:

$$GPA = \beta_0 + \beta_1 PC + u,$$

donde *PC* es una variable binaria que indica la posesión de una PC.

i) ¿Por qué la posesión de una PC podría estar correlacionada con u?

ii) Explique por qué es probable que *PC* esté relacionada con el ingreso anual de los padres. ¿Esto significa que el ingreso de los padres es una buena VI para *PC*? ¿Por qué?

iii) Suponga que, hace cuatro años, la universidad otorgó subsidios para comprar computadoras para casi la mitad de los estudiantes de nuevo ingreso, y se eligió de manera aleatoria a quienes los recibieron. Explique de manera detallada cómo se usaría esta información para construir una variable instrumental para *PC*.

15.2 Suponga que se desea estimar el efecto de la asistencia a clases sobre el desempeño de los estudiantes, como en el ejemplo 6.3. Un modelo básico es

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u,$$

donde las variables se definen como en el capítulo 6.

i) Sea *dist* la distancia de los dormitorios de los estudiantes al aula. ¿Piensa que *dist* no está correlacionada con u?

ii) En el supuesto de que *dist* y u no se correlacionen, ¿qué otros supuestos debe satisfacer *dist* para ser una VI válida para *atndrte*?

iii) Suponga que, como en la ecuación (6.18), se agrega el término de interacción *priGPA·atndrte*:

$$stndfnl = \beta_0 + \beta_1 atn.drte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \cdot atndrte + u.$$

Si *atndrte* está correlacionada con u, entonces, en general, también *priGPA·atndrte*. ¿Cuál sería una buena VI para *priGPA·atndrte*? [*Sugerencia:* si $\mathrm{E}(u|priGPA, ACT, dist) = 0$, como sucede cuando *priGPA*, *ACT*, y *dist* son exógenas, entonces cualquier función de *priGPA* y *dist* no está correlacionada con u.]

15.3 Considere el modelo de regresión simple

$$y = \beta_0 + \beta_1 x + u$$

y sea z una variable instrumental *binaria* para x. Use (15.10) para mostrar que el estimador de VI $\hat{\beta}_1$ se puede escribir como

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0),$$

donde $\bar{y}_0$ y $\bar{x}_0$ son los promedios muestrales de y_i y x_i sobre la parte de la muestra con $z_i = 0$, y donde $\bar{y}_1$ y $\bar{x}_1$ son los promedios muestrales de y_i y x_i sobre la parte de la muestra con $z_i = 1$. Wald (1940) fue el primero en sugerir este estimador, conocido como *estimador de agrupación*.

15.4 Suponga que para un estado determinado en Estados Unidos, se desea utilizar datos de series de tiempo anuales para estimar el efecto del salario mínimo, a nivel estatal, sobre el empleo de las personas entre 18 y 25 años de edad (*EMP*). Un modelo simple es

$$gEMP_t = \beta_0 + \beta_1 gMIN_t + \beta_2 gPOP_t + \beta_3 gGSP_t + \beta_4 gGDP_t + u_t,$$

donde MIN_t es el salario mínimo, en dólares reales, POP_t es la población de 18 a 25 años de edad, GSP_t es el producto bruto del estado y GDP_t es el producto interno bruto estadounidense. El prefijo g indica la tasa de crecimiento del año $t - 1$ al año t, la cual por lo general se aproxima mediante la diferencia en los logaritmos.

i) Si es de preocupar que el estado elija su salario mínimo con base, en parte, en factores inobservables (para nosotros) que afectan el empleo juvenil, ¿cuál es el problema con la estimación de MCO?

ii) Sea $USMIN_t$ el salario mínimo estadounidense, que también se mide en términos reales. ¿Considera que $gUSMIN_t$ no está correlacionada con u_t?

iii) Por ley, el salario mínimo de cualquier estado debe ser al menos el monto del mínimo estadounidense. Explique por qué esto hace que $gUSMIN_t$ sea un candidato potencial como VI para $gMIN_t$.

15.5 Refiérase a las ecuaciones (15.19) y (15.20). Suponga que $\sigma_u = \sigma_x$, de manera que la variación poblacional en el término de error sea la misma que en x. Suponga que la variable instrumental, z, está ligeramente correlacionada con u: Corr(z,u) = .1. Suponga también que z y x tienen una correlación un poco más fuerte: Corr(z,x) = .2.

i) ¿Cuál es el sesgo asintótico en el estimador de VI?

ii) ¿Cuánta correlación tendría que existir entre x y u antes de que MCO tuviera más sesgo asintótico que MC2E?

15.6 i) En el modelo con una variable explicativa endógena, una variable explicativa exógena y una variable exógena adicional, tome la forma reducida para y_2, e insértela en la ecuación estructural (15.22). Esto da la forma reducida de y_1:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1.$$

Calcule las α_j en términos de las β_j y las π_j.

ii) Calcule el error de la forma reducida, v_1, en términos de u_1, v_2 y los parámetros.

iii) ¿Cómo se estimarían consistentemente las a_j?

15.7 El siguiente es un modelo simple para medir el efecto de un programa de elección escolar sobre el desempeño en un examen estandarizado [vea Rouse (1998)]:

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

donde *score* es la puntuación en un examen a nivel estatal, *choice* es una variable binaria que indica si un estudiante asistió a la escuela de su elección el año pasado, y *faminc* es el ingreso familiar. La VI para *choice* es *grant*, la cantidad de dólares otorgada a los estudiantes para matricularse en las escuelas de su elección. La cantidad de la beca difería según el nivel del ingreso familiar, razón por la que se controla *faminc* en la ecuación.

i) Aun con *faminc* en la ecuación, ¿por qué *choice* estaría correlacionada con u_1?

ii) Si dentro de cada clase de ingreso, los montos de las becas se asignaran de manera aleatoria, ¿la beca *grant* está correlacionada con u_1?

iii) Escriba la ecuación en forma reducida para *choice*. ¿Qué se necesita para que *grant* se correlacione parcialmente con *choice*?

iv) Escriba la ecuación en forma reducida para *score*. Explique por qué esto es útil. (*Sugerencia:* ¿Cómo se interpreta el coeficiente de *grant*?)

15.8 Suponga que se desea probar si las jóvenes que asisten al bachillerato femenino tienen un mejor rendimiento en matemáticas que aquellas que asisten a las escuelas mixtas. Se tiene una muestra aleatoria de jóvenes de los últimos años del bachillerato de un estado en Estados Unidos, y *score* es la calificación en un examen de matemáticas estandarizado. Sea *girlhs* una variable binaria que indique si una estudiante asiste a un bachillerato femenil.

i) ¿Qué otros factores se pudieran controlar en la ecuación? (Debe poder recabar datos sobre estos factores.)

ii) Escriba una ecuación que relacione *score* con *girlhs* y los otros factores que se listaron en la parte i).

iii) Suponga que el apoyo y la motivación que ofrecen los padres son factores no medidos en el término de error en la parte ii). ¿Es probable que se correlacionen con *girlhs*? Explique este punto.

iv) Analice los supuestos que se necesitan para que el número de bachilleratos femeniles dentro de un radio de 20 millas de la casa de una joven sea una VI válida para *girlhs*.

15.9 Suponga que, en la ecuación (15.8), no se tiene una buena candidata como variable instrumental para *skipped*. Pero se tienen otras dos piezas de información sobre los estudiantes: puntuación del examen de admisión a la universidad (SAT) y promedio de calificaciones (GPA) antes del semestre. ¿Qué podría hacer en lugar de la estimación de VI?

15.10 En un artículo, Evans y Schwab (1995) estudiaron los efectos de asistir a un bachillerato católico sobre la probabilidad de asistir a la universidad. En concreto, sea *college* una variable binaria igual a la unidad si un estudiante asiste a la universidad y cero en caso contrario. Sea *CathHS* una variable binaria igual a uno si el estudiante asiste a una preparatoria católica. Un modelo de probabilidad lineal es

$$college = \beta_0 + \beta_1 CathHS + otros\ factores + u,$$

donde los otros factores incluyen género, raza, ingreso familiar y educación de los padres.

i) ¿Por qué *CathHS* puede estar correlacionada con u?

ii) Evans y Schwab tienen datos sobre la puntuación de un examen estandarizado que se realizó cuando cada estudiante estaba en el segundo año universitario. ¿Qué se puede hacer con esta variable para mejorar la estimación *ceteris paribus* de asistir a un colegio católico?

iii) Sea *CathRel* una variable binaria igual a uno si el estudiante es católico. Analice los dos requisitos necesarios para que ésta sea una VI válida para *CathHS* en la ecuación anterior. ¿Cuáles de estos requisitos se pueden probar?

iv) No es de sorprender que ser católico tenga un efecto significativo en asistir a un bachillerato católico. ¿Piensa que *CathRel* es un instrumento convincente para *CathHS*?

15.11 Considere un modelo simple de series de tiempo donde la variable explicativa tenga un error de medición clásico:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t^* + u_t \\ x_t &= x_t^* + e_t, \end{aligned} \tag{15.58}$$

donde u_t tiene una media de cero y no está correlacionada con x_t^* ni con e_t. Sólo se observa y_t y x_t. Suponga que e_t tiene una media de cero y que no está correlacionada con x_t^* y que x_t^* también tiene una media de cero (este último supuesto es sólo para simplificar el cálculo algebraico).

i) Escriba $x_t^* = x_t - e_t$ e inserte esto en (15.58). Muestre que el término de error en la nueva ecuación, v_t, está correlacionado negativamente con x_t si $\beta_1 > 0$. ¿Qué implica esto acerca del estimador de MCO de β_1 obtenido de la regresión de y_t sobre x_t?

ii) Además de los supuestos previos, suponga que u_t y e_t no están correlacionados con ninguno de los valores pasados de x_t^* ni de e_t; en particular, ni con x_{t-1}^* ni con e_{t-1}. Muestre que $\mathrm{E}(x_{t-1}v_t) = 0$, donde v_t es el término de error en el modelo de la parte i).

iii) ¿Es probable que x_t y x_{t-1} estén correlacionadas? Explique.

iv) ¿Qué sugieren las partes ii) y iii) como una estrategia útil para estimar de forma consistente β_0 y β_1?

EJERCICIOS EN COMPUTADORA

C15.1 Use los datos en WAGE2.RAW para este ejercicio.

i) En el ejemplo 15.2, usando *sibs* como instrumento para *educ*, la estimación de VI del rendimiento de la educación es .122. Para convencerse de que usar *sibs* como una VI para *educ no* es lo mismo que sólo insertar *sibs* en lugar de *educ* y estimar la regresión por MCO, realice la regresión de log(*wage*) sobre *sibs* y explique sus hallazgos.

ii) La variable *brthord* es el orden de nacimiento (*brthord* es uno para el niño que nació primero, dos para el segundo, y así sucesivamente). Explique por qué *educ* y *brthord* pueden estar negativamente correlacionadas. Realice la regresión de *educ* sobre *brthord* para determinar si existe una correlación negativa estadísticamente significativa.

iii) Use *brthord* como una VI para *educ* en la ecuación (15.1). Reporte e interprete los resultados.

iv) Ahora suponga que se incluye el número de hermanos como una variable explicativa en la ecuación del salario; esto controla los antecedentes familiares, en cierto grado:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u.$$

Suponga que se quiere usar *brthord* como una VI para *educ*, suponiendo que *sibs* sea exógena. La forma reducida para *educ* es

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v.$$

Plantee y pruebe el supuesto de identificación.

v) Estime la ecuación de la parte iv) usando *brthord* como una VI para *educ* (y *sibs* como su propia VI). Comente sobre los errores estándar de $\hat{\beta}_{educ}$ y $\hat{\beta}_{sibs}$.

vi) Utilizando los valores ajustados de la parte iv), $\widehat{educ}$, calcule la correlación entre $\widehat{educ}$ y *sibs*. Use este resultado para explicar sus hallazgos de la parte v).

C15.2 Los datos en FERTIL2.RAW incluyen, para las mujeres en Bostwana durante 1988, información sobre el número de niños (*children*), años de educación (*educ*), edad (*age*) y variables del estatus religioso y económico.

i) Estime el modelo

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

mediante MCO e interprete las estimaciones. En particular, si se mantiene *age* fija, ¿cuál es el efecto estimado de un año más de educación sobre la fertilidad? Si 100 mujeres reciben otro año de educación, ¿cuántos niños menos se espera que tengan?

ii) La variable *frsthalf* es una variable binaria igual a uno si la mujer nació durante los primeros seis meses del año. En caso de que *frsthalf* no esté correlacionada con el término de error de la parte i), muestre que *frsthalf* es una candidata razonable como VI para *educ*. (*Sugerencia:* es necesario realizar una regresión.)

iii) Estime el modelo de la parte i) usado *frsthalf* como una VI para *educ*. Compare el efecto estimado de la educación con la estimación por MCO de la parte i).

iv) Agregue las variables binarias, *electric* (*electricidad*), *tv* (*televisión*) y *bicycle* (*bicicleta*) al modelo y suponga que son exógenas. Estime la ecuación por MCO y MC2E y compare los coeficientes estimados de *educ*. Interprete el coeficiente de *tv* y explique por qué tener televisor tiene un efecto negativo en la fertilidad.

C15.3 Para este ejercicio use los datos en CARD.RAW.

i) La ecuación que se estimó en el ejemplo 15.4 se puede escribir como

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \ldots + u,$$

donde las demás variables explicativas se listan en la tabla 15.1. Con el fin de que VI sean consistentes, la VI para *educ*, *nearc4*, no debe estar correlacionada con *u*. ¿Estaría *nearc4* correlacionada con cuestiones en el término de error, como la capacidad inobservable? Explique esto.

ii) Para una submuestra de hombres en el conjunto de datos, se dispone de las puntuaciones IQ. Realice una regresión de *IQ* sobre *nearc4* para revisar si el promedio de las puntuaciones IQ varía si el hombre se crió cerca de una universidad con programas de cuatro años. ¿Qué se concluye?

iii) Ahora, realice una regresión de *IQ* sobre *nearc4*, *smsa66*, y las variables binarias regionales de 1966, *reg662*, …, *reg669*. ¿Están *IQ* y *nearc4* relacionadas después de haber descontado el efecto parcial de las variables binarias geográficas? Compare esto con sus resultados de la parte ii).

iv) A partir de las partes ii) y iii), ¿qué concluye acerca de la importancia de controlar *smsa66* y las variables binarias regionales de 1966 en la ecuación del log(*wage*)?

C15.4 En este ejercicio use los datos en INTDEF.RAW. Una ecuación simple que relaciona la tasa de bonos del Tesoro a tres meses ($i3_t$) con la tasa de inflación (construida a partir del Índice de Precios al Consumidor) es

$$i3_t = \beta_0 + \beta_1 inf_t + u_t.$$

i) Estime esta ecuación por MCO, omitiendo el primer periodo para comparaciones posteriores. Reporte los resultados como de costumbre.

ii) Algunos economistas consideran que el Índice de Precios al Consumidor no representa la verdadera tasa de inflación, de manera que los MCO en la parte i) sufren de un sesgo de error de medición. Vuelva a estimar la ecuación de la parte i), usando inf_{t-1} como una VI para inf_t. Compare el estimador de VI para β_1 con el de MCO.

iii) Ahora obtenga las primeras diferencias de la ecuación:

$$\Delta i3_t = \beta_0 + \beta_1 \Delta inf_t + \Delta u_t.$$

Estime esto mediante MCO y compare la estimación de β_1 con las estimaciones previas.

iv) ¿Se puede usar Δinf_{t-1} como una VI para Δinf_t en la ecuación diferenciada en la parte iii)? Explique. (*Sugerencia:* ¿Están Δinf_t y Δinf_{t-1} suficientemente correlacionadas?

C15.5 Use los datos de CARD.RAW para este ejercicio.

i) En la tabla 15.1 la diferencia entre las estimaciones de VI y de MCO del rendimiento de la educación es económicamente importante. Obtenga los residuales de la forma reducida, $\hat{v}_2$, de (15.32). (Vea la tabla 15.1 para las otras variables a incluir en la regresión.) Use éstos para probar si *educ* es exógena; es decir, determine si la diferencia entre MCO y VI es *estadísticamente* significativa.

ii) Estime la ecuación por MC2E, agregue *nearc2* como instrumento. ¿El coeficiente de *educ* cambia mucho?

iii) Pruebe la única restricción de sobreidentificación del punto ii).

C15.6 Use los datos en MURDER.RAW para este ejercicio. La variable *mrdrte* es la tasa de homicidios, es decir, el número de homicidios por cada 100,000 personas. La variable *exec* es el número total de prisioneros ejecutados en el año en curso y los últimos dos años; *unem* es la tasa de desempleo estatal.

i) ¿Cuántos estados ejecutaron al menos a un prisionero en 1991, 1992 o 1993? ¿Qué estado (*id*) es el que tiene más ejecuciones?

ii) Usando los dos años 1990 y 1993, efectúe una regresión combinada de *mrdrte* sobre *d93* (variable binaria igual a uno si el año es 1993), *exec* y *unem*. ¿Qué interpretación tiene del coeficiente sobre *exec*?

iii) Tomando los cambios de 1990 a 1993 (para un total de 51 observaciones), estime la ecuación

$$\Delta mrdrte = \delta_0 + \beta_1 \Delta exec + \beta_2 \Delta unem + \Delta u$$

mediante MCO y reporte los resultados en la forma usual. Ahora, ¿la pena capital parece tener un efecto disuasivo?

iv) El cambio en las ejecuciones puede estar al menos parcialmente relacionado con los cambios en el índice esperado de homicidios, así que $\Delta exec$ está correlacionada con Δu en la parte iii). Podría ser razonable suponer que $\Delta exec_{-1}$ no está correlacionada con Δu. (Después de todo, $\Delta exec_{-1}$ depende de las ejecuciones que ocurrieron tres o más años atrás.) Realice una regresión de $\Delta exec$ sobre $\Delta exec_{-1}$ para ver si están suficientemente correlacionadas; interprete el coeficiente de $\Delta exec_{-1}$.

v) Vuelva a estimar la ecuación de la parte iii) usando $\Delta exec_{-1}$ como una VI para $\Delta exec$. Suponga que $\Delta unem$ es exógena. ¿Cómo cambiaron sus conclusiones del punto iii)?

C15.7 Use los datos en PHILLIPS.RAW para este ejercicio.

i) En el ejemplo 11.5 se estimó una curva de Phillips aumentada por expectativas de la forma

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t,$$

donde $\Delta inf_t = inf_t - inf_{t-1}$. Para estimar esta ecuación por MCO, se supuso que el choque de la oferta, e_t, no estaba correlacionado con $unem_t$. Si éste no fuera el caso, ¿qué se puede decir acerca del estimador de MCO de β_1?

ii) Suponga que e_t es impredecible dada toda la información pasada: $E(e_t|inf_{t-1}, unem_{t-1}, \ldots) = 0$. Explique por qué esto hace que $unem_{t-1}$ sea una buena candidata como VI para $unem_t$.

iii) Realice la regresión de $unem_t$ sobre $unem_{t-1}$. ¿Están $unem_t$ y $unem_{t-1}$ correlacionadas de forma significativa?

iv) Estime la curva de Phillips aumentada por expectativas a través de VI. Reporte los resultados de la forma acostumbrada y compárelos con las estimaciones de MCO del ejemplo 11.5.

C15.8 Use los datos en 401KSUBS.RAW para este ejercicio. La ecuación de interés es un modelo de probabilidad lineal:

$$pira = \beta_0 + \beta_1 p401k + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u.$$

La meta es probar si existe un efecto de sustitución entre participar en un plan de pensión 401(k) (*p401k*) y tener una cuenta individual de retiro (*pira*). Por tanto, se quiere estimar β_1.

i) Estime la ecuación mediante MCO y analice el efecto estimado de *p401k*.

ii) Para estimar el efecto de sustitución *ceteris paribus* entre la participación en dos tipos de planes de ahorro para el retiro, ¿cuál podría ser un problema con usar mínimos cuadrados ordinarios?

iii) La variable *e401k* es una variable binaria igual a uno si un trabajador es *elegible* para participar en un plan 401(k). Explique qué se requiere para que *e401k* sea una VI válida para *p401k*. ¿Estos supuestos parecen razonables?

iv) Estime la forma reducida para *p401k* y verifique que *e401k* tenga una correlación parcial significativa con *p401k*. Dado que la forma reducida también es un modelo de probabilidad lineal, use el error estándar robusto a la heterocedasticidad.

v) Ahora, estime la ecuación estructural mediante VI y compare la estimación de β_1 con la estimación de MCO. Una vez más, se deben obtener errores estándar robustos a la heterocedasticidad.

vi) Pruebe la hipótesis nula de que *p401k* es exógena, usando una prueba robusta a la heterocedasticidad.

C15.9 El propósito de este ejercicio es comparar las estimaciones y errores estándar obtenidos al usar correctamente MC2E con los que se obtuvieron usando procedimientos inadecuados. Use el archivo de datos WAGE2.RAW.

i) Emplee una rutina de MC2E para estimar la ecuación

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u,$$

donde *sibs* (número de hermanos) es la VI para *educ*. Reporte los resultados de la forma acostumbrada.

ii) Ahora, manualmente realice MC2E. Es decir, primero efectúe una regresión de $educ_i$ sobre $sibs_i$, $exper_i$, $tenure_i$ y $black_i$ para obtener los valores ajustados $\widehat{educ}_i$, $i = 1, ..., n$. Después, ejecute la regresión de la segunda etapa $\log(wage_i)$ sobre $\widehat{educ}_i$, $exper_i$, $tenure_i$ y $black_i$, $i = 1, ..., n$. Verifique que las $\hat{\beta}_j$ son idénticas a las obtenidas de la parte i), pero que los errores estándar son un tanto diferentes. Los errores estándar obtenidos de la regresión de la segunda etapa, cuando se realiza manualmente MC2E, por lo general son inadecuados.

iii) Ahora use el siguiente procedimiento en dos etapas, que generalmente produce estimaciones inconsistentes de los parámetros β_j, y no sólo errores estándar inconsistentes. En la etapa uno haga una regresión de $educ_i$ sólo sobre $sibs_i$ y obtenga los valore ajustados, $\widetilde{educ}_i$. (Observe que esta es una regresión incorrecta para la primera etapa.) Entonces, en la segunda etapa, realice la regresión de log($wage_i$) sobre $\widetilde{educ}_i$, $exper_i$, $tenure_i$ y $black_i$, $i = 1, \ldots, n$. ¿ Cómo se compara la estimación de este procedimiento incorrecto de dos etapas con la estimación correcta de MC2E del rendimiento de la educación?

C15.10 Use los datos en HTV.RAW para este ejercicio.

i) Realice una simple regresión de MCO de log(*wage*) sobre *educ*. Sin controlar otros factores, ¿cuál es el intervalo de confianza al 95% para el rendimiento de un año adicional de educación?

ii) La variable *ctuit*, es el cambio en la matrícula universitaria, en miles de dólares, que deben pagar los estudiantes al pasar de 17 a 18 años de edad. Demuestre que *educ* y *ctuit* no están correlacionadas. ¿Qué dice esto acerca de *ctuit* como una VI posible para *educ* en un análisis de regresión simple?

iii) Ahora, agregue al modelo de regresión simple en la parte i) una función cuadrática de la experiencia y un conjunto completo de variables binarias regionales para la residencia actual (*nc*, *ne*, *south*, *west*) y la residencia a la edad de 18 (*nc18*, *ne18*, *south18*, west18). También incluya identificadores urbanos para las residencias actuales y de 18 años (*urban*, *urban18*). ¿Cuál es el rendimiento esperado de un año de educación?

iv) Nuevamente, usando *ctuit* como una VI potencial para *educ*, estime la ecuación en la forma reducida de *educ*. [Naturalmente, la forma reducida de *educ* ahora incluye las variables explicativas en la parte iii).] Muestre que *ctuit* es ahora estadísticamente significativa en la forma reducida para *educ*.

v) Estime el modelo de la parte iii) mediante VI, usando *ctuit* como una VI para *educ*. ¿Cómo se compara el intervalo de confianza para el rendimiento de la educación con el de MCO de la parte iii)?

vi) ¿Piensa que el procedimiento de VI en la parte v) sea convincente?

Apéndice 15A

Supuestos de mínimos cuadrados en dos etapas

Este apéndice cubre los supuestos bajo los cuales MC2E tienen propiedades deseables de muestras grandes. Primero se establecen los supuestos para las aplicaciones de corte transversal basados en el muestreo aleatorio. Después, se analizan los que se necesitan agregar con el fin de aplicarse a series de tiempo y datos de panel.

Supuesto MC2E.1 (Lineal en parámetros)

El modelo poblacional se puede escribir como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

donde β_0, β_1, …, β_k son los parámetros (constantes) desconocidos de interés y u es un error aleatorio inobservable o término de perturbación aleatoria. Las variables instrumentales se denotan como z_j.

Vale la pena enfatizar que el supuesto MC2E.1 es prácticamente idéntico a RLM.1 (con la excepción menor de que MC2E.1 menciona la notación para las variables instrumentales, z_j).

En otras palabras, el modelo que aquí interesa es el mismo que el de la estimación por MCO de las β_j. Algunas veces es fácil perder de vista el hecho de que se pueden aplicar diferentes métodos de estimación al mismo modelo. Por desgracia, no es poco común escuchar a los investigadores decir: "Estimé un modelo por MCO" o "Utilicé un modelo con MC2E". Tales expresiones no tienen significado. MCO y MC2E son diferentes métodos de *estimación* que se aplican al *mismo* modelo. Es verdad que tienen propiedades estadísticas deseables bajo diferentes conjuntos de supuestos sobre el modelo, pero la relación que están estimando está dada por la ecuación en MC2E.1 (o RLM.1). El punto es similar al establecido para el modelo de datos de panel de efectos inobservables que se cubrió en los capítulos 13 y 14: MCO combinado, las primeras diferencias, efectos fijos y efectos aleatorios son diferentes métodos de estimación para el mismo modelo.

Supuesto MC2E.2 (Muestreo aleatorio)

Se tiene una muestra aleatoria para y, las x_j y las z_j.

Supuesto MC2E.3 (Condición de rango)

i) No existen relaciones lineales perfectas entre las variables instrumentales. ii) La condición de rango para la identificación se mantiene.

Con una sola variable explicativa endógena, como en la ecuación (15.42), la condición de rango se describe fácilmente. Sean $z_1, \ldots, z_m$ las variables exógenas, donde $z_k, \ldots, z_m$ no aparecen en el modelo estructural (15.42). La forma reducida de y_2 es

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \ldots + \pi_{k-1} z_{k-1} + \pi_k z_k + \ldots + \pi_m z_m + v_2.$$

Después, se necesita que al menos una de $\pi_k, \ldots, \pi_m$ sea diferente de cero. Esto requiere al menos una variable exógena que no aparezca en (15.42) (condición de orden). Expresar la condición de rango con dos o más variables explicativas endógenas requiere de álgebra matricial. [Vea Wooldridge (2002, capítulo 5).]

Supuesto MC2E.4 (Variables instrumentales exógenas)

El término de error u tiene una media de cero, y ninguna VI se correlaciona con u.

Recuerde que toda x_j que no está correlacionada con u también actúa como una VI.

Teorema 15A.1

Bajo los supuestos MC2E.1 a MC2E.4, el estimador de MC2E es consistente.

Supuesto MC2E.5 (Homocedasticidad)

Sea **z** el conjunto de todas las variables instrumentales. Entonces, $E(u^2|\mathbf{z}) = \sigma^2$.

Teorema 15A.2

Bajo los supuestos MC2E.1 a MC2E.5, los estimadores de MC2E se distribuyen asintóticamente de manera normal. Los estimadores consistentes de la varianza asintótica se dan como en la ecuación (15.43), donde σ^2 se remplaza con $\hat{\sigma}^2 = (n - k - 1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2$, y las $\hat{u}_i$ son los residuales de MC2E.

El estimador de MC2E también es el mejor estimador de VI bajo los cinco supuestos dados. Aquí se enuncia el resultado. Una prueba se puede encontrar en Wooldridge (2002, capítulo 5).

Teorema 15A.3

Bajo los supuestos MC2E.1 a MC2E.5, el estimador de MC2E es asintóticamente eficiente en la clase de los estimadores de VI que emplean combinaciones lineales de las variables exógenas como instrumentos.

Si el supuesto de homocedasticidad no es válido, los estimadores de MC2E seguirán siendo asintóticamente normales, pero los errores estándar (y los estadísticos t y F) necesitarán ajustarse; varios paquetes econométricos realizan esto de manera rutinaria. Además, el estimador de MC2E ya no es el estimador asintóticamente eficiente de VI, en general. No se estudiarán estimadores más eficientes aquí [vea Wooldridge (2002, capítulo 8)].

Para aplicaciones de series de tiempo se deben agregar algunos supuestos. Primero, como con MCO, se debe suponer que todas las series (incluidas las VI) son débilmente dependientes: esto asegura que la ley de los grandes números y el teorema del límite central sean válidos. Para que los errores estándar y los estadísticos de prueba usuales sean válidos, así como para la eficiencia asintótica, se debe agregar un supuesto de no correlación serial.

Supuesto MC2E.6 (No correlación serial)

La ecuación (15.54) se mantiene.

Un supuesto similar de no correlación serial se necesita en las aplicaciones de datos de panel. Las pruebas y correcciones para la correlación serial se analizaron en la sección 15.7.

CAPÍTULO 16

Modelos de ecuaciones simultáneas

En el capítulo anterior se mostró que el método de variables instrumentales puede resolver dos tipos de problemas: las variables omitidas y el error de medición. En términos conceptuales, estos problemas son sencillos. En el caso de las variables omitidas, existe una variable (o más de una) que se desearía mantener fija al estimar el efecto *ceteris paribus* de una o más de las variables explicativas observadas. En el caso del error de medición, lo que interesa estimar es el efecto de ciertas variables explicativas sobre y, pero que se midieron una o más variables de manera equivocada. En ambos casos, se podría estimar el parámetro de interés mediante MCO si fuera posible recabar mejores datos.

Otra forma importante de endogeneidad de las variables explicativas es la **simultaneidad**. Ésta surge cuando una o más de las variables explicativas se *determina conjuntamente* con la variable dependiente, por lo general, a través de un mecanismo de equilibrio (como se verá más adelante). En este capítulo se estudiarán los métodos para estimar los modelos de ecuaciones simultáneas simples (MES). Aunque un análisis completo de este tema rebasa el alcance de este libro, se estudiarán los modelos más conocidos.

El método más importante para estimar modelos de ecuaciones simultáneas es el de variables instrumentales. Por tanto, la solución para el problema de simultaneidad en esencia, es el mismo que las soluciones de VI a problemas de variables omitidas y problemas de error de medición. Sin embargo, la interpretación de los MES es desafiante. Así, se comenzará con el análisis de la naturaleza y alcance de los modelos de ecuaciones simultáneas en la sección 16.1. En la sección 16.2 se confirma que MCO aplicado a una ecuación en un sistema simultáneo es, por lo general, sesgado e inconsistente.

La sección 16.3 ofrece una descripción general de la identificación y estimación en un sistema de dos ecuaciones, mientras que en la sección 16.4 se cubren modelos con más de dos ecuaciones. Los modelos de ecuaciones simultáneas se emplean para representar series de tiempo agregadas y en la sección 16.5 se incluye un análisis de algunas cuestiones especiales que surgen de tales modelos. La sección 16.6 cubre los modelos de ecuaciones simultáneas con datos de panel.

16.1 Naturaleza de los modelos de ecuaciones simultáneas

El punto más importante para recordar cuando se utilizan modelos de ecuaciones simultáneas es que cada ecuación en el sistema debe tener una interpretación causal, *ceteris paribus*. Ya que sólo se observan los resultados en equilibrio, se requiere que se emplee el razonamiento hipotético al construir ecuaciones de un modelo de ecuaciones simultáneas. Es necesario pensar en términos de resultados potenciales así como reales.

El ejemplo clásico de un MES es una ecuación de oferta y demanda para algunas mercancías o insumos de producción (como mano de obra). Concretamente, sea h_s las horas anuales de mano de obra proporcionada por campesinos, medida a nivel nacional, y w el salario por hora promedio que se ofrece a tales trabajadores. Una función simple de oferta de mano de obra es

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1, \tag{16.1}$$

donde z_1 es alguna variable observable que afecta la oferta de mano de obra, por ejemplo, el salario manufacturero promedio en la demarcación. El término de error, u_1, contiene otros factores que afectan a la mano de obra. [Muchos de estos factores son observables y podrían incluirse en la ecuación (16.1); para ilustrar los conceptos básicos, se incluye uno de tales factores, z_1.] La ecuación (16.1) es un ejemplo de **ecuación estructural**. Este nombre proviene del hecho de que la función de mano de obra es derivable de la teoría económica y tiene una interpretación causal. El coeficiente α_1 mide cómo cambia la oferta de mano de obra cuando el salario también cambia; si h_s y w están en forma logarítmica, α_1 es la elasticidad de la oferta de mano de obra. Por lo general, se espera que α_1 sea positiva (aunque la teoría económica no descarta $\alpha_1 \leq 0$). Las elasticidades de la mano de obra son importantes para determinar cómo cambiará el número de horas que los trabajadores desean trabajar cuando las tasas impositivas sobre el ingreso salarial cambien. Si z_1 es el salario manufacturero, se espera que $\beta_1 \leq 0$: todos los demás factores iguales, si el salario de manufactura aumenta, más trabajadores se dedicarán a la manufactura que a la agricultura.

Cuando se grafica la oferta de mano de obra, se trazan las horas en función del salario, con z_1 y u_1 constantes. Un cambio en z_1 desplaza la función de mano de obra, como lo hace un cambio en u_1. La diferencia es que z_1 es observable, mientras que u_1 no. Algunas veces, z_1 recibe el nombre de *desplazador observable de la oferta*, y u_1 recibe el nombre de *desplazador inobservable de la oferta*.

¿En qué difiere la ecuación (16.1) de las que se han estudiado antes? La diferencia es sutil. A pesar de que se supone que la ecuación (16.1) sea válida para todos los valores salariales posibles, el salario no puede estar variando de manera exógena a través de un corte transversal de demarcaciones. Si se pudiera realizar un experimento donde se variara el nivel de salarios agrícolas y manufactureros a través de una muestra de demarcaciones y se encuestara a los trabajadores para obtener la oferta de mano de obra h_s para cada demarcación, entonces se podría estimar (16.1) mediante MCO. Por desgracia, éste no es un experimento realizable; en lugar de esto, se recaban datos sobre los salarios promedio en estos dos sectores además de cuántas horas hombre se invierten en la producción agrícola. Al decidir cómo analizar estos datos, se debe comprender que se describen mejor por la interacción de la oferta *y* la demanda de mano de obra. De acuerdo con el supuesto de la igualación de la oferta y la demanda en los mercados de mano de obra, los valores de los salarios y las horas trabajadas están en *equilibrio*.

Para describir cómo se determinan los salarios y las horas en equilibrio, es necesario introducir la demanda de mano de obra, la cual se supone está dada por

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2, \tag{16.2}$$

donde h_d son las horas demandadas. Como con la función de la oferta, se grafican las horas demandadas en función del salario, w, y z_2 y u_2 se mantienen fijas. La variable z_2 —digamos, la extensión de tierra cultivable— es un *desplazador observable de la demanda*, mientras que u_2 es un *desplazador inobservable de la demanda*.

Tal como con la ecuación de la oferta de mano de obra, la de la demanda es una ecuación estructural, pues se puede obtener de las consideraciones de maximización de utilidades de los agricultores. Si h_d y w están en forma logarítmica, α_2 es la elasticidad de la demanda de mano de obra. La teoría económica dice que $\alpha_2 < 0$. Dado que la mano de obra y la tierra son complementos en la producción, se espera que $\beta_2 > 0$.

Observe cómo las ecuaciones (16.1) y (16.2) describen relaciones completamente diferentes. La oferta de mano de obra es una ecuación del comportamiento de los trabajadores y la de la demanda lo es del comportamiento de los patrones. Cada ecuación tiene una interpretación *ceteris paribus* y se sostiene por sí misma. En un análisis econométrico, ambas se vinculan sólo debido a que el salario y las horas *observados* se determinan por el intercepto de la oferta y la demanda. En otras palabras, para cada país i, las horas observadas h_i y el salario observado w_i están determinados por la condición de equilibrio

$$h_{is} = h_{id}. \tag{16.3}$$

Debido a que se observan sólo horas de equilibrio en cada demarcación i, las horas observadas se denotan con h_i.

Cuando se combina la condición de equilibrio en (16.3) con las ecuaciones de oferta y demanda, se obtiene

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1} \tag{16.4}$$

y

$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}, \tag{16.5}$$

donde explícitamente se incluye el subíndice i para enfatizar que h_i y w_i son los valores de equilibrio observados en cada demarcación i. Estas dos ecuaciones constituyen un **modelo de ecuaciones simultáneas (MES)**, el cual tiene muchas características importantes. Primero, dados z_{i1}, z_{i2}, u_{i1} y u_{i2}, estas dos ecuaciones determinan h_i y w_i. (En realidad, se debe suponer que $\alpha_1 \neq \alpha_2$, lo cual significa que las pendientes de las funciones de la oferta y la demanda difieren; vea el problema 16.1.) Por esta razón, h_i y w_i son las **variables endógenas** en este MES. ¿Qué hay de z_{i1} y z_{i2}? Debido a que se determinan fuera del modelo, se consideran como **variables exógenas**. Desde un punto de vista estadístico el supuesto clave concerniente a z_{i1} y z_{i2} es que ninguno está correlacionado con los errores de la oferta y la demanda, u_{i1} y u_{i2}, respectivamente. Éstos son ejemplos de **errores estructurales**, debido a que aparecen en ecuaciones estructurales.

Un segundo punto importante es que, sin incluir z_1 ni z_2 en el modelo, no hay forma de determinar qué ecuación es la función de la oferta y cuál la de demanda. Cuando z_1 representa el salario manufacturero, el razonamiento económico afirma que es un factor de la oferta de mano de obra agrícola, debido a que mide el costo de oportunidad de trabajar en la agricultura; cuando z_2 representa el área de tierras agrícolas, la teoría de la producción implica que aparece en la función de la demanda de mano obra. Por tanto, se sabe que (16.4) representa la mano de obra y (16.5) representa la demanda. Si z_1 y z_2 son iguales (por ejemplo, el nivel promedio de educación de adultos en la demarcación que puede afectar tanto a la oferta como a la demanda), entonces las ecuaciones parecerán idénticas, y no será posible estimar ninguna. En síntesis, esto ilustra el problema de identificación en modelos de ecuaciones simultáneas, que se analizará con mayor detalle en la sección 16.3.

Los ejemplos más convincentes de MES comparten la misma esencia de los ejemplos de oferta y demanda. Cada ecuación, en sí misma, debe ser una interpretación *ceteris paribus* del comportamiento. Como sólo se observan resultados en equilibrio, especificar un MES requiere que se planteen preguntas hipotéticas como: ¿cuánta mano de obra *proporcionarían* los trabajadores si el salario fuera diferente de su valor de equilibrio? El ejemplo 16.1 ofrece otra ilustración de un MES donde cada ecuación tiene una interpretación *ceteris paribus*.

Ejemplo 16.1

[Índice de homicidios y tamaño de la fuerza policiaca]

Las ciudades suelen precisar cuánta más fuerza policiaca necesitan para disminuir sus índices de homicidios. Un modelo simple de corte transversal para abordar esta cuestión es

$$murdpc = \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1, \quad \mathbf{16.6}$$

Donde *murdpc* son los homicidios, *polpc* es el número de oficiales de policía e *incpc* es el ingreso, todos ellos *per cápita*. (Por tanto, no se incluye un subíndice *i*.) El ingreso *per cápita* se considera exógeno en esta ecuación. En la práctica, se incluirían otros factores, como distribuciones de edad y sexo, niveles educativos, quizá variables geográficas y otras variables que midieran la severidad del castigo. Para esclarecer este punto, considere la ecuación (16. 6).

La pregunta que se desea contestar es: si una ciudad incrementa, de manera exógena, su fuerza policiaca, ¿ese incremento hará descender el promedio del índice de homicidios? Si se pudiera elegir de manera exógena la cantidad de fuerza policiaca para una muestra aleatoria de ciudades, se estimaría (16.16) mediante MCO. Por supuesto, tal experimento no se puede realizar. Pero, de cualquier forma, ¿es posible pensar que la cantidad de fuerza policiaca se determina de manera exógena? Probablemente no. El gasto de una ciudad en las fuerzas del orden está determinado, al menos parcialmente, por su índice de homicidios esperado. Para reflejar esto, se postula una segunda relación:

$$polpc = \alpha_2 murdpc + \beta_{20} + otros\ factores. \quad \mathbf{16.7}$$

Se espera que $\alpha_2 > 0$: manteniendo los otros factores constantes, las ciudades con índices de homicidios (esperados) más altos tengan más policías *per cápita*. Una vez que se especifican los demás factores en (16.7), se tiene un modelo de dos ecuaciones simultáneas. En realidad sólo interesa la ecuación (16.6) pero, como se verá en la sección 16.3, es necesario saber con precisión cómo se especifica la segunda ecuación con el fin de estimar la primera.

Un punto importante es que (16.7) describe el comportamiento de los policías, mientras que (16.6) describe las acciones de los potenciales homicidas. Esto da a cada ecuación una interpretación *ceteris paribus* clara, lo cual hace que las ecuaciones (16.6) y (16.7) sean un modelo de ecuaciones simultáneas apropiado.

A continuación se presenta un ejemplo de un uso inadecuado de MES.

Ejemplo 16.2

[Gasto y ahorro en vivienda]

Suponga que una familia aleatoria en la población tiene gastos y ahorros en vivienda anuales que están determinados de manera conjunta por

$$housing = \alpha_1 saving + \beta_{10} + \beta_{11} inc + \beta_{12} educ + \beta_{13} age + u_1 \quad \mathbf{16.8}$$

y

$$saving = \alpha_2 housing + \beta_{20} + \beta_{21} inc + \beta_{22} educ + \beta_{23} age + u_2, \quad \mathbf{16.9}$$

donde *inc* es el ingreso anual y *educ* y *age* se miden en años. En un principio puede parecer que estas ecuaciones son una forma sensible de ver cómo se determinan los gastos y ahorros en vivienda. Pero es necesario preguntar, ¿qué valor tendría una de estas ecuaciones sin el otro? Ninguna tiene una interpretación *ceteris paribus* debido a que *housing* (vivienda) y *saving* (ahorro) se eligieron para la misma familia. Por ejemplo, no tiene sentido preguntar: si el ingreso anual aumenta en $10,000, ¿cómo cambiarían los gastos de la vivienda, si *saving se mantuviera fijo*? Si el ingreso familiar aumenta, una familia, por lo general, cambiará la mezcla óptima de gastos y ahorro en vivienda. Pero la ecuación (16.8) hace parecer como si se quisiera conocer el efecto de un cambio en *inc*, *educ* o *age* mientras se mantiene fijo *saving*. Este difícil experimento no es interesante. Cualquier modelo basado en principios económicos, en particular, la maximización de utilidades, haría que las familias eligieran de manera óptima *housing* y *saving* como funciones de *inc* y los precios relativos de la vivienda y el ahorro. Las variables *educ* y *age* afectarían las preferencias de consumo, ahorro y riesgo. Por tanto, *housing* y *saving* serían cada una funciones del ingreso, educación, edad y otras variables que afectan al problema de maximización de utilidades (como tasas de rendimiento diferentes sobre vivienda y otros ahorros).

Incluso si se decidiera que MES en (16.8) y (16.9) tienen sentido, no hay forma de estimar los parámetros. (Se analizará este problema más ampliamente en la sección 16.3.) Las dos ecuaciones son indistintas, a menos que se suponga que el ingreso, la educación o la edad aparecen en una ecuación pero no en la otra, lo cual no tendría sentido.

Aunque este es un ejemplo pobre del MES, lo que podría resultar interesante es probar si, en otros factores fijos, existe una compensación entre los gastos de vivienda y ahorro. Pero entonces sólo se estaría estimando (16.8) mediante MCO, a menos que existiera una variable omitida o problema de error de medición.

El ejemplo 16.2 tiene las características de todas las numerosas aplicaciones de MES. El problema es que el mismo agente económico es el que elige a las dos variables endógenas. Por tanto, ninguna ecuación puede sostenerse por sí misma. Otro ejemplo de uso inadecuado de un MES sería modelar las horas semanales que se pasan estudiando y las que se pasan trabajando. Cada estudiante elegirá estas variables de forma simultánea, quizás en función del salario que pueda ganar trabajando, su capacidad como estudiante, su entusiasmo por la universidad, etc. Tal como en el ejemplo 16.2, no tiene sentido especificar dos ecuaciones que están en función una de la otra. La lección importante de esto es: sólo debido a que dos variables están determinadas de forma simultánea *no* significa que un modelo de ecuaciones simultáneas sea idóneo. Para que un MES tenga sentido, cada una de sus ecuaciones debe tener una interpretación *ceteris paribus* independientemente de la otra. Como se analizó antes, los ejemplos de la oferta y la demanda y el ejemplo 16.1 tienen esta característica. Por lo general, el razonamiento económico básico, apoyado en algunos casos por modelos económicos simples, puede ser de ayuda para utilizar los MES de manera inteligente (además de saber cuándo no emplearlos).

Pregunta 16.1

Pindyck y Rubinfeld (1992, sección 11.6) describen un modelo de publicidad donde las empresas monopólicas elijen niveles de maximización de utilidades de precio y gastos publicitarios. ¿Esto significa que se deba emplear un MES para modelar estas variables a nivel de la empresa?

16.2 Sesgo de simultaneidad en MCO

Es útil ver, en un modelo simple, que una variable explicativa que está determinada de forma simultánea con la variable dependiente por lo general, está correlacionada con el término de error, lo cual genera un sesgo e inconsistencia en MCO. Considere el modelo estructural de dos ecuaciones

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \qquad \textbf{16.10}$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \qquad \textbf{16.11}$$

y enfocarse en estimar la primera ecuación. Las variables z_1 y z_2 son exógenas, de manera que ninguna de ellas se correlaciona con u_1 y u_2. Por simplicidad, se suprime el intercepto en cada ecuación.

Para mostrar que y_2 generalmente se correlaciona con u_1, se resuelven las dos ecuaciones para y_2 en términos de las variables exógenas y el término de error. Si se inserta el lado derecho de (16.10) para y_1, en (16.11) se obtiene

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

o

$$(1 - \alpha_2\alpha_1)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2. \quad \boxed{16.12}$$

Ahora, se debe plantear un supuesto acerca de los parámetros con el fin de resolver para y_2:

$$\alpha_2\alpha_1 \neq 1. \quad \boxed{16.13}$$

Que este supuesto sea restrictivo dependerá de la aplicación. En el ejemplo 16.1 se piensa que $\alpha_1 \leq 0$ y $\alpha_2 \geq 0$, lo cual implica que $\alpha_1\alpha_2 \leq 0$; por tanto (16.13) es bastante razonable para el ejemplo 16.1.

Si la condición (16.13) se mantiene, se puede dividir (16.12) entre $(1 - \alpha_2\alpha_1)$ y se escribe y_2 como

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2, \quad \boxed{16.14}$$

donde $\pi_{21} = \alpha_2\beta_1/(1 - \alpha_2\alpha_1)$, $\pi_{22} = \beta_2/(1 - \alpha_2\alpha_1)$ y $v_2 = (\alpha_2 u_1 + u_2)/(1 - \alpha_2\alpha_1)$. La ecuación (16.14), que expresa y_2 en términos de las variables exógenas y los términos de error, es la **ecuación en la forma reducida** para y_2, un concepto que se presentó en el capítulo 15 en el contexto de la estimación de variables instrumentales. Los parámetros π_{21} y π_{22} reciben el nombre de **parámetros de la forma reducida**; observe que son funciones no lineales de los **parámetros estructurales** que aparecen en las ecuaciones estructurales (16.10) y (16.11).

El **error de la forma reducida**, v_2, es una función lineal de los términos de error estructurales, u_1 y u_2. Como ni u_1 ni u_2 están correlacionados con z_1 ni z_2, v_2 tampoco está correlacionada con z_1 ni z_2. Por tanto, se puede estimar de forma consistente π_{21} y π_{22} mediante MCO, el cual se emplea para estimar mínimos cuadrados en dos etapas (cuestión que se abordará en la siguiente sección). Además, en ocasiones, los parámetros de la forma reducida son de interés directo, aunque el interés aquí resida en estimar la ecuación (16.10).

También existe una forma reducida para y_1 bajo el supuesto (16.13); el álgebra es similar a la que se emplea para obtener (16.14). Tiene las mismas propiedades que la ecuación en la forma reducida para y_2.

Se puede emplear la ecuación (16.14) para mostrar que, salvo en supuestos especiales, la estimación de la ecuación (16.10) mediante MCO producirá estimadores inconsistentes y sesgados de α_1 y β_1 en la ecuación (16.10). Ya que z_1 y u_1 no están correlacionadas mediante el supuesto, la cuestión es saber si y_2 y u_1 no están correlacionadas. A partir de la forma reducida en (16.14), se observa que y_2 y u_1 están correlacionadas si, y sólo si, v_2 y u_1 están correlacionadas (porque se supone que z_1 y z_2 son exógenos). Pero v_2 es una función lineal de u_1 y u_2, de manera que suelen estar correlacionados con u_1. De hecho, si se supone que u_1 y u_2 no están correlacionados, entonces v_2 y u_1 *deben* estar correlacionados siempre que $\alpha_2 \neq 0$. Incluso si α_2 fuera igual a cero, lo cual significaría que y_1 no aparece en la ecuación (16.11); v_2 y u_1 se correlacionarán si u_1 y u_2 están correlacionados.

Cuando $\alpha_2 = 0$ y u_1 y u_2 no están correlacionados, y_2 y u_1 tampoco están correlacionados. Éstos son requisitos muy fuertes: si $\alpha_2 = 0$, y_2 no se determina simultáneamente con y_1. Si se agrega una correlación cero entre u_1 y u_2, esto descarta las variables omitidas o los errores de medición en u_1 que estén correlacionados con y_2. No es de sorprender que la estimación de MCO de la ecuación (16.10) funcione en este caso.

Cuando y_2 está correlacionada con u_1 debido a la simultaneidad, se dice que MCO sufre de **sesgo de simultaneidad**. Obtener la dirección del sesgo en los coeficientes por lo general es complicado, como se vio con el sesgo de las variables omitidas en los capítulos 3 y 5. Pero en modelos simples, esto puede hacerse. Por ejemplo, suponga que la ecuación (16.10) se simplifica al descartar z_1 de la ecuación, y se supone que u_1 y u_2 no están correlacionados. Entonces, la covarianza entre y_2 y u_1 es

$$\begin{aligned}\mathrm{Cov}(y_2,u_1) &= \mathrm{Cov}(v_2,u_1) = [\alpha_2/(1-\alpha_2\alpha_1)]\mathrm{E}(u_1^2)\\ &= [\alpha_2/(1-\alpha_2\alpha_1)]\sigma_1^2,\end{aligned}$$

donde $\sigma_1^2 = \mathrm{Var}(u_1) > 0$. Por tanto, el sesgo asintótico (o inconsistencia) en el estimador de MCO de α_1 tiene el mismo signo que el de $\alpha_2/(1-\alpha_2\alpha_1)$. Si $\alpha_2 > 0$ y $\alpha_2\alpha_1 < 1$, el sesgo asintótico es positivo. (Por desgracia, tal como en el cálculo del sesgo en variables omitidas de la sección 3.3, las conclusiones no se pueden extender a modelos más generales, pero sirven como una guía.) Como ilustración, en el ejemplo 16.1, se piensa que $\alpha_2 > 0$ y $\alpha_2\alpha_1 \leq 0$, lo cual significa que el estimador de MCO de α_1 tendría un sesgo positivo. Si $\alpha_1 = 0$, MCO estimarían, en promedio, un impacto *positivo* de más policías sobre el índice de homicidios; en general, el estimador de α_1 está sesgado hacia arriba. Debido a que se espera que el incremento en el número de policías tienda a reducir los índices de homicidios (*ceteris paribus*), el sesgo hacia arriba significa que MCO subestimará la eficacia de una fuerza policiaca mayor.

16.3 Identificar y estimar una ecuación estructural

Como se vio en la sección anterior, MCO es sesgado e inconsistente cuando se aplica a una ecuación estructural en un sistema de ecuaciones simultáneas. En el capítulo 15 se explicó que el método de mínimos cuadrados en dos etapas se puede utilizar para resolver el problema de las variables explicativas endógenas. Ahora se muestra cómo aplicar MC2E a MES.

La mecánica de MC2E es similar a la que se plantea en el capítulo 15. La diferencia es que, debido a que se especifica una ecuación estructural para cada variable endógena, puede verse de inmediato si hay suficientes VI para estimar cualquier ecuación. Se comienza por analizar el problema de identificación.

Identificación en un sistema de dos ecuaciones

En el capítulo 15 se mencionó la noción de identificación. Cuando se estima un modelo mediante MCO, la condición clave de identificación es que ninguna variable explicativa esté correlacionada con el término de error. Como se demostró en la sección 16.2, esta condición fundamental ya no aplica para los MES en general. No obstante, si se tienen algunas variables instrumentales, aún se pueden identificar (o estimar de manera consistente) los parámetros en una ecuación de MES, tal como con las variables omitidas o el error de medición.

Antes de considerar un MES general de dos ecuaciones, es útil considerar un simple ejemplo de la oferta y la demanda para obtener una mayor comprensión. Escriba el sistema en su forma de equilibrio (que es, con $q_s = q_d = q$ impuesta) como

$$q = \alpha_1 p + \beta_1 z_1 + u_1 \tag{16.15}$$

y

$$q = \alpha_2 p + u_2. \tag{16.16}$$

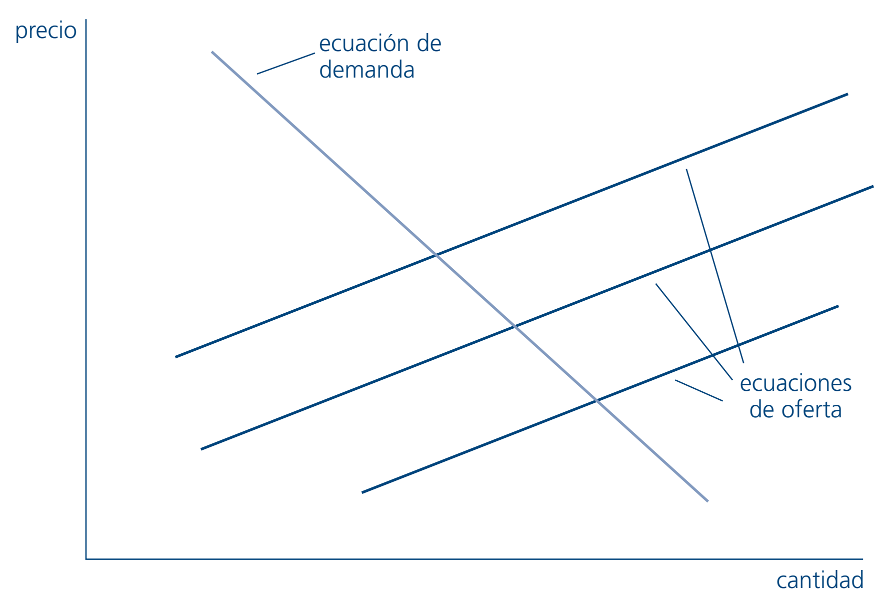
Concretamente, sea q el consumo *per cápita* de leche a nivel nacional, sea p el precio promedio por galón de leche en la demarcación y sea z_1 el precio del forraje para ganado, que se supone exógeno a las ecuaciones de la oferta y la demanda para la leche. Esto significa que (16.15) debe ser la función de la oferta, puesto que el precio del forraje para ganado desplazaría a la oferta ($\beta_1 < 0$) pero no a la demanda. La función de la demanda no contiene desplazadores observables.

Dada una muestra aleatoria sobre (q, p, z_1), ¿cuál de estas ecuaciones puede estimarse? Es decir, ¿cuál es una **ecuación identificada**? Resulta que la ecuación de la *demanda*, (16.16), está identificada, pero la ecuación de la oferta, no. Esto puede verse con facilidad mediante las reglas para la estimación de VI del capítulo 15: se puede emplear z_1 como VI para el precio en la ecuación (16.16). Sin embargo, debido a que z_1 aparece en la ecuación (16.15), no se tiene VI alguna para el precio en la ecuación de la oferta.

Intuitivamente, el hecho de que se identifique la ecuación de la demanda se desprende del hecho de tener una variable observable, z_1, que desplaza a la ecuación de la oferta mientras que no afecta a la ecuación de la demanda. Dada la variación en z_1 y la falta de errores, se puede trazar la curva de demanda, como se muestra en la figura 16.1. La presencia de un desplazador inobservable de la demanda u_2 ocasiona que se estime la ecuación de demanda con error, pero los estimadores serán consistentes, siempre y cuando z_1 no esté correlacionado con u_2.

FIGURA 16.1

Al desplazar las ecuaciones de oferta se traza la ecuación de demanda. Cada ecuación se traza para un valor diferente de la variable exógena, z_1.



La ecuación de oferta no se puede trazar, pues no existen factores exógenos observables que desplacen la curva de demanda. Que existan factores inobservables que hagan esto no ayuda; se necesita algo observable. Si, como en la función de la demanda de mano de obra (16.2), se tiene un desplazador observable de la demanda exógeno, como el ingreso en la función de la demanda de la leche, entonces también se podría identificar la función de la oferta.

Para resumir: *en el sistema de (16.15) y (16.16), es la presencia de una variable exógena en la ecuación de oferta lo que permite estimar la ecuación de demanda.*

Extender el análisis de identificación a un modelo general de dos ecuaciones no es difícil; las dos ecuaciones se escriben

$$y_1 = \beta_{10} + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1 \tag{16.17}$$

y

$$y_2 = \beta_{20} + \alpha_2 y_1 + \mathbf{z}_2\boldsymbol{\beta}_2 + u_2, \tag{16.18}$$

donde y_1 y y_2 son variables endógenas, y u_1 y u_2 son términos de error estructurales. El intercepto en la primera ecuación es β_{10}, y en la segunda es β_{20}. La variable $\mathbf{z}_1$ denota un conjunto de k_1 variables exógenas que aparecen en la primera ecuación: $\mathbf{z}_1 = (z_{11}, z_{12}, \ldots, z_{1k_1})$. Asimismo, $\mathbf{z}_2$ es el conjunto de k_2 variables exógenas en la segunda ecuación: $\mathbf{z}_2 = (z_{21}, z_{22}, \ldots, z_{2k_2})$. En muchos casos, $\mathbf{z}_1$ y $\mathbf{z}_2$ se traslaparán. En forma abreviada, se empleará la notación

$$\mathbf{z}_1\boldsymbol{\beta}_1 = \beta_{11}z_{11} + \beta_{12}z_{12} + \ldots + \beta_{1k_1}z_{1k_1}$$

y

$$\mathbf{z}_2\boldsymbol{\beta}_2 = \beta_{21}z_{21} + \beta_{22}z_{22} + \ldots + \beta_{2k_2}z_{2k_2};$$

es decir, $\mathbf{z}_1\boldsymbol{\beta}_1$ representa todas las variables exógenas en la primera ecuación, con cada una multiplicada por un coeficiente, y de manera similar para $\mathbf{z}_2\boldsymbol{\beta}_2$. (Algunos autores usan, en cambio, la notación $\mathbf{z}_1'\boldsymbol{\beta}_1$ y $\mathbf{z}_2'\boldsymbol{\beta}_2$. Si se está interesado en el método de álgebra matricial en econometría, vea el apéndice E.)

El hecho de que $\mathbf{z}_1$ y $\mathbf{z}_2$ contengan, por lo general, diferentes variables exógenas significa que se han impuesto **restricciones de exclusión** en el modelo. En otras palabras, se *supone* que ciertas variables exógenas no aparecen en la primera ecuación y que otras están ausentes en la segunda. Como se pudo apreciar en los ejemplos anteriores de oferta y demanda, esto permite distinguir entre las dos ecuaciones estructurales.

¿Se pueden resolver las ecuaciones (16.17) y (16.18) para y_1 y y_2 (como funciones lineales de todas las variables exógenas y los errores estructurales, u_1 y u_2)? La condición es la misma que en (16.13), a saber, $\alpha_2\alpha_1 \neq 1$. La prueba es prácticamente idéntica al modelo simple en la sección 16.2. Con base en este supuesto, existen formas reducidas para y_1 y y_2.

La pregunta clave es: ¿Con base en qué supuestos es posible estimar los parámetros en, por ejemplo, (16.17)? Este es un problema de identificación. La **condición de rango** para la identificación de la ecuación (16.17) es fácil de enunciar.

Condición de rango para la identificación de una ecuación estructural. La primera ecuación en un modelo de dos ecuaciones simultáneas se identifica si, y sólo si, la *segunda* ecuación contiene al menos una variable exógena (con un coeficiente diferente de cero) excluida de la primera.

Esta es la condición necesaria y suficiente para que se pueda identificar la ecuación (16.17). La **condición de orden**, que se analizó en el capítulo 15, es necesaria para la condición de rango. La condición de orden para identificar la primera ecuación establece que al menos una variable exógena se excluye de esta ecuación. Una vez que ambas ecuaciones se han especificado es trivial verificar la condición de orden. La condición de rango requiere más: al menos una de las variables exógenas excluidas de la primera ecuación debe tener un coeficiente de población diferente de cero en la segunda ecuación. Esto asegura que al menos una de las variables exógenas omitidas de la primera ecuación, en realidad, aparezca en la forma reducida de y_2, así que se pueden usar estas variables como instrumentos de y_2. Esto se puede verificar mediante una prueba *t* o *F*, como en el capítulo 15. A continuación se muestran algunos ejemplos.

La identificación de la segunda ecuación es, naturalmente, tan solo la imagen reflejo del enunciado de la primera ecuación. Por otra parte, si se escriben las ecuaciones como en el ejemplo de oferta y demanda de mano de obra en la sección 16.1, de manera que y_1 aparezca del lado izquierdo de *ambas* ecuaciones y y_2 en el lado derecho, la condición de identificación es idéntica.

Ejemplo 16.3

[Oferta de mano de obra de mujeres trabajadoras y casadas]

Para ilustrar la cuestión de la identificación, considere la oferta de mano de obra de trabajadoras casadas que ya forman parte del mercado laboral. En lugar de la función de la demanda, se escribe la oferta salarial en función de las horas y las variables de productividad acostumbradas. Con la condición de equilibrio impuesta, las dos ecuaciones estructurales son

$$\begin{aligned} hours = {} & \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 \\ & + \beta_{14} nwifeinc + u_1 \end{aligned} \tag{16.19}$$

y

$$\begin{aligned} \log(wage) = {} & \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper \\ & + \beta_{23} exper^2 + u_2. \end{aligned} \tag{16.20}$$

La variable *age* es la edad de la mujer, *kidslt6* es el número de niños que tienen menos de seis años de edad, *nwifeinc* es el ingreso no salarial de la mujer (que incluye los ingresos de su cónyuge) y *educ* (educación) y *exper* (experiencia) son los años de educación y experiencia previa, respectivamente. Se supone que todas las variables, salvo *hours* y log(*wage*) son exógenas. (Este es un supuesto débil, puesto que *educ* podría estar correlacionado con la capacidad omitida en su ecuación. Pero con fines de ilustración, se ignora el problema de capacidad omitida.) La forma funcional en este sistema, donde *hours* aparece en forma nivelada pero *wage* (salario) está en forma logarítmica, es común en economía laboral. Se puede escribir este sistema como en las ecuaciones (16.17) y (16.18) al definir $y_1 = hours$ y $y_2 = \log(wage)$.

La primera ecuación es la función de la oferta y satisface la condición de orden, pues las dos variables exógenas, *exper* y $exper^2$, se omitieron en la ecuación de oferta de mano de obra. Estas restricciones de exclusión son supuestos cruciales: se supone que, una vez que se controlan el salario, la educación, la edad, el número de niños pequeños y otros ingresos, la experiencia pasada no tiene efecto sobre la oferta actual de mano de obra. Desde luego, este supuesto es muy cuestionable, pero se utiliza aquí a modo de ilustración.

Dadas las ecuaciones (16.19) y (16.20), la condición de rango para identificar la primera ecuación es que al menos *exper* o $exper^2$ tiene un coeficiente diferente de cero en la ecuación (16.20). Si $\beta_{22} = 0$ y

$\beta_{23} = 0$, no existen variables exógenas que aparezcan en la segunda ecuación y que tampoco aparezcan en la primera (*educ* aparece en ambas). Se puede decir que la condición de rango para la identificación de (16.19) es equivalente en términos en la forma reducida de log(*wage*), que es

$$\log(wage) = \pi_{20} + \pi_{21}educ + \pi_{22}age + \pi_{23}kidslt6 + \pi_{24}nwifeinc + \pi_{25}exper + \pi_{26}exper^2 + v_2. \quad \textbf{16.21}$$

Para la identificación, se necesita que $\pi_{25} \neq 0$ o $\pi_{26} \neq 0$, algo que se puede probar mediante un estadístico F estándar, como se analizó en el capítulo 15.

La ecuación de la oferta salarial, (16.20), se identifica si al menos una de las variables *age*, *kidslt6*, o *nwifeinc* tiene un coeficiente diferente de cero en (16.19). Esto es idéntico a suponer que la forma reducida de *hours*, que tiene la misma forma que el lado derecho de (16.21), depende de al menos una de las variables *age*, *kidslt6* o *nwifeinc*. Cuando se especifica la ecuación de la oferta salarial, se *supone* que ni *age*, *kidslt6* o *nwifeinc* tienen efecto alguno en el salario ofrecido, una vez que se ha dado cuenta de las horas, la educación y la experiencia. Éstos serían supuestos pobres si las variables tienen, de alguna manera, efectos directos en la productividad, o si las mujeres padecieran discriminación basada en su edad o en el número de hijos pequeños.

En el ejemplo 16.3, se toma como población de interés a las casadas que trabajan (de manera que las horas de equilibrio sean positivas). Esto excluye el grupo de mujeres casadas que eligen no trabajar fuera del hogar. Incluir a estas mujeres en el modelo plantea algunas dificultades. Por ejemplo, si una mujer no trabaja, no se puede observar su oferta salarial. En el capítulo 17 se abordan estos temas; pero por ahora, se pensará que las ecuaciones (16.19) y (16.20) comprenden a sólo una mujer que tenga $hours > 0$.

Ejemplo 16.4

[Inflación y apertura]

Romer (1993) propone modelos teóricos sobre la inflación que sugieren que los países más "abiertos" deberían tener tasas de inflación más bajas. Su análisis empírico explica las tasas de inflación anuales promedio (desde 1973) en términos de la participación promedio de las importaciones en el producto interno bruto (o nacional) desde 1973, la cual es su medida de apertura. Además de estimar la ecuación clave mediante MCO, utiliza variables instrumentales. Si bien Romer no especifica ambas ecuaciones en un sistema simultáneo, tiene en mente un sistema de dos ecuaciones:

$$inf = \beta_{10} + \alpha_1 open + \beta_{11}\log(pcinc) + u_1 \quad \textbf{16.22}$$

$$open = \beta_{20} + \alpha_2 inf + \beta_{21}\log(pcinc) + \beta_{22}\log(land) + u_2, \quad \textbf{16.23}$$

donde *pcinc* es el ingreso *per cápita* (en dólares de 1980 y se supone exógena) y *land* es la extensión territorial del país, en millas cuadradas (también supuestamente exógena). La ecuación (16.22) es de interés, por su hipótesis de que $\alpha_1 < 0$. (Las economías más abiertas tienen tasas de inflación menores.) La segunda ecuación refleja el hecho de que el grado de apertura puede depender de la tasa de inflación promedio, así como de otros factores. La variable log(*pcinc*) aparece en ambas ecuaciones, pero se

Pregunta 16.2

Si se tiene un crecimiento en la oferta de dinero desde 1973 en cada país, lo que supuestamente es exógeno, ¿esto ayuda a identificar la ecuación (16.23)?

supone que log(*land*) aparece sólo en la segunda ecuación. La idea es que, *ceteris paribus*, un país más pequeño tiende a ser más abierto (así que $\beta_{22} < 0$).

Mediante la regla de identificación que se mencionó antes, se identifica la ecuación (16.22), siempre y cuando $\beta_{22} \neq 0$. La ecuación (16.23) *no* se identifica pues contiene dos variables exógenas. Pero lo que interesa es (16.22).

Estimación mediante MC2E

Una vez que se determina que una ecuación está identificada, se puede estimar mediante mínimos cuadrados en dos etapas. Las variables instrumentales consisten en las variables exógenas que aparecen en cada ecuación.

Ejemplo 16.5

[Oferta de mano de obra de trabajadoras casadas]

Se utilizan los datos de las mujeres trabajadoras casadas en MROZ.RAW para estimar la ecuación de oferta de mano de obra (16.19) mediante MC2E. El conjunto completo de variables instrumentales incluye a *educ*, *age*, *kidslt6*, *nwifeinc*, *exper* y $exper^2$. La curva estimada de oferta de mano de obra es

$$\begin{aligned}\widehat{hours} = {} & \underset{(574.56)}{2{,}225.66} + \underset{(470.58)}{1{,}639.56}\log(wage) - \underset{(59.10)}{183.75}\,educ \\ & - \underset{(9.38)}{7.81}\,age - \underset{(182.93)}{198.15}\,kidslt6 - \underset{(6.61)}{10.17}\,nwifeinc,\ n = 428,\end{aligned} \tag{16.24}$$

la cual muestra que la curva de oferta de mano de obra tiene pendiente positiva. El coeficiente estimado en log(*wage*) tiene la siguiente interpretación, todos los factores constantes, $\widehat{\Delta hours} \approx 16.4(\%\Delta wage)$. Las elasticidades de la oferta de mano de obra se pueden calcular si se multiplican ambos lados de esta última ecuación por 100/*hours*:

$$100\cdot(\widehat{\Delta hours/hours}) \approx (1{,}640/hours)(\%\Delta wage)$$

o

$$\%\widehat{\Delta hours} \approx (1{,}640/hours)(\%\Delta wage),$$

lo cual implica que la elasticidad de la oferta de mano de obra (respecto al salario) simplemente es de 1,640/*hours*. [La elasticidad no es constante en este modelo, pues la variable dependiente en (16.24) es *hours* y no log(*hours*).] Al promedio de horas trabajadas, 1,303, la elasticidad estimada es $1{,}640/1{,}303 \approx 1.26$, lo cual implica un incremento mayor que 1% en las horas trabajadas dado un incremento de 1% en el salario. Esta es una elasticidad estimada grande. A más horas, la elasticidad será menor; a menos horas, como con $hours = 800$, la elasticidad será superior a dos.

Para fines comparativos, cuando (16.19) se estima mediante MCO, el coeficiente de log(*wage*) es -2.05 (ee $= 54.88$), lo cual implica que no hay efecto en las horas trabajadas. Para confirmar que log(*wage*) en realidad es endógena en (16.19), se puede realizar la prueba de la sección 15.5. Cuando se agregan los residuales de la forma reducida $\hat{v}_2$ a la ecuación y se estima mediante MCO, el estadístico *t* en $\hat{v}_2$ es -6.61, lo cual es muy significativo y por tanto log(*wage*) parece ser endógeno.

La ecuación de oferta salarial (16.20) también se puede estimar mediante MC2E. El resultado es

$$
\begin{aligned}
\widehat{\log(wage)} = &-.656 + .00013\ hours + .110\ educ \\
&(.338)\quad (.00025) \qquad\quad (.016) \\
&+ .035\ exper - .00071\ exper^2, \quad n = 428. \\
&\ \ (.019) \qquad\quad (.00045)
\end{aligned}
\tag{16.25}
$$

Esto difiere de las ecuaciones de salario anteriores en que *hours* se incluye como variable explicativa y MC2E se utiliza para dar cuenta de la endogeneidad de *hours* (y se supone que *educ* y *exper* son exógenas). El coeficiente en *hours* es estadísticamente insignificante, lo cual significa que no hay evidencia de que la oferta salarial aumente con las horas trabajadas. Los demás coeficientes son similares a los que se obtuvieron al eliminar *hours* y estimar la ecuación mediante MCO.

Estimar el efecto de la apertura sobre la inflación mediante variables instrumentales también es un procedimiento fácil.

Ejemplo 16.6

[Inflación y apertura]

Antes de estimar (16.22) mediante los datos en OPENNESS.RAW, se revisa para ver si *open* (apertura) tiene una correlación parcial suficiente con la VI propuesta, log(*land*). La regresión en la forma reducida es

$$
\begin{aligned}
\widehat{open} = &\ 117.08 + .546 \log(pcinc) - 7.57 \log(land) \\
&\ (15.85)\quad (1.493) \qquad\qquad\qquad (.81) \\
n = &\ 114,\ R^2 = .449.
\end{aligned}
$$

El estadístico t sobre log(*land*) es superior a nueve en valor absoluto, lo cual verifica la aseveración de Romer de que los países más pequeños son más abiertos. El hecho de que log(*pcinc*) sea tan insignificante en esta regresión es irrelevante.

Estimar (16.22) mediante log(*land*) como una VI para *open* da

$$
\begin{aligned}
\widehat{inf} = &\ 26.90 - .337\ open + .376 \log(pcinc), \quad n = 114. \\
&(15.40)\quad (.144) \qquad\quad (2.015)
\end{aligned}
\tag{16.26}
$$

Pregunta 16.3

¿Cómo se comprobaría si la diferencia entre las estimaciones de MCO y VI en *open* son estadísticamente diferentes?

El coeficiente en *open* es estadísticamente significativo al nivel aproximadamente de 1% en comparación con la alternativa ($\alpha_1 < 0$). El efecto también es económicamente importante: para cada incremento de punto porcentual en la participación de importaciones en el producto interno bruto, la inflación anual será menor aproximadamente en una tercera parte de un punto porcentual. A modo de comparación, la estimación de MCO es de $-.215$ (ee $=$ .095).

16.4 Sistemas con más de dos ecuaciones

Los modelos de ecuaciones simultáneas están compuestos de más de dos ecuaciones. Estudiar la identificación general de estos modelos es difícil y requiere cálculos de álgebra matricial. Una vez que se ha mostrado una ecuación en un sistema general para ser identificada, se puede estimar mediante MC2E.

Identificación en sistemas con tres o más ecuaciones

Se empleará un sistema de tres ecuaciones para ilustrar las cuestiones que surgen de la identificación de MES complicados. Suprimiendo los interceptos, se escribe el modelo de la siguiente manera

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}z_1 + u_1 \tag{16.27}$$

$$y_2 = \alpha_{21}y_1 + \beta_{21}z_1 + \beta_{22}z_2 + \beta_{23}z_3 + u_2 \tag{16.28}$$

$$y_3 = \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3, \tag{16.29}$$

donde y_g son las variables endógenas y z_j son exógenas. El primer subíndice en los parámetros indica el número de la ecuación y el segundo, el número de la variable; se usa α para los parámetros de las variables endógenas y β para las variables exógenas.

¿Cuál de de estas ecuaciones se puede estimar? Por lo general, es difícil mostrar que una ecuación en un MES con más de dos ecuaciones está identificada, pero es fácil notar cuando ciertas ecuaciones no están identificadas. En el sistema (16.27) a (16.29), se puede apreciar fácilmente que (16.29) cae dentro de esta categoría. Debido a que toda variable exógena aparece en esta ecuación, no se tiene ninguna VI para y_2. Por tanto, no se pueden estimar de manera consistente los parámetros de esta ecuación. Por las razones que se analizaron en la sección 16.2, la estimación de MCO, por lo general, no será consistente.

¿Qué sucede con la ecuación (16.27)? Las cosas parecen prometedoras debido a que z_2, z_3 y z_4 se excluyen de la ecuación, este es otro ejemplo de *restricciones de exclusión*. Aunque existen dos variables endógenas en esta ecuación, se tienen tres potenciales VI para y_2 y y_3. Por tanto, la ecuación (16.27) satisface la condición de orden. Para completar, se enunciará la condición de orden para MES generales.

Condición de orden para identificación. Una ecuación en cualquier MES satisface la condición de orden para identificación si el número de variables exógenas *excluidas* de la ecuación es al menos el mismo que el número de las variables endógenas del lado derecho.

La segunda ecuación, (16.28), también satisface la condición de orden, pues hay una variable exógena excluida, z_4, y una variable endógena del lado derecho, y_1.

Como se analizó en el capítulo 15 y en la sección anterior, la condición de orden sólo es necesaria, no suficiente, para la identificación. Por ejemplo, si $\beta_{34} = 0$, z_4 no aparece en ningún lugar del sistema, significaría que no se relaciona con y_1, y_2 o y_3. Si $\beta_{34} = 0$, entonces la segunda ecuación no está identificada, porque z_4 es inútil así como una VI para y_1. Esto ilustra nuevamente que la identificación de una ecuación depende de los valores de los parámetros (lo cual nunca se sabe con seguridad) en las demás ecuaciones.

Existen muchas formas sutiles en que la identificación puede fallar en MES complicados. Para obtener suficientes condiciones, se necesita ampliar la condición de rango para la identificación en sistemas de dos ecuaciones. Esto es posible, pero requiere álgebra matricial [vea, por ejemplo, Wooldridge (2002, capítulo 9)]. En muchas aplicaciones se supone que, a menos que exista una falla evidente de identificación, una ecuación que satisface la condición de orden está identificada.

La nomenclatura de las ecuaciones sobreidentificadas o exactamente identificadas del capítulo 15 se originó en los MES. En términos de la condición de orden, (16.27) es una **ecuación sobreidentificada** debido a que sólo se necesitan dos VI (para y_2 y y_3) pero se dispone de tres (z_2, z_3 y z_4); no existe restricción de sobreidentificación en esta ecuación. En general, el número de restricciones es igual al número total de variables exógenas en el sistema, menos el número total de variables explicativas en la ecuación. Esto se puede probar mediante la prueba de sobreidentificación de la sección 15.5. La ecuación (16.28) es una **ecuación exactamente identificada**, y la tercera de ellas es una **ecuación no identificada**.

Estimación

Sin importar el número de ecuaciones en un MES, cada ecuación identificada puede estimarse mediante MC2E. Los instrumentos de una ecuación en particular consisten en variables exógenas que aparecen en cualquier parte del sistema. Las pruebas para endogeneidad, heterocedasticidad, correlación serial y restricciones sobreidentificadas se obtienen tal como en el capítulo 15.

Resulta que, cuando se especifica de forma correcta cualquier sistema con dos o más ecuaciones, y ciertos supuestos adicionales se mantienen, los *métodos de estimación de sistemas* suelen ser más eficientes que estimar cada ecuación mediante MC2E. El método de estimación de sistemas más común en el contexto de MES es el de *mínimos cuadrados en tres etapas*. Estos métodos, con o sin variables exploratorias endógenas van más allá del alcance de este libro [vea, por ejemplo, Wooldridge (2002, capítulos 7 y 8).]

16.5 Modelos de ecuaciones simultáneas con series de tiempo

Entre las primeras aplicaciones de MES están las ecuaciones simultáneas de grandes sistemas que se utilizaron para describir la economía de un país. Un modelo keynesiano simple de la demanda agregada (que ignora las importaciones y exportaciones) es

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1} \tag{16.30}$$

$$I_t = \gamma_0 + \gamma_1 r_t + u_{t2} \tag{16.31}$$

$$Y_t \equiv C_t + I_t + G_t, \tag{16.32}$$

donde

C_t = consumo.
Y_t = ingreso.
T_t = impuestos.
r_t = la tasa de interés.
I_t = inversión.
G_t = gasto gubernamental.

[Vea, por ejemplo, Mankiw (1994, capítulo 9).] En concreto, suponga que t representa el año.

La primera ecuación es una función del consumo agregado, donde el consumo depende del ingreso disponible, la tasa de interés y el error estructural inobservable u_{t1}. La secunda ecuación es una función muy simple de la inversión. La ecuación (16.32) es una *identidad* que resulta de la contabilidad nacional del ingreso: por definición es válida, sin error. Por tanto, no tiene sentido que se estime (16.32), pero se necesita esta ecuación para redondear el modelo.

Debido a que existen tres ecuaciones en el sistema, también debe haber tres variables endógenas. Dadas las dos primeras ecuaciones, es claro que se intenta que C_t e I_t sean endógenas. Además, debido a la identidad contable, Y_t es endógena. Se supondría que, al menos en este modelo T_t, r_t y G_t sean exógenas, así que no están correlacionadas con u_{t1} ni u_{t2}. (Más adelante se analizarán problemas con este tipo de supuesto.)

Si r_t es exógena, entonces la estimación mediante MCO de la ecuación (16.31) es natural. Sin embargo, la función del consumo, depende del ingreso disponible, que es endógeno debido a que Y_t lo es. Se tienen dos instrumentos disponibles con base en los supuestos de exogeneidad: T_t y G_t. Por tanto, si se sigue la prescripción para estimar ecuaciones de corte transversal, se estimaría (16.30) por MC2E mediante los instrumentos (T_t,G_t,r_t).

Ahora, modelos como (16.30) a (16.32) se estiman de forma aleatoria, por varias buenas razones. Primero, es muy difícil justificar, a nivel agregado, el supuesto de que los impuestos, tasas de interés y gasto gubernamental sean variables exógenas. Es claro que los impuestos dependen directamente del ingreso; por ejemplo, con una tasa impositiva sobre el ingreso marginal de τ_t en el año t, $T_t = \tau_t Y_t$. Se puede permitir esto fácilmente al remplazar $(Y_t - T_t)$ con $(1 - \tau_t)Y_t$ en (16.30), y también se puede estimar la ecuación mediante MC2E si se supone que el gasto gubernamental es exógeno. Se puede también agregar la tasa impositiva a la lista de instrumentos, si esta es exógena. Pero, ¿el gasto gubernamental y la tasa impositiva en realidad son exógenos? Desde luego, en principio, así podría ser, si el gobierno fijara las tasas impositivas y el gasto independientemente de lo que esté sucediendo en la economía. Pero es difícil que suceda en la realidad: el gasto gubernamental suele depender del nivel de ingreso y a altos niveles de ingreso, se recaudan los mismos montos fiscales para las tasas impositivas marginales menores. Además, suponer que las tasas de interés sean exógenas es muy cuestionable. Se puede especificar un modelo más realista que incluya la demanda y la oferta de dinero, y entonces, las tasas de interés podrían determinarse de manera conjunta con C_t, I_t, y Y_t. Pero entonces hallar suficientes variables exógenas para identificar las ecuaciones se vuelve muy difícil (y los siguientes problemas relativos a estos modelos aún persisten).

Algunos han argumentado que ciertos componentes del gasto gubernamental, como los de defensa [vea, por ejemplo, Hall (1988) y Ramey (1991)] son exógenos en una variedad de aplicaciones de ecuaciones simultáneas. Pero esto no es universalmente aceptado, y, en cualquier caso, el gasto de defensa no siempre se correlaciona de manera apropiada con las variables explicativas endógenas [vea Shea (1993) para un análisis y el ejercicio para computadora C16.6 para un ejemplo].

Un segundo problema con un modelo como el (16.30) a (16.32) es que es completamente estático. En especial con datos mensuales o trimestrales, pero incluso con datos anuales, suelen esperarse rezagos de ajuste. (Un argumento a favor de los modelos estáticos del tipo keynesiano es que tienen la intención de describir el largo plazo sin preocuparse de las dinámicas del corto plazo.) Permitir este tipo de dinámicas no es difícil. Por ejemplo, se puede agregar un ingreso rezagado a la ecuación (16.31):

$$I_t = \gamma_0 + \gamma_1 r_t + \gamma_2 Y_{t-1} + u_{t2}. \tag{16.33}$$

En otras palabras, se agrega una **variable endógena rezagada** (pero no I_{t-1}) a la ecuación de inversión. ¿Se puede tratar a Y_{t-1} como exógena en esta ecuación? Con base en ciertos supuestos sobre u_{t2}, la respuesta es sí. Pero por lo general, una variable endógena rezagada en un MES recibe el nombre de **variable predeterminada**. Los rezagos en las variables exógenas también son predeterminadas. Si se supone que u_{t2} no está correlacionada con las variables exógenas actuales ni con todas las variables endógenas y exógenas *pasadas*, entonces Y_{t-1} no está correlacionada con u_{t2}. Dada la exogeneidad de r_t, se puede estimar (16.33) mediante MCO.

Si se agrega un consumo rezagado a (16.30), se puede tratar a C_{t-1} como exógena en esta ecuación con base en los mismos supuestos sobre u_{t1} que se hicieron para u_{t2} en el párrafo anterior. El ingreso disponible actual sigue siendo endógeno en

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + \beta_3 C_{t-1} + u_{t1}, \tag{16.34}$$

así que se podría estimar esta ecuación mediante MC2E y con los instrumentos (T_t, G_t, r_t, C_{t-1}); si la inversión se determina por (16.33), Y_{t-1} debe agregarse a la lista de instrumentos. [Para ver por qué, use (16.32), (16.33) y (16.34) para hallar la forma reducida de Y_t en términos de las variables exógenas y predeterminadas: T_t, r_t, G_t, C_{t-1} y Y_{t-1}. Como Y_{t-1} se muestra en esta forma reducida, se debe emplear como una VI.]

La presencia de dinámica en el MES agregado es, al menos para fines de pronóstico, una clara mejora sobre los MES estáticos. Pero aún existen algunos importantes problemas cuando se estiman MES con datos de series de tiempo agregados, algunos de los cuales se analizaron en los capítulos 11 y 15. Recuerde que la validez de los procedimientos de inferencia con MCO o MC2E habituales en aplicaciones de series de tiempo gira en torno a la noción de *dependencia débil*. Por desgracia, series como el consumo agregado, el ingreso, la inversión e incluso las tasas de interés parecen violar los requisitos de dependencia débil. (En la terminología del capítulo 11 tiene *raíces unitarias*.) Estas series también suelen tener tendencias exponenciales, aunque esto puede superarse parcialmente mediante la transformación logarítmica y suponiendo diferentes formas funcionales. Por lo general, incluso en muestras grandes, y no se diga en las pequeñas, las propiedades de MCO y MC2E son complicadas y dependen de varios supuestos cuando se aplican a ecuaciones con variables I(1). Se abordan de manera breve estas cuestiones en el capítulo 18, pero un tratamiento general más avanzado lo encontrará en Hamilton (1994).

¿El análisis anterior significa que la aplicación de MES no es útil en los datos de series de tiempo? En absoluto. Los problemas de las tendencias y la alta persistencia se pueden evitar al especificar los sistemas en las primeras diferencias o tasas de crecimiento; pero es necesario reconocer que esta es un MES diferente del que se especifica en niveles. [Por ejemplo, si se especifica el aumento del consumo como una función del aumento del ingreso disponible y los cambios de las tasas de interés, esto es diferente de (16.30).] Por otra parte, como antes se discutió, incorporar la dinámica no es especialmente difícil. Por último, el problema de encontrar variables verdaderamente exógenas por incluir en MES suele ser más fácil con datos no agregados. Por ejemplo, para las industrias manufactureras, Shea (1993) describe cómo la producción (o, de forma más precisa, el crecimiento en la producción) en otras industrias puede utilizarse como un instrumento para estimar las funciones de la oferta. Ramey (1991) también tiene un análisis convincente de la estimación de las funciones de costo en la industria mediante variables instrumentales que emplean datos de series de tiempo.

El siguiente ejemplo muestra cómo se pueden utilizar los datos agregados para probar una importante teoría económica, la teoría del ingreso permanente del consumo, que por lo regular se conoce como *hipótesis del ingreso permanente* (HIP). El método que se utiliza en este ejemplo no está basado, en términos estrictos, en un modelo de ecuaciones simultáneas, pero se puede pensar en el crecimiento del consumo y el ingreso (así como en las tasas de interés) como determinados conjuntamente.

Ejemplo 16.7

[Prueba de hipótesis del ingreso permanente]

Campbell y Mankiw (1990) emplearon métodos de variables instrumentales para probar varias versiones de hipótesis de ingreso permanente. Se usarán los datos anuales de 1959 a 1995 en CONSUMP.RAW para imitar uno de sus análisis. Campbell y Mankiw emplearon datos trimestrales hasta 1985.

Una ecuación que Campbell y Mankiw estimaron (escrita con nuestra notación) es

$$gc_t = \beta_0 + \beta_1 gy_t + \beta_2 r3_t + u_t, \quad \textbf{16.35}$$

donde

$gc_t = \Delta\log(c_t)$ = crecimiento anual en consumo real *per cápita* (excluyendo a los no perecederos).

gy_t = crecimiento en el ingreso real disponible.

$r3_t$ = la tasa de interés real (ex post) medida por el rendimiento en las tasas trimestrales de los bonos del Tesoro: $r3_t = i3_t - inf_t$, donde la tasa de inflación está basada en el Índice de precios al consumidor.

Las tasas de crecimiento de consumo e ingreso disponible no poseen tendencia y son débilmente dependientes; se supondrá que este también es el caso para $r3_t$ así que se puede aplicar la teoría estándar asintótica.

La característica clave de la ecuación (16.35) es que la HIP implica que el término de error u_t tiene una media condicional de cero en toda la información observada en el tiempo $t - 1$ o antes: $\mathrm{E}(u_t|I_{t-1}) = 0$. Sin embargo, *no* es necesario que u_t *no* esté correlacionado con gy_t o $r3_t$; una forma tradicional de pensar esto es que dichas variables están determinadas en forma conjunta, pero no se escribirá un sistema completo de tres ecuaciones.

Debido a que u_t no está correlacionado con todas las variables de fecha $t - 1$ o anterior, los instrumentos válidos para estimar (16.35) son los valores rezagados de gc, gy y $r3$ (y los rezagos de las otras variables, pero no se usarán aquí). ¿Cuáles son las hipótesis de interés? La forma pura de HIP tiene $\beta_1 = \beta_2 = 0$. Campbell y Mankiw argumentan que β_1 es positiva si alguna fracción de la población consume su ingreso actual, más que su ingreso permanente. La HIP con una tasa de interés real y no constante implica que $\beta_2 > 0$.

Cuando se estima (16.35) mediante MC2E, usando los instrumentos gc_{-1}, gy_{-1} y $r3_{-1}$ como variables endógenas gy_t y $r3_t$, se obtiene

$$\begin{aligned} \widehat{gc}_t &= \underset{(.0032)}{.0081} + \underset{(.135)}{.586}\, gy_t - \underset{(.00076)}{.00027}\, r3_t \\ n &= 35,\ R^2 = .678. \end{aligned} \quad \textbf{16.36}$$

Por tanto, la forma pura de HIP se rechaza fuertemente debido a que el coeficiente en gy es económicamente grande (un incremento de 1% en el ingreso disponible aumenta el consumo en más de .5%) y es estadísticamente significativo ($t = 4.34$). En contraste, el coeficiente de interés real es muy pequeño y estadísticamente insignificante. Estos hallazgos son cualitativamente iguales que los de Campbell y Mankiw.

La HIP también implica que los errores $\{u_t\}$ no están correlacionados serialmente. Después de la estimación de MC2E, se obtienen los residuales, $\hat{u}_t$, y se incluye $\hat{u}_{t-1}$ como una variable explicativa adicional en (16.36); aún se emplearán los instrumentos gc_{t-1}, gy_{t-1}, $r3_{t-1}$, y $\hat{u}_{t-1}$ actúa como su propio instrumento (vea la sección 15.7). El coeficiente en $\hat{u}_{t-1}$ es $\hat{\rho} = .187$ (ee = .133), así que existe cierta evidencia de correlación serial positiva, aunque no al nivel de significancia de 5%. Campbell y Mankiw analizan el por qué, con los datos trimestrales disponibles, la correlación serial positiva puede encontrarse en errores aun cuando la HIP fuera válida; algunas de estas cuestiones también se relacionan con los datos anuales.

Pregunta 16.4

Suponga que para una ciudad particular se tienen datos mensuales sobre el consumo de pescado *per cápita*, el ingreso *per cápita*, el precio del pescado y los precios de la carne de res y pollo; el ingreso y los precios de la carne de res y el pollo son exógenos. Suponga que no existe estacionalidad en la función de la demanda del pescado, pero que existe en la oferta de pescado. ¿Cómo se puede emplear esta información para estimar una ecuación de elasticidad constante de la demanda de pescado? Especifique la ecuación y analice su identificación. (*Sugerencia*: se deben tener 11 variables instrumentales para el precio del pescado.)

Utilizar las tasas de crecimiento de las tendencias o variables I(1) en MES es muy común en aplicaciones de series de tiempo. Por ejemplo, Shea (1993) estima curvas de oferta en la industria especificadas en términos de las tasas de crecimiento.

Si un modelo estructural contiene una tendencia de tiempo, lo cual puede capturar factores exógenos de tendencia que no están modelados directamente, entonces la tendencia actúa como su propia VI.

16.6 Modelos de ecuaciones simultáneas con datos de panel

Los modelos de ecuaciones simultáneas también surgen en contextos de datos de panel. Por ejemplo, se puede imaginar que se estiman ecuaciones de oferta salarial y de mano de obra, como en el ejemplo 16.3, para un grupo de personas que trabaja durante un periodo determinado. Además de permitir la determinación simultánea de las variables dentro de cada periodo, se pueden permitir efectos inobservables en cada ecuación. En una función de la oferta de mano de obra, sería útil considerar el gusto por el ocio, que es inobservable y que no cambia con el tiempo.

El método básico para estimar MES con datos de panel implica dos pasos: 1) eliminar los efectos inobservables de las ecuaciones de interés mediante la transformación de efectos fijos o la primera diferencia y 2) encontrar variables instrumentales para aquellas endógenas en la ecuación transformada. Esto puede ser muy difícil, pues para un análisis convincente es necesario hallar instrumentos que cambien con el tiempo. Para ver por qué, escriba un MES para datos de panel como

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1} \quad \mathbf{16.37}$$

$$y_{it2} = \alpha_2 y_{it1} + \mathbf{z}_{it2}\boldsymbol{\beta}_2 + a_{i2} + u_{it2}, \quad \mathbf{16.38}$$

donde i denota el corte transversal, t el periodo y $\mathbf{z}_{it1}\boldsymbol{\beta}_1$ o $\mathbf{z}_{it2}\boldsymbol{\beta}_2$ denotan funciones lineales de un conjunto de variables explicativas exógenas en cada ecuación. El análisis más general permite que los efectos inobservables, a_{i1} y a_{i2}, estén correlacionados con *todas* las variables explicativas, incluso los elementos en z. No obstante, se supone que los errores estructurales idiosincráticos, u_{it1} y u_{it2}, no estén correlacionados con las $\mathbf{z}$ en ambas ecuaciones y en todos los periodos; este es el sentido en el que las $\mathbf{z}$ son exógenas. Salvo casos especiales, y_{it2} está correlacionada con u_{it1}, y y_{it1} está correlacionada con u_{it2}.

Suponga que interesa la ecuación (16.37). No se puede estimar por MCO, pues el error compuesto $a_{i1} + u_{it1}$ se correlaciona potencialmente con todas las variables explicativas. Suponga que se realiza la diferencia en el tiempo para eliminar el efecto inobservable, a_{i1}:

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \Delta\mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \quad \mathbf{16.39}$$

(Como con las diferencias o las desviaciones respecto a la media temporal, sólo se pueden estimar los efectos de las variables que cambian con el tiempo en al menos algunas unidades de corte transversal.) Ahora, el término de error de esta ecuación no está correlacionado con $\Delta\mathbf{z}_{it1}$ por el supuesto. Pero Δy_{it2} y Δu_{it1} están correlacionados positivamente. Por tanto, se necesita una VI para Δy_{it2}.

Como en el caso de los datos puros de corte transversal o de series de tiempo, las VI posibles provienen de la *otra* ecuación: los elementos en $\mathbf{z}_{it2}$ que no están tampoco en $\mathbf{z}_{it1}$. En la práctica, se necesitan elementos que *varíen con el tiempo* en $\mathbf{z}_{it2}$ y que no lo hagan en $\mathbf{z}_{it1}$. Esto se debe a que se requiere un instrumento para Δy_{it2}, y es poco probable que un cambio en una variable de un periodo al siguiente esté altamente correlacionada con el *nivel* de variables exógenas. De hecho, si se diferencia (16.38), se observa que las VI naturales para Δy_{it2} son los elementos de $\Delta \mathbf{z}_{it2}$ que no están en $\Delta \mathbf{z}_{it1}$.

Como ejemplo de los problemas que pueden surgir, considere una versión de datos de panel de la función de la oferta de mano de obra en el ejemplo 16.3. Después de diferenciar, suponga que se tiene la ecuación

$$\Delta hours_{it} = \beta_0 + \alpha_1 \Delta \log(wage_{it}) + \Delta(other\ factors_{it}),$$

y que se desea usar $\Delta exper_{it}$ como instrumento para $\Delta \log(wage_{it})$. El problema es que, debido a que se buscan personas que trabajen en todos los periodos, $\Delta exper_{it} = 1$ para toda i y t. (Cada persona obtiene otro año de experiencia después de transcurrido un año.) No se puede emplear una VI que tenga el mismo valor para toda i y t, por lo que se deberá buscar en otra parte.

Con frecuencia, la participación en un programa experimental puede ser utilizado para obtener VI en contextos de paneles de datos. En el ejemplo 15.10, se emplearon los subsidios para la capacitación laboral como una VI para el cambio en las horas de capacitación como determinante de los efectos de la capacitación laboral sobre la productividad del trabajador. De hecho, se puede ver que en un contexto MES: la capacitación laboral y la productividad se determinan de forma conjunta, pero la recepción del subsidio para capacitación laboral es exógena en la ecuación (15.57).

En ocasiones, se pueden idear variables instrumentales más inteligentes y convincentes en las aplicaciones de paneles de datos, como lo ilustra el siguiente ejemplo.

Ejemplo 16.8

[Efecto de la población carcelaria en los índices de delitos violentos]

Con el fin de estimar el efecto causal del incremento de la población carcelaria en el índice delictivo a nivel estatal, Levitt (1996) utilizó casos de pleitos en prisiones sobrepobladas como instrumentos del crecimiento en la población de la prisión. La ecuación que Levitt estimó es en primeras diferencias; se puede escribir un modelo subyacente de efectos fijos como

$$\log(crime_{it}) = \theta_t + \alpha_1 \log(prison_{it}) + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1}, \tag{16.40}$$

donde θ_t denota diferentes interceptos de tiempo, y *crime* (delitos) y *prison* (población carcelaria) se miden por cada 100,000 personas. (La variable de la población de la prisión se mide sobre el último día del año anterior.) El vector $\mathbf{z}_{it1}$ contiene el logaritmo de policías *per cápita*, el logaritmo de ingreso *per cápita*, la tasa de desempleo, las proporciones de negros y personas que viven en áreas metropolitanas y las proporciones de distribución por edades.

Al diferenciar (16.40) se obtiene la ecuación estimada por Levitt:

$$\Delta \log(crime_{it}) = \xi_t + \alpha_1 \Delta \log(prison_{it}) + \Delta \mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \tag{16.41}$$

La simultaneidad entre los índices delictivos y la población carcelaria, o de manera más precisa en las tasas de crecimiento, hace que la estimación de MCO de (16.41) sea por lo general, más inconsistente. Mediante el índice de delitos violentos y un subconjunto de datos de Levitt (en PRISON.RAW, para los años de 1980 a 1993, para $51 \cdot 14 = 714$ observaciones totales), se obtiene la estimación combinada de MCO de α_1, que es $-.181$ (ee $= .048$). También se estima (16.41) por MC2E combinados, donde los instrumentos para $\Delta \log(prison)$ son dos variables binarias, cada una para si se llegó a la decisión final sobre las peleas por sobrepoblación en el año presente o en los pasados dos años. La estimación combinada de MC2E de α_1 es -1.032 (ee $=$ .370). Por tanto, el efecto estimado de MC2E es mucho mayor; no es de sorprender que sea mucho menos preciso también. Levitt encontró resultados similares cuando usó un periodo de tiempo mayor (pero sin las primeras observaciones para algunos estados) y más instrumentos.

Probar la correlación serial AR(1) en $r_{it1} = \Delta u_{it1}$ es fácil. Después de la estimación combinada de MC2E, obtener los residuales, $\hat{r}_{it1}$. Después incluir un rezago de estos residuales en la ecuación original y estimar la ecuación por MC2E, donde $\hat{r}_{it1}$ actúa como su propio instrumento. El primer año se pierde debido al rezago. Entonces, el estadístico t usual de MC2E sobre el residual rezagado es una prueba válida de la correlación serial. En el ejemplo 16.8 el coeficiente en $\hat{r}_{it1}$ es aproximadamente de .076 con $t = 1.67$. Con un coeficiente tan pequeño y un estadístico t modesto, se puede suponer de manera segura una independencia serial.

Un método alternativo para estimar MES con datos de panel es utilizar la transformación de efectos fijos y después aplicar una técnica de VI como MC2E combinado. Un procedimiento simple es estimar la ecuación en desviaciones respecto a su media temporal mediante MC2E combinados, lo cual sería

$$\ddot{y}_{it1} = \alpha_1 \ddot{y}_{t2} + \ddot{\mathbf{z}}_{it1}\boldsymbol{\beta}_1 + \ddot{u}_{it1}, \quad t = 1, 2, \ldots, T, \tag{16.42}$$

donde $\ddot{\mathbf{z}}_{it1}$ y $\ddot{\mathbf{z}}_{it2}$ son VI. Esto es equivalente a emplear MC2E en la formulación de la variable binaria, donde las variables binarias específicas de unidad actúan como sus propios instrumentos. Ayres y Levitt (1988) aplicaron MC2E a una ecuación en desviaciones respecto a la media temporal para estimar el efecto de los dispositivos electrónicos de prevención de robos LoJack sobre los índices de robo de automóviles en las ciudades. Si (16.42) se estima directamente, entonces se necesita corregir los gl a $N(T - 1) - k_1$, donde k_1 es el número total de elementos en α_1 y $\boldsymbol{\beta}_1$. Incluidas las variables binarias específicas de unidad y MC2E combinado aplicado a los procedimientos de datos originales para corregir los gl.

RESUMEN

Los modelos de ecuaciones simultáneas son apropiados cuando cada ecuación en el sistema tiene una interpretación *ceteris paribus*. Un buen ejemplo es cuando las ecuaciones separadas describen diferentes lados de un mismo mercado o de las relaciones de comportamiento de diferentes agentes económicos. Los ejemplos de oferta y demanda son los casos principales, pero existen muchas otras aplicaciones de MES en las ciencias económicas y sociales.

Una característica adecuada de los MES es que, al especificar totalmente el sistema, resulta evidente qué variables se suponen como exógenas y cuáles aparecen en cada ecuación. Dado un sistema completo, se podrá determinar qué ecuaciones se pueden identificar (es decir, se pueden estimar). En el importante caso de un sistema de dos ecuaciones, la identificación de (por ejemplo) la primera ecuación es fácil de expresar: al menos una variable exógena debe excluirse de la primera ecuación que aparece con un coeficiente diferente de cero en la segunda ecuación.

Como se sabe por los capítulos anteriores, la estimación de MCO de una ecuación que contiene una variable explicativa endógena, por lo general, produce estimadores sesgados e inconsistentes. En lugar de ello, MC2E se puede emplear para estimar cualquier ecuación identificada en un sistema. Se disponen de métodos de sistemas más avanzados, pero rebasan el alcance de este libro.

La distinción entre variables omitidas y simultaneidad en las aplicaciones no siempre es evidente. Ambos problemas, sin mencionar el error de medición, pueden aparecer en la misma ecuación. Un buen ejemplo es la oferta de mano de obra de las trabajadoras. Los años de educación (*educ*) aparecen tanto en las funciones de oferta de mano de obra y de salario [vea las ecuaciones (16.19) y (16.20)]. Si la capacidad omitida está en el término de error de la función de la oferta de mano de obra, entonces el salario y la educación serán endógenos. Lo importante es que una ecuación estimada mediante MC2E pueda sostenerse por sí misma.

Los MES se pueden aplicar también a datos de series de tiempo. Al igual que la estimación de MCO, se deben tomar en cuenta los procesos integrados cuando se aplica MC2E. Los problemas como la correlación serial se pueden manejar como en la sección 15.7. También se dio un ejemplo de cómo estimar datos de panel usando MES, donde la ecuación tiene primera diferencia mediante MC2E combinados, tal como en el capítulo 15. Por otra parte, en algunos casos, se pueden calcular las desviaciones de todas las variables, incluidas las VI y después aplicar MC2E combinados; esto es lo mismo que incluir variables binarias por cada observación de corte transversal y utilizar MC2E, donde las binarias actúen como sus propios instrumentos. Las aplicaciones de MES con datos de panel son muy poderosas, pues permiten controlar la heterogeneidad inobservada mientras se maneja la simultaneidad. Se están volviendo cada vez más comunes y no son difíciles de estimar.

TÉRMINOS CLAVE

Condición de orden
Condición de rango
Ecuación en la forma reducida
Ecuación estructural
Ecuación exactamente identificada
Ecuación identificada
Ecuación no identificada
Ecuación sobreidentificada
Error de la forma reducida
Errores estructurales
Modelo de ecuaciones simultáneas (MES)
Parámetros de la forma reducida
Parámetros estructurales
Restricciones de exclusión
Sesgo de simultaneidad
Simultaneidad
Variable endógena rezagada
Variable predeterminada
Variables endógenas

PROBLEMAS

16.1 Escriba un sistema de dos ecuaciones en la "forma de oferta y demanda", es decir, con la misma variable y_1 (por lo general, "cantidad") en el lado izquierdo:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_1 = \alpha_2 y_2 + \beta_2 z_2 + u_2.$$

i) Si $\alpha_1 = 0$ o $\alpha_2 = 0$, explicar por qué existe una forma reducida para y_1. (Recuerde que una forma reducida expresa a y_1 como una función lineal de las variables exógenas y los errores estructurales.) Si $\alpha_1 \neq 0$ y $\alpha_2 = 0$, determine la forma reducida de y_2.

ii) Si $\alpha_1 \neq 0$, $\alpha_2 \neq 0$ y $\alpha_1 \neq \alpha_2$, determine la forma reducida de y_1. ¿y_2 tiene una forma reducida en este caso?

iii) ¿Es probable que la condición $\alpha_1 \neq \alpha_2$ satisfaga los ejemplos de oferta y demanda? Explicar.

16.2 Sea *corn* el consumo de maíz *per cápita* en fanegas, a nivel de demarcación, sea *price* el precio por fanega de maíz, sea *income* el ingreso *per cápita* en la demarcación, y *rainfall* las pulgadas de lluvia durante la última temporada de cultivo de maíz. El siguiente modelo de ecuaciones simultáneas impone la condición de equilibrio de que la demanda es igual a la oferta:

$$corn = \alpha_1 price + \beta_1 income + u_1$$
$$corn = \alpha_2 price + \beta_2 rainfall + \gamma_2 rainfall^2 + u_2.$$

¿Cuál es la ecuación de la oferta y cuál la de demanda? Explique.

16.3 En el problema 3.3 del capítulo 3, se estimó una ecuación para analizar la compensación entre minutos por semana que se invertían durmiendo (*sleep*) y minutos por semana que se invertían trabajando (*totwrk*) entre una muestra aleatoria de individuos. También se incluyó la educación y la edad en la ecuación. Debido a que el individuo elige conjuntamente *sleep* y *totwrk*, ¿la compensación estimada entre dormir y trabajar está sujeta a la crítica del "sesgo de simultaneidad"? Explique.

16.4 Suponga que las ganancias anuales y el consumo de alcohol están determinados por el MES

$$\log(earnings) = \beta_0 + \beta_1 alcohol + \beta_2 educ + u_1$$
$$alcohol = \gamma_0 + \gamma_1 \log(earnings) + \gamma_2 educ + \gamma_3 \log(price) + u_2,$$

donde *price* es el índice local de precios del alcohol, que incluye impuestos locales y estatales. Suponga que *educ* y *price* son exógenos. Si β_1, β_2, γ_1, γ_2 y γ_3 son distintos a cero, ¿qué ecuación se identifica? ¿Cómo se estimaría dicha ecuación?

16.5 Un modelo simple para determinar la efectividad del uso del preservativo para reducir las enfermedades de transmisión sexual entre estudiantes sexualmente activos del bachillerato es

$$infrate = \beta_0 + \beta_1 conuse + \beta_2 percmale + \beta_3 avginc + \beta_4 city + u_1,$$

donde

infrate = es el porcentaje de estudiantes sexualmente activos que contrajeron una enfermedad venérea.
conuse = es el porcentaje de jóvenes que afirman usar regularmente preservativos.
avginc = es el ingreso familiar promedio.
city = es la variable binaria que indica si una escuela está en una ciudad.

El modelo se encuentra a nivel de la escuela.

i) Si la ecuación anterior se interpreta en una forma causal, *ceteris paribus*, ¿cuál sería el signo de β_1?

ii) ¿Por qué *infrate* y *conuse* se determinan conjuntamente?

iii) Si el uso del preservativo aumenta con la tasa de enfermedades sexuales, de manera que $\gamma_1 > 0$ en la ecuación

$$conuse = \gamma_0 + \gamma_1 infrate + otros\ factores,$$

¿cuál es el sesgo probable cuando se estima β_1 mediante MCO?

iv) Sea *condis* una variable binaria igual a la unidad si una escuela tiene un programa para distribuir preservativos. Explique cómo se puede utilizar esto para estimar β_1 (y las demás betas) mediante VI. ¿Qué se tiene que suponer acerca de *condis* en cada ecuación?

16.6 Considere un modelo de probabilidad lineal que describa si los empleadores ofrecen un plan de retiro basado en el porcentaje de trabajadores que pertenecen a un sindicato, así como otros factores:

$$pension = \beta_0 + \beta_1 percunion + \beta_2 avgage + \beta_3 avgeduc$$
$$+ \beta_4 percmale + \beta_5 percmarr + u_1.$$

i) ¿Por qué *percunion* se determinaría conjuntamente con *pension*?

ii) Suponga que es posible encuestar a los trabajadores de las empresas y recabar información sobre sus familias. ¿Qué información se podría utilizar para construir una VI para *percunion*?

iii) ¿Cómo se probaría si la variable es al menos una candidata para VI razonable para *percunion*?

16.7 En una universidad grande, se pide estimar la demanda de boletos para los juegos de básquetbol femenil. Es posible recabar datos de series de tiempo de 10 temporadas, para un total aproximado de 150 observaciones. Un modelo posible es

$$lATTEND_t = \beta_0 + \beta_1 lPRICE_t + \beta_2 WINPERC_t + \beta_3 RIVAL_t$$
$$+ \beta_4 WEEKEND_t + \beta_5 t + u_t,$$

donde

$PRICE_t$ = el precio de admisión, medido probablemente en términos reales, por ejemplo, deflactado por el Índice regional de precios al consumidor.
$WINPERC_t$ = el porcentaje actual de victorias del equipo.
$RIVAL_t$ = variable binaria que indica un juego contra un rival.
$WEEKEND_t$ = una variable binaria que indica si el juego es en un fin de semana.

l denota el logaritmo natural, así que la función de la demanda tiene elasticidad precio constante.

i) ¿Por qué es buena idea tener una tendencia temporal en la ecuación?

ii) La oferta de boletos está limitada por la capacidad del estadio; suponga que esto no ha cambiado durante 10 años. Esto indica que la cantidad proporcionada no varía con el precio. ¿Esto significa que el precio es necesariamente exógeno en la ecuación de la demanda? (*Sugerencia:* la respuesta es no.)

iii) Suponga que el precio nominal de la admisión cambia lentamente, por ejemplo, al principio de cada temporada. La directiva elige basar el precio en la asistencia promedio de la última temporada, así como en el éxito del equipo en la temporada pasada. ¿Con base en qué supuestos el porcentaje de victorias en la última temporada ($SEASPERC_{t-1}$) es una variable instrumental válida para $lPRICE_t$?

iv) ¿Parece razonable incluir el (logaritmo del) precio real de los juegos de básquetbol varonil en la ecuación? Explique. ¿Qué signo predice la teoría económica para este coeficiente? ¿Qué otra variable relacionada con el básquetbol varonil podría incluirse en la ecuación de asistencia a los partidos femeniles?

v) Si preocupa que algunas de las series, en particular *lATTEND* y *lPRICE*, tengan raíces unitarias, ¿cómo se cambiará la ecuación estimada?

vi) Si algunos juegos se agotaran, ¿qué problemas causaría esto para estimar la función de la demanda? (*Sugerencia:* si un juego está agotado, ¿necesariamente observaría la demanda verdadera?)

16.8 ¿Qué tan grande es el efecto de los gastos escolares por estudiante en el valor de las viviendas locales? Sea *HPRICE* el precio medio de la vivienda en un distrito escolar y sea *EXPEND* los gastos. Mediante los datos de panel para los años 1992, 1994 y 1996, se postula el modelo

$$lHPRICE_{it} = \theta_t + \beta_1 lEXPEND_{it} + \beta_2 lPOLICE_{it} + \beta_3 lMEDINC_{it} + \beta_4 PROPTAX_{it} + a_{i1} + u_{it1},$$

donde $POLICE_{it}$ son los gastos *per cápita* en fuerza policial, $MEDINC_{it}$ es el ingreso medio, y $PROPTAX_{it}$ es el impuesto sobre la propiedad; *l* denota el logaritmo natural. Los gastos y el precio de la vivienda están determinados simultáneamente debido a que el valor de las casas afecta directamente a los ingresos disponibles para financiar escuelas.

Suponga que, en 1994, la forma en que las escuelas se financiaban cambió drásticamente: en lugar de financiarse con impuestos prediales locales, su financiamiento se determinaba en gran parte a nivel estatal. Sea $lSTATEALL_{it}$ el logaritmo de la asignación estatal para el distrito *i* en el año *t*, la cual es exógena en la ecuación anterior, una vez que se controlan los gastos y un efecto fijo del distrito. ¿Cómo estimaría β_j?

EJERCICIOS EN COMPUTADORA

C16.1 Use SMOKE.RAW para este ejercicio.

i) Un modelo para estimar los efectos del tabaquismo en el ingreso anual (quizás a través de ausencias laborales debido a enfermedades o efectos de productividad) es

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1,$$

donde *cigs* es el número de cigarros que se consumen al día, en promedio. ¿Cómo se interpreta β_1?

ii) Para reflejar el hecho de que el consumo de cigarros podría estar determinado conjuntamente por el ingreso, una ecuación de la demanda de cigarros es

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2,$$

donde *cigpric* es el precio de una cajetilla de cigarros (en centavos) y *restaurn* es una variable binaria igual a la unidad si la persona vive en un estado cuyos restaurantes restringen el consumo del cigarro. Si se supone que éstas son variables exógenas al individuo, ¿qué signos se esperarían para γ_5 y γ_6?

iii) ¿Con base en qué supuesto la ecuación del ingreso de la parte i) se identificó?

iv) Estime la ecuación de ingreso mediante MCO y analice la estimación de β_1.

v) Estime la forma reducida para *cigs*. (Recuerde que esto implica realizar la regresión de *cigs* sobre todas las variables exógenas.) ¿log(*cigpric*) y *restaurn* son significativas en la forma reducida?

vi) Ahora, estime la ecuación de ingreso mediante MC2E. Analice cómo se compara la estimación de β_1 con la estimación de MCO.

vii) ¿Se piensa que los precios del cigarro y las restricciones en los restaurantes para fumar son exógenas en la ecuación de ingreso?

C16.2 Use MROZ.RAW para este ejercicio.

i) Vuelva a estimar la función de la oferta de mano de obra en el ejemplo 16.5, mediante log(*hours*) como la variable dependiente. Compare la elasticidad estimada (que ahora es constante) con la estimación obtenida de la ecuación (16.24) en el promedio de horas trabajadas.

ii) En la ecuación de oferta de mano de obra de la parte i), sea *educ* endógena debido a la capacidad omitida. Use *motheduc* y *fatheduc* como VI para *educ*. Recuerde, ahora se tiene dos variables endógenas en la ecuación.

iii) Pruebe las restricciones de sobreidentificación en la estimación de MC2E de la parte ii). ¿Las VI pasan la prueba?

C16.3 Use los datos en OPENNESS.RAW para este ejercicio.

i) Ya que log(*pcinc*) es insignificante en (16.22) y también en la forma reducida para *open*, descártela del análisis. Estime (16.22) mediante MCO y VI sin log(*pcinc*). ¿Cambia alguna conclusión importante?

ii) Aun con log(*pcinc*) fuera del análisis, ¿cuál es un mejor instrumento para open: *land* o log(*land*)? (*Sugerencia*: realizar una regresión de *open* sobre cada una de ellas de manera separada y conjunta.)

iii) Ahora, vuelva a (16.22). Agregue la variable aleatoria *oil* a la ecuación y trate como exógena. Estime la ecuación mediante VI. ¿Ser un productor de petróleo tiene un efecto *ceteris paribus* en la inflación?

C16.4 Use los datos en CONSUMP.RAW para este ejercicio.

i) En el ejemplo 16.7, use el método de la sección 15.5 para probar la restricción de sobreidentificación para estimar (16.35). ¿Qué se concluye?

ii) Campbell y Mankiw (1990) usan rezagos de *segundo orden* de todas las variables como VI debido a los posibles problemas de medición de datos y rezagos en la información. Vuelva a estimar (16.35), usando sólo gc_{t-2}, gy_{t-2} y $r3_{t-2}$ como VI. ¿Cómo se comparan las estimaciones con las de (16.36)?

iii) Realice una regresión de gy_t sobre las VI de la parte ii) y pruebe si gy_t está suficientemente correlacionada con ellas. ¿Por qué es esto importante?

C16.5 Use el *Economic Report of the President* (2005 o posterior) para actualizar los datos en CONSUMP.RAW, al menos a 2003. Vuelva a estimar la ecuación (16.35). ¿Cambia alguna conclusión importante?

C16.6 Use los datos en CEMENT.RAW para este ejercicio.

i) Una función estática (inversa) de la oferta para el crecimiento mensual en el precio del cemento (*gprc*) en función del crecimiento en la cantidad (*gcem*) es

$$gprc_t = \alpha_1 gcem_t + \beta_0 + \beta_1 gprcpet + \beta_2 feb_t + \ldots + \beta_{12} dec_t + u_t^s,$$

donde *gprcpet* (aumento en el precio del petróleo) se supone exógena y *feb*, …, *dec* son variables binarias mensuales. ¿Qué signos espera para α_1 y β_1? Estime la ecuación mediante MCO. ¿La pendiente de la función de la oferta es ascendente?

ii) La variable *gdefs* es el aumento mensual en el gasto real en defensa de Estados Unidos. ¿Qué se necesita suponer acerca de *gdefs* para que sea una buena VI para *gcem*? Pruebe si *gcem* está correlacionada parcialmente con *gdefs*. (No debe preocupar una posible correlación serial en la forma reducida.) ¿Puede emplear *gdefs* como una VI para estimar la función de la oferta?

iii) Shea (1993) argumenta que el aumento en la producción de la construcción residencial (*gres*) y no residencial (*gnon*) son instrumentos válidos para *gcem*. La idea es que estos desplazadores de la demanda, en general, no deben estar correlacionados con el error de la oferta u_t^s. Pruebe si *gcem* está correlacionado parcialmente con *gres* y *gnon*; de nuevo, no debe preocupar la correlación serial en la forma reducida.

iv) Estime la función de la oferta, utilizando *gres* y *gnon* como VI para *gcem*. ¿Qué se concluye acerca de la función de la oferta estática para el cemento? [La función de la oferta dinámica tiene, en apariencia, una pendiente ascendente; vea Shea (1993).]

C16.7 Refiérase al ejemplo 13.9 y a los datos en CRIME4.RAW.

i) Suponga que, después de diferenciar para eliminar el efecto inobservado, se piensa que $\Delta\log(polpc)$ es simultáneamente determinado con $\Delta\log(crmrte)$; en particular, los aumentos en la delincuencia están asociados con aumentos en el número de policías. ¿Cómo ayuda esto a explicar el coeficiente positivo en $\Delta\log(polpc)$ en la ecuación (13.33)?

ii) La variable *taxpc* son los impuestos recaudados por persona en la demarcación. ¿Parece razonable excluirla de la ecuación del índice delictivo?

iii) Estime la forma reducida para $\Delta\log(polpc)$ mediante MCO combinado, incluida la VI potencial, $\Delta\log(taxpc)$. ¿Parece que $\Delta\log(taxpc)$ es un buen candidato para VI? Explique.

iv) Suponga que, en varios años, el estado de Carolina del Norte concedió subsidios en algunas demarcaciones para incrementar la fuerza policial en esos lugares. ¿Cómo se emplearía esta información para estimar el efecto de policías adicionales en el índice delictivo?

C16.8 Use los datos en FISH.RAW, provenientes de Graddy (1995), para este ejercicio. La base de datos también se empleó en el ejercicio para computadora C12.9. Ahora, se usará para estimar una función de la demanda del pescado.

i) Suponga que la ecuación de demanda se puede escribir, en equilibrio para cada periodo de tiempo como

$$\log(totqty_t) = \alpha_1\log(avgprc_t) + \beta_{10} + \beta_{11}mon_t + \beta_{12}tues_t + \beta_{13}wed_t + \beta_{14}thurs_t + u_{t1},$$

así que se permite que la demanda difiera a través de los días de la semana. Si se trata la variable del precio como endógena, ¿qué información adicional será necesaria para estimar de manera consistente los parámetros de la ecuación de demanda?

ii) Las variables $wave2_t$ y $wave3_t$ son medidas de las alturas de olas oceánicas en los días pasados. ¿Cuáles dos supuestos son necesarios para usar $wave2_t$ y $wave3_t$ como VI para $\log(avgprc_t)$ con el fin de estimar la ecuación de demanda?

iii) Realice una regresión de $\log(avgprc_t)$ sobre las binarias del día de la semana y las dos medidas de altura de olas oceánicas. ¿Son conjuntamente significativas $wave2_t$ y wave3_t? ¿Cuál es el valor-*p* de la prueba?

iv) Ahora, estime la ecuación de demanda por MC2E. ¿Cuál es el intervalo de confianza a 95% para la elasticidad precio de la demanda? ¿La elasticidad estimada es razonable?

v) Obtenga los residuales de MC2E, $\hat{u}_{t1}$. Agregue un solo rezago, $\hat{u}_{t-1,1}$ para estimar la ecuación de demanda por MC2E. Recuerde, $\hat{u}_{t-1,1}$ es su propio instrumento. ¿Existe evidencia de la correlación serial AR(1) en los errores de la ecuación de demanda?

vi) Dado que la ecuación de la oferta depende evidentemente de las variables de olas, ¿qué par de supuestos se necesitarían formular para estimar la elasticidad precio de la oferta?

vii) En la ecuación en la forma reducida para log(*avgprct*), ¿las binarias conjuntas del día de la semana son significativas? ¿Qué se concluye acerca de poder estimar la elasticidad de la oferta?

C16.9 En este ejercicio use los datos en AIRFARE.RAW, pero sólo para el año 1997.

i) Una simple función de la demanda para los asientos del avión en rutas para Estados Unidos es

$$\log(passen) = \beta_{10} + \alpha_1 \log(fare) + \beta_{11}\log(dist) + \beta_{12}[\log(dist)]^2 + u_1,$$

donde

$passen$ = promedio de pasajeros por día.
$fare$ = tarifa aérea promedio.
$dist$ = distancia de la ruta (en millas).

Esta es verdaderamente una función de la demanda, ¿cuál será el signo de α_1?

ii) Estime la ecuación de la parte i) mediante MCO. ¿Cuál es la elasticidad precio estimada?

iii) Considere la variable *concen*, que es la medida de la concentración del mercado. (En específico, es la participación del negocio que posee el transportista más grande.) Explicar, con palabras, qué se debe suponer para tratar *concen* como exógena en la ecuación de la demanda

iv) Ahora suponga que *concen* es exógena a la ecuación de la demanda. Estime la forma reducida para log(*fare*) y confirme que *concen* tiene un efecto positivo (parcial) sobre log(*fare*).

v) Estime la función de la demanda mediante VI. Ahora, ¿cuál es la elasticidad precio estimada de la demanda? ¿Cómo se compara esto con la estimación de MCO?

vi) Mediante las estimaciones de VI, describa cómo depende la demanda de asientos de la distancia de la ruta.

C16.10 Use toda la base de datos de panel en AIRFARE.RAW para este ejercicio. La ecuación de la demanda en un modelo de efectos inobservables de ecuaciones simultáneas es

$$\log(passen_{it}) = \theta_{t1} + \alpha_1\log(fare_{it}) + a_{i1} + u_{it1},$$

donde se absorben las variables de distancia en a_{i1}.

i) Estime la función de la demanda mediante efectos fijos y asegurándose de incluir variable binarias de años para dar cuenta de los diferentes interceptos. ¿Cuál es la elasticidad estimada?

ii) Use efectos fijos para estimar en la forma reducida

$$\log(fare_{it}) = \theta_{t2} + \pi_{21}concen_{it} + a_{i2} + v_{it2}.$$

Realice la prueba adecuada para asegurarse de que $concen_{it}$ se puede emplear como VI para log($fare_{it}$).

iii) Ahora estime la función de la demanda mediante la transformación de efectos fijos junto con VI, como en la ecuación (16.42). Ahora, ¿cuál es la elasticidad estimada? ¿Es estadísticamente significativa?

CAPÍTULO 17

Modelos de variable dependiente limitada y correcciones a la selección muestral

En el capítulo 7 se estudió el modelo de probabilidad lineal, el cual simplemente es una aplicación del modelo de regresión múltiple a una variable dependiente binaria. Una variable dependiente binaria es un ejemplo de una **variable dependiente limitada (VDL)**. Una VDL se define en sentido amplio como una variable dependiente cuyo rango de valores está restringido de forma importante. Una variable binaria asume sólo dos valores, el cero y el uno. Se han analizado otros varios ejemplos de variables dependientes limitadas: el porcentaje de participación en un plan de pensiones está entre 0 y 100, el número de veces que un individuo es arrestado en un año determinado es un entero no negativo, y el promedio de calificaciones universitarias es de entre 0 y 4.0 en la mayoría de las universidades estadounidenses.

En cierto sentido, la mayoría de las variables económicas que se desea explicar son limitadas, debido a que con frecuencia deben ser positivas. Por ejemplo, el salario por hora, el precio de la vivienda y las tasas de interés nominales deben ser mayores que cero. Pero no todas estas variables requieren un tratamiento especial. Si una variable estrictamente positiva asume muchos valores diferentes, en raras ocasiones será necesario un modelo econométrico especial. Cuando y es discreta y asume un pequeño número de valores, no tiene sentido tratarla como una variable aproximadamente continua. Por sí misma, la discrecionalidad de y no significa que los modelos lineales sean inapropiados. No obstante, como se vio en el capítulo 7 para la respuesta binaria, el modelo de probabilidad lineal tiene ciertas desventajas. En la sección 17.1, se analizan los modelos logit y probit, que superan las desventajas del MPL; la desventaja es que son más difíciles de interpretar.

En el análisis econométrico surgen otros tipos de variables dependientes limitadas, en especial cuando se está modelando el comportamiento de individuos, familias o empresas. El comportamiento optimizador suele provocar una **respuesta de solución de esquina** para una fracción no trivial de la población. Es decir, lo mejor es elegir una cantidad o valor en dólares igual a cero, por ejemplo. En un año cualquiera, un número significativo de familias hará cero contribuciones de caridad. Por tanto, las contribuciones caritativas familiares anuales tienen una distribución poblacional dispersa entre el gran rango de valores positivos, pero con una acumulación en el valor cero. Aunque un modelo lineal podría ser adecuado para capturar el valor esperado de las contribuciones de caridad, es probable que un modelo lineal lleve a predicciones negativas para algunas familias. Obtener el logaritmo natural no es posible debido a que muchas observaciones son de cero. El modelo Tobit, que se abordará en la sección 17.2, está diseñado explícitamente para modelar las variables dependientes con solución de esquina.

Otro tipo importante de VDL es una variable de conteo, que asume valores enteros no negativos. La sección 17.3 ilustra por qué los modelos de regresión Poisson son adecuados para modelar variables de conteo.

En algunos casos, se observan variables dependientes limitadas debido a la censura de datos, un tema que se presenta en la sección 17.4. El problema general de la selección muestral, donde se observa una muestra no aleatoria de la población subyacente, se analiza en la sección 17.5.

Los modelos de variable dependiente limitada se pueden usar para series de tiempo y datos de panel, pero suelen aplicarse a los datos de corte transversal. Los problemas de selección muestral suelen estar confinados a los datos de corte transversal o de panel. Este capítulo se centrará en las aplicaciones de corte transversal. En Wooldridge (2002) se presentan estos problemas en el contexto de modelos de datos de panel y ofrece mayores detalles para las aplicaciones de corte transversal y de datos de panel.

17.1 Modelos logit y probit para respuesta binaria

Estimar y utilizar el modelo de probabilidad lineal es simple, pero tiene algunas desventajas que se analizaron en la sección 7.5. Las dos desventajas más importantes son que las probabilidades ajustadas pueden ser menores que cero o mayores que uno y el efecto parcial de cualquier variable explicativa (si aparece en la ecuación en su nivel) es constante. Estas limitaciones del MPL pueden superarse si se usan **modelos de respuesta binaria** más sofisticados.

En un modelo de respuesta binaria, el interés yace principalmente en la **probabilidad de respuesta**

$$P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, \ldots, x_k), \tag{17.1}$$

donde $\mathbf{x}$ denota el conjunto total de variables explicativas. Por ejemplo, cuando y es un indicador del empleo, $\mathbf{x}$ podría contener varias características individuales como la educación, edad, estado civil y otros factores que afectan el estado del empleo, incluida una variable de indicador binario para la participación en un reciente programa de empleo.

Especificación de modelos logit y probit

En el MPL, se supone que la probabilidad de respuesta es lineal en un conjunto de parámetros, β_j; vea la ecuación (7.27). Para evitar las limitaciones del MPL, considere una clase de modelos de respuesta binaria de la forma

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \tag{17.2}$$

donde G es una función que asume valores estrictamente entre cero y uno: $0 < G(z) < 1$, para todos los números reales z. Esto asegura que las probabilidades de respuesta estimada sean estrictamente entre cero y uno. Como en los primeros capítulos, se escribe $\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$.

Se han sugerido varias funciones no lineales para la función G a fin de asegurar que las probabilidades estén entre cero y uno. Las dos que se estudiarán aquí se usan en la mayoría de las aplicaciones (junto con el MPL). En el **modelo logit**, G es la función logística:

$$G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z), \tag{17.3}$$

que está entre cero y uno para todos los números reales z. Esta es la función de distribución acumulada (fda) para una variable aleatoria logística estándar. En el **modelo probit**, G es la función de distribución acumulada normal estándar, que se expresa como una integral:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)\mathrm{d}v, \tag{17.4}$$

donde $\phi(z)$ es la densidad normal estándar

$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2). \tag{17.5}$$

Esta elección de G nuevamente asegura que (17.2) esté estrictamente entre cero y uno para todos los valores de los parámetros y las x_j.

Las funciones G en (17.3) y (17.4) son funciones crecientes. Cada una aumenta con más rapidez en $z = 0$, $G(z) \rightarrow 0$ a medida que $z \rightarrow -\infty$, y $G(z) \rightarrow 1$ a medida que $z \rightarrow \infty$. La función logística está graficada en la figura 17.1. La fda normal estándar tiene una forma muy similar a la de la fda logística.

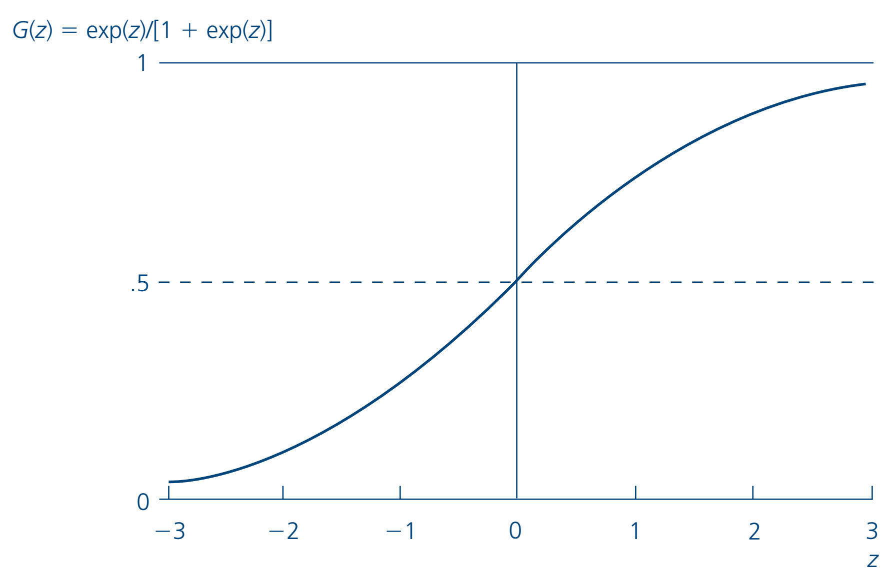
Los modelos logit y probit pueden derivarse a partir de un **modelo de variable latente** subyacente. Sea y^* una variable inobservable, o *latente*, determinada por

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e, \quad y = 1[y^* > 0], \tag{17.6}$$

donde se introduce la notación $1[\cdot]$ para definir un resultado binario. La función $1[\cdot]$ recibe el nombre de *función de indicador*, que asume el valor de uno si el evento dentro de los corchetes es verdadero y de cero si no lo es. Por tanto, y es uno si $y^* > 0$ y y es cero si $y^* \leq 0$. Se supone

FIGURA 17.1

Gráfica de la función logística $G(z) = \exp(z)/[1 + \exp(z)]$.



que e es independiente de $\mathbf{x}$ y que e tiene la distribución logística estándar o la distribución normal estándar. En cualquier caso, e se distribuye simétricamente en torno a cero, lo cual significa que $1 - G(-z) = G(z)$ para todos los números reales z. Los economistas tienden a favorecer el supuesto de normalidad para e, lo cual es la razón por la que en econometría el modelo probit es más popular que el logit. Además, varios problemas de especificación, que se tratarán después, se analizan fácilmente mediante probit debido a las propiedades de la distribución normal.

A partir de (17.6) y de los supuestos establecidos, se puede calcular la probabilidad de respuesta para y:

$$\begin{aligned} \mathrm{P}(y = 1|\mathbf{x}) &= \mathrm{P}(y^* > 0|\mathbf{x}) = \mathrm{P}[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \end{aligned}$$

lo cual es exactamente lo mismo que (17.2).

En la mayoría de las aplicaciones de los modelos de respuesta binaria, la meta principal es explicar los efectos de las x_j sobre la probabilidad de respuesta $\mathrm{P}(y = 1|\mathbf{x})$. La formulación de la variable latente tiende a dar la impresión de que lo que principalmente interesa son los efectos de cada x_j sobre y^*. Como se verá, para logit y probit, la *dirección* del efecto de x_j sobre $\mathrm{E}(y^*|\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ y sobre $\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ es siempre la misma. Pero la variable latente y^* rara vez tiene una unidad de medición bien definida. (Por ejemplo, y^* puede ser la diferencia en niveles de utilidad de dos acciones diferentes.) Por tanto, las magnitudes de cada β_j no son, por sí mismas, especialmente útiles (en contraste con el modelo de probabilidad lineal). Para la mayoría de los propósitos, se quiere estimar el efecto de x_j sobre la probabilidad de éxito $\mathrm{P}(y = 1|\mathbf{x})$, pero esto se complica por la naturaleza no lineal de $G(\cdot)$.

Para hallar el efecto parcial de las variables aproximadamente continuas sobre la probabilidad de respuesta, es necesario recurrir al cálculo. Si x_j es una variable aproximadamente continua, su efecto parcial sobre $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$ se obtiene de la derivada parcial:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j, \quad \text{donde } g(z) \equiv \frac{dG}{dz}(z). \qquad \boxed{17.7}$$

Debido a que G es la fda de una variable aleatoria continua, g es una función de densidad de probabilidad. En los casos logit y probit, $G(\cdot)$ es una fda estrictamente creciente y, por tanto $g(z) > 0$ para toda z. Por consiguiente, el efecto parcial de x_j sobre $p(\mathbf{x})$ depende de $\mathbf{x}$ a través de la cantidad positiva $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})$, lo cual significa que el efecto parcial siempre tiene el mismo signo que β_j.

La ecuación (17.7) muestra que los efectos *relativos* de cualesquiera dos variables explicativas continuas no dependen de $\mathbf{x}$: la razón de los efectos parciales de x_j y x_h es β_j/β_h. En el caso típico de que g sea una densidad simétrica en torno de cero, con una única moda en cero, el mayor efecto ocurre cuando $\beta_0 + \mathbf{x}\boldsymbol{\beta} = 0$. Por ejemplo, en el caso probit con $g(z) = \phi(z)$, $g(0) = \phi(0) = 1/\sqrt{2\pi} \approx .40$. En el caso logit, $g(z) = \exp(z)/[1 + \exp(z)]^2$, y por tanto $g(0) = .25$.

Si, por ejemplo, x_1 es una variable explicativa binaria, entonces el efecto parcial de cambiar x_1 de cero a uno, manteniendo todas las demás variables fijas, simplemente es

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_k x_k). \qquad \boxed{17.8}$$

De nuevo, esto depende de todos los valores de las otras x_j. Por ejemplo, si y es un indicador de empleo y x_1 es una variable binaria que indica la participación en un programa de capacitación laboral, entonces (17.8) es el cambio en la probabilidad del empleo debido a este programa de capacitación; esto depende de las demás características que afectan la posibilidad de obtener el empleo, como la educación y la experiencia. Observe que saber el signo de β_1 es suficiente para determinar si el programa tuvo un efecto positivo o negativo. Pero para hallar la *magnitud* del efecto, se tiene que estimar la cantidad en (17.8).

También se puede usar la diferencia en (17.8) para otros tipos de variables discretas (como el número de niños). Si x_k denota esta variable, el efecto sobre la probabilidad de que x_k cambie de c_k a $c_k + 1$ es simplemente

$$G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k(c_k + 1)] - G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k c_k). \tag{17.9}$$

Es muy fácil incluir las formas funcionales usuales de las variables explicativas. Por ejemplo, en el modelo

$$\mathrm{P}(y = 1|\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3),$$

el efecto parcial de z_1 sobre $\mathrm{P}(y = 1|\mathbf{z})$ es $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_1 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$ y el efecto parcial de z_2 sobre la probabilidad de respuesta es $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_2 = g(\beta_0 + \mathbf{x}\beta)(\beta_3/z_2)$, donde $\mathbf{x}\boldsymbol{\beta} = \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3$. Por tanto, $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/100)$ es el cambio aproximado en la probabilidad de respuesta cuando z_2 aumenta 1%.

Algunas veces se quiere calcular la elasticidad de la probabilidad de respuesta respecto a una variable explicativa, aunque se deben interpretar de manera cuidadosa los cambios porcentuales en las probabilidades. Por ejemplo, un cambio en una probabilidad de .04 a .06 representa un incremento de 2 *puntos* porcentuales en la probabilidad, pero un incremento de 50% relativo al valor inicial. Mediante el cálculo, en el modelo anterior se puede mostrar que la elasticidad de $\mathrm{P}(y = 1|\mathbf{z})$ respecto a z_2 es $\beta_3[g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. La elasticidad respecto a z_3 es $(\beta_4 z_3)[g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. En el primer caso la elasticidad siempre es del mismo signo que β_2, pero por lo general depende de todos los parámetros y todos los valores de las variables explicativas. Si $z_3 > 0$, la segunda elasticidad siempre tiene el mismo signo que el parámetro β_4.

Los modelos con interacciones entre las variables explicativas pueden ser un tanto delicados, pero se deben calcular las derivadas parciales y después evaluar los efectos parciales resultantes en valores interesantes. Cuando se miden los efectos de las variables discretas, sin importar qué tan complicado sea el modelo, se debe usar (17.9). Se analizará esto con mayor profundidad en la subsección sobre interpretación de las estimaciones de la página 580.

Estimación de máxima verosimilitud de los modelos logit y probit

¿Cómo se deben estimar los modelos no lineales de respuesta binaria? Para estimar los MPL, se pueden usar mínimos cuadrados ordinarios (vea la sección 7.5) o, en algunos casos, mínimos cuadrados ponderados (vea la sección 8.5). Debido a la naturaleza no lineal de $\mathrm{E}(y|\mathbf{x})$, MCO y MCP no son aplicables. Se podrían usar las versiones no lineales de estos métodos, pero no es más difícil usar la **estimación de máxima verosimilitud (EMV)** (vea el apéndice 17A para un breve análisis). Hasta ahora se ha tenido poca necesidad de la EMV, aunque sí se estableció que, bajo los supuestos del modelo lineal clásico, el estimador de MCO es el estimador de máxima verosimilitud (condicional en las variables explicativas). Para estimar los modelos de variables dependientes limitadas, los métodos de máxima verosimilitud son indispensables. Como la estimación de máxima verosimilitud está basada en la distribución de y dada $\mathbf{x}$, la heterocedasticidad en $\mathrm{Var}(y|\mathbf{x})$ automáticamente se toma en cuenta.

Suponga que se tiene una muestra aleatoria de tamaño n. Para obtener el estimador de máxima verosimilitud, condicional sobre las variables explicativas, se necesita la densidad de y_i dada $\mathbf{x}_i$. Esto se puede escribir como

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y}, \; y = 0, 1, \qquad \textbf{17.10}$$

donde, por simplicidad, se absorbe el intercepto en el vector $\mathbf{x}_i$. Se puede ver con facilidad que cuando $y = 1$, se obtiene $G(\mathbf{x}_i\boldsymbol{\beta})$ y cuando $y = 0$, se obtiene $1 - G(\mathbf{x}_i\boldsymbol{\beta})$. La **función de log-verosimilitud** para la observación i es una función de los parámetros y los datos $(\mathbf{x}_i, y_i)$, y se obtiene al aplicar el log a la ecuación (17.10):

$$\ell_i(\boldsymbol{\beta}) = y_i\log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]. \qquad \textbf{17.11}$$

Debido a que $G(\cdot)$ está estrictamente entre cero y uno para logit y probit, $\ell_i(\boldsymbol{\beta})$ está bien definida para todos los valores de $\boldsymbol{\beta}$.

La log-verosimilitud para un tamaño de muestra de n se obtiene al sumar (17.11) a través de todas las observaciones: $\mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$. La EMV de $\boldsymbol{\beta}$, denotada como $\hat{\boldsymbol{\beta}}$, maximiza esta log-verosimilitud. Si $G(\cdot)$ es la fda logit estándar, entonces $\hat{\boldsymbol{\beta}}$ es el *estimador logit*; si $G(\cdot)$ es la fda normal estándar, entonces $\hat{\boldsymbol{\beta}}$ es el *estimador probit*.

Debido a la naturaleza no lineal del problema de maximización, no se pueden escribir fórmulas para las estimaciones de máxima verosimilitud logit o probit. Además de los problemas computacionales, esto hace que la teoría estadística para logit y probit sea mucho más difícil que MCO o incluso que MC2E. Sin embargo, la teoría general de EMV para muestras aleatorias implica que, bajo condiciones muy generales, la EMV es consistente, asintóticamente normal y asintóticamente eficiente. [Vea Wooldridge (2002, capítulo 13) para un análisis general.] Aquí sólo se usarán los resultados; aplicar los modelos logit y probit es muy fácil, siempre y cuando se comprenda lo que significan los estadísticos.

Cada $\hat{\beta}_j$ viene con un error estándar (asintótico), cuya fórmula es complicada y se presenta en el apéndice del capítulo. Una vez que se tengan los errores estándar, y que éstos se reporten junto con las estimaciones de los coeficientes mediante cualquier paquete que soporte logit y probit, será posible construir pruebas t e intervalos de confianza (asintóticos), tal como con en MCO, MC2E y otros estimadores. En particular, para probar H_0: $\beta_j = 0$, se forma el estadístico $t = \hat{\beta}_j/\mathrm{ee}(\hat{\beta}_j)$ y se realiza la prueba en la forma usual, una vez que se ha decidido una alternativa de una o dos colas.

Prueba de hipótesis múltiples

También es posible probar restricciones múltiples en los modelos probit y logit. En la mayoría de los casos, se trata de pruebas de restricciones de exclusión múltiple, como en la sección 4.5. Esta sección se enfocará en las restricciones de exclusión.

Existen tres formas de probar las restricciones de exclusión para modelos logit y probit. El multiplicador de Lagrange o el estadístico de puntuación sólo requieren estimar el modelo bajo la hipótesis nula, tal como en el caso lineal de la sección 5.2; aquí no se abordará el estadístico de puntuación, dado que rara vez se necesita probar las restricciones de exclusión. [Vea Wooldridge (2002, capítulo 15) para otros usos de la prueba de puntuación en los modelos de respuesta binaria.]

La prueba de Wald requiere la estimación sólo del modelo no restringido. En el caso del modelo lineal, el **estadístico de Wald**, después de una simple transformación, es esencialmente el estadístico F, así que no hay necesidad de abordar el estadístico de Wald por separado. La fórmula para el estadístico de Wald se da en Wooldridge (2002, capítulo 15). Los paquetes econométricos que permiten que las restricciones de exclusión se prueben después de que se ha estimado el modelo no restringido pueden calcular este estadístico. Éste tiene una distribución ji-cuadrada asintótica, con *gl* igual al número de restricciones que se están probando.

Si tanto el modelo restringido como el no restringido son fáciles de estimar, como suele ser el caso con las restricciones de exclusión, entonces la *prueba de la razón de verosimilitudes* (*RV*) se vuelve muy atractiva. La prueba *RV* está basada en el mismo concepto que la prueba *F* en un modelo lineal. La prueba *F* mide el incremento en la suma de los residuales cuadrados cuando las variables se desechan del modelo. La prueba *RV* está basada en la diferencia en las funciones de log-verosimilitud para los modelos restringido y no restringido. La idea es la siguiente. Dado que la EMV maximiza la función de log-verosimilitud, omitir variables por lo general ocasiona una log-verosimilitud *menor*, o al menos no mayor. (Esto es similar al hecho de que la *R*-cuadrada nunca aumenta cuando las variables se omiten de una regresión.) La cuestión es si la disminución de la log-verosimilitud es lo bastante grande para concluir que las variables omitidas son importantes. Se puede tomar esta decisión una vez que se tiene un estadístico de prueba y un conjunto de valores críticos.

El **estadístico de razón de verosimilitudes** es *dos veces* la diferencia en las log-verosimilitudes:

$$RV = 2(\mathscr{L}_{nr} - \mathscr{L}_r), \tag{17.12}$$

donde $\mathscr{L}_{nr}$ es el valor de la log-verosimilitud para el modelo no restringido y $\mathscr{L}_r$ es el valor de la log-verosimilitud para el modelo restringido. Debido a que $\mathscr{L}_{nr} \geq \mathscr{L}_r$, *RV* es no negativa y suele ser estrictamente positiva. Al calcular el estadístico *RV* para modelos de respuesta binaria, es importante saber que la función de log-verosimilitud siempre es un número negativo. Esto se desprende de la ecuación (17.11), debido a que y_i es cero o uno, y ambas variables dentro de la función log están estrictamente entre cero y uno, lo cual significa que sus logaritmos naturales son negativos. Que las funciones de log-verosimilitud sean ambas negativas no cambia la forma en que se calcula el estadístico *RV*; tan sólo se preservan los signos negativos en la ecuación (17.12).

La multiplicación por dos en (17.12) es necesaria de manera que *RV* tenga una distribución *ji*-cuadrada aproximada bajo H_0. Si se están probando q restricciones de exclusión, $RV \overset{a}{\sim} \chi^2_q$. Esto significa que, para probar H_0 al nivel de 5%, se usa como valor crítico el percentil 95 en la distribución χ^2_q. Calcular los valores-*p* es fácil con la mayoría de los paquetes de software.

Pregunta 17.1

Un modelo probit para explicar si una empresa es absorbida por otra durante un año determinado es

$$\begin{aligned} \mathrm{P}(takeover = 1|\mathbf{x}) = \Phi(&\beta_0 + \beta_1 avgprof + \beta_2 mktval \\ &+ \beta_3 debtearn + \beta_4 ceoten \\ &+ \beta_5 ceosal + \beta_6 ceoage), \end{aligned}$$

donde *takeover* es una variable de respuesta binaria, *avgprof* es el margen de utilidades promedio en los años previos, *mktval* es el valor de mercado de la empresa, *debtearn* es la razón deuda a ingresos y *ceoten*, *ceosal* y *ceoage* son la antigüedad, sueldo anual y edad del director general, respectivamente. Establezca la hipótesis nula de que, manteniendo todos los demás factores igual, las variables del director general no tienen efecto sobre la probabilidad de absorción. ¿Cuántos *gl* hay en la distribución ji-cuadrada para la prueba *RV* o la prueba de Wald?

Interpretación de las estimaciones logit y probit

Dadas las computadoras modernas, desde una perspectiva práctica el aspecto más difícil de los modelos logit y probit es presentar e interpretar los resultados. Las estimaciones de coeficientes, sus errores estándar y el valor de la función de log-verosimilitud se pueden obtener mediante todos los paquetes de software que realicen logit y probit, y se deben reportar en cualquier aplicación. Los coeficientes dan los signos de los efectos parciales de cada x_j sobre la probabilidad de respuesta y la significancia estadística de x_j está determinada por si se puede rechazar H_0: $\beta_j = 0$ a un nivel de significancia suficientemente pequeño.

Como se analizó brevemente en la sección 7.5 para el modelo de probabilidad lineal, se puede calcular una medida de la bondad de ajuste llamada **porcentaje correctamente predicho**. Como antes, se define un predictor binario de y_i como uno si la probabilidad predicha es de al menos .5, y cero en caso contrario. En términos matemáticos, $\tilde{y}_i = 1$ si $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .5$ y $\tilde{y}_i = 0$ si $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) < .5$. Dada $\{\tilde{y}_i: i = 1, 2, ..., n\}$, se puede ver qué tan bien predice $\tilde{y}_i$ a y_i a través de todas las observaciones. Hay cuatro resultados posibles en cada par, $(y_i, \tilde{y}_i)$; cuando ambos son cero o ambos son uno, se hace la predicción correcta. En los dos casos en que un componente del par es cero y el otro es uno, la predicción es incorrecta. El porcentaje predicho correctamente es el porcentaje de veces en que $\tilde{y}_i = y_i$.

Aunque el porcentaje predicho correctamente es útil como una medida de la bondad de ajuste, puede ser confuso. En particular, es posible obtener porcentajes muy altos predichos con precisión aun cuando el resultado menos probable esté predicho de manera muy deficiente. Por ejemplo, suponga que $n = 200$, 160 observaciones tienen $y_i = 0$, y de estas 160 observaciones, 140 de las $\tilde{y}_i$ son también cero (así que se predice correctamente 87.5% de los resultados cero). Aunque *ninguna* de las predicciones sea correcta cuando $y_i = 1$, aún se predicen con precisión 70% de todos los resultados (140/200 = .70). Con frecuencia, se espera tener alguna capacidad de predecir el resultado menos probable (por ejemplo, si alguien es arrestado por cometer un delito) y así se debe ser transparente acerca de qué tan bien se predice cada resultado. Por tanto, es lógico también calcular el porcentaje predicho correctamente para cada uno de los resultados. El problema 17.1 le pide que demuestre que el porcentaje general predicho correctamente es un promedio ponderado de $\hat{q}_0$ (el porcentaje predicho para $y_i = 0$) y $\hat{q}_1$ (el porcentaje predicho para $y_i = 1$), donde las ponderaciones son las proporciones de ceros y de unos en la muestra, respectivamente.

Algunos han criticado la regla de predicción que se acaba de describir por usar un valor umbral de .5, en especial cuando uno de los resultados es improbable. Por ejemplo, si $\bar{y} = .08$ (sólo 8% de "éxitos" en la muestra) podría ser que *nunca* se prediga $y_i = 1$ debido a que la probabilidad estimada de éxito nunca es mayor que .5. Una alternativa es usar la fracción de éxitos en la muestra como el umbral: .08 en el ejemplo anterior. En otras palabras, definir $\tilde{y}_i = 1$ cuando $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .08$ y cero de otra manera. Mediante esta regla seguramente se incrementa el número de éxitos predichos, pero no sin costo: necesariamente se cometerán más errores, quizá muchos más, en predecir ceros ("fallas"). En términos del porcentaje general predicho correctamente, el desempeño puede ser peor que si se usara el umbral de .5.

Una tercera posibilidad es elegir el umbral de tal manera que la fracción de $\tilde{y}_i = 1$ en la muestra sea la misma que (o muy cercana a) $\bar{y}$. En otras palabras, buscar a través de valores umbral τ, $0 < \tau < 1$, tal que si se define $\tilde{y}_i = 1$ cuando $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq \tau$, entonces $\sum_{i=1}^{n} \tilde{y}_i \approx \sum_{i=1}^{n} y_i$. (La prueba y error requeridos para encontrar el valor deseado de τ puede ser tedioso pero es factible. En algunos casos no será posible hacer que el número de éxitos predicho sea exactamente el mismo que el número de éxitos en la muestra.) Ahora, dado este conjunto de $\tilde{y}_i$, se puede calcular el porcentaje correctamente predicho para cada uno de los dos resultados así como el porcentaje general predicho correctamente.

Existen varias medidas de **pseudo *R*-cuadradas** para la respuesta binaria. McFadden (1974) sugiere la medida $1 - \mathcal{L}_{nr}/\mathcal{L}_o$, donde $\mathcal{L}_{nr}$ es la función de log-verosimilitud para el modelo estimado, y $\mathcal{L}_o$ es la función de probabilidad log en el modelo con sólo un intercepto. ¿Por qué es lógica esta medida? Recuerde que las log-verosimilitudes son negativas y, por tanto $\mathcal{L}_{nr}/\mathcal{L}_o = |\mathcal{L}_{nr}|/|\mathcal{L}_o|$. Además, $|\mathcal{L}_{nr}| \leq |\mathcal{L}_o|$. Si las covariadas no tienen poder explicativo, entonces $\mathcal{L}_{nr}/\mathcal{L}_o = 1$, y la pseudo *R*-cuadrada es cero, tal como la *R*-cuadrada usual es cero en una regresión lineal cuando las covariadas no tienen poder explicativo. Por lo general, $|\mathcal{L}_{nr}| < |\mathcal{L}_o|$, en cuyo caso $1 - \mathcal{L}_{nr}/\mathcal{L}_o > 0$. Si $\mathcal{L}_{nr}$ fuera cero, la pseudo R-cuadrada sería igual a la unidad. De hecho, $\mathcal{L}_{nr}$ no puede llegar a cero en un modelo probit o logit, ya que esto requeriría que las probabilidades estimadas cuando $y_i = 1$ fueran iguales a la unidad y que las probabilidades estimadas cuando $y_i = 0$ fueran todas iguales a cero.

Las pseudo R-cuadradas alternas para probit y logit están relacionadas más directamente con la R-cuadrada usual de la estimación de MCO de un modelo de probabilidad lineal. Para probit o logit, sean $\hat{y}_i = G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ las probabilidades ajustadas. Dado que estas probabilidades también son estimaciones de $\mathrm{E}(y_i|\mathbf{x}_i)$, se puede basar una R-cuadrada en qué tan cerca están las $\hat{y}_i$ de las y_i. Una posibilidad que se sugiere del análisis de regresión estándar es calcular la correlación cuadrada entre y_i y $\hat{y}_i$. Recuerde, en un marco de regresión lineal, esta es una forma algebraica equivalente de obtener la R-cuadrada usual; vea la ecuación (3.29). Por tanto, se puede calcular una pseudo R-cuadrada para probit y logit que sea directamente comparable con la R-cuadrada usual a partir de la estimación de un modelo de probabilidad lineal. En cualquier caso, la bondad de ajuste suele ser menos importante que intentar obtener estimaciones convincentes de los efectos *ceteris paribus* de las variables explicativas.

Con frecuencia se desean estimar los efectos de las x_j sobre las probabilidades de respuesta, $\mathrm{P}(y = 1|\mathbf{x})$. Si x_j es (aproximadamente) continua, entonces

$$\Delta\hat{\mathrm{P}}(y = 1|\mathbf{x}) \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j, \tag{17.13}$$

para cambios "pequeños" en x_j. Así que, para $\Delta x_j = 1$, el cambio en la probabilidad estimada de éxito es aproximadamente $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j$. En comparación con el modelo de probabilidad lineal, el costo de usar los modelos probit y logit es que los efectos parciales en la ecuación (17.13) son más difíciles de resumir debido a que el factor de escala, $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$, depende de $\mathbf{x}$ (es decir, de todas las variables explicativas). Una posibilidad es insertar valores interesantes para las x_j —como las medias, las medianas, los mínimos, los máximos y los cuartiles superiores e inferiores— y ver cómo cambia $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$. Aunque atractivo, éste puede ser un proceso tedioso y dar como resultado demasiada información aun si el número de variables explicativas es moderado.

Como resumen rápido para obtener las magnitudes de los efectos parciales, es útil tener un factor escalar único que se pueda usar para multiplicar cada $\hat{\beta}_j$ (o al menos aquellos coeficientes de variables aproximadamente continuas). Un método que suele usarse en los paquetes econométricos que estiman de forma rutinaria los modelos probit y logit, es remplazar cada variable explicativa con su promedio muestral. En otras palabras, el factor de ajuste es

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k), \tag{17.14}$$

donde $g(\cdot)$ es la densidad normal estándar en el caso probit $g(z) = \exp(z)/[1 + \exp(z)]^2$ en el caso logit. La idea en que se basa (17.14) es que cuando ésta se multiplica por $\hat{\beta}_j$, se obtiene el efecto parcial de x_j para la persona "promedio" en la muestra. Por tanto, si se multiplica un coeficiente por (17.14), por lo general se obtiene el **efecto parcial en el promedio (EPeP)**.

Existen al menos dos posibles problemas con el uso de los EPeP para resumir los efectos parciales de las variables explicativas. Primero, si algunas de las variables explicativas son discretas, sus promedios no representan a nadie en la muestra (o población, en tal caso). Por ejemplo, si $x_1 = mujeres$ y 47.5% de la muestra son mujeres, ¿qué sentido tiene insertar $\bar{x}_1 = .475$ para representar a la persona "promedio"? Segundo, si una variable explicativa continua aparece como una función no lineal, por ejemplo, como un log natural o en una cuadrática, no es claro si se quiere promediar la función no lineal o insertar el promedio en la función no lineal. Por ejemplo ¿Se debe usar $\overline{\log(sales)}$ o $\log(\overline{sales})$ para representar el tamaño promedio de una empresa? Los paquetes econométricos que calculan el factor escalar en (17.14) se quedan en el primero: el paquete está programado para calcular los promedios de los regresores incluidos en la estimación probit o logit.

Un método diferente para calcular un factor escalar elude la cuestión de qué valores insertar para las variables explicativas. En lugar de ello, el segundo factor escalar resulta al promediar los efectos parciales individuales a través de la muestra, lo que genera el algunas veces llamado **efecto parcial promedio (EPP)**. Para una variable explicativa continua x_j, el efecto parcial promedio es $n^{-1}\sum_{i=1}^{n}[g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\hat{\beta}_j] = \left[n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\right]\hat{\beta}_j$. El término que multiplica a $\hat{\beta}_j$ actúa como un factor escalar:

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}). \tag{17.15}$$

La ecuación (17.15) se calcula fácilmente después de la estimación probit o logit, donde $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \phi(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ en el caso probit y $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})]^2$ en el caso logit. Los dos factores escalares difieren, y posiblemente son muy diferentes, debido a que en (17.15) se usa el promedio de la función no lineal en lugar de la función no lineal en el promedio [como en (17.14)].

Debido a que los dos factores escalares que se acaban de describir dependen de la aproximación de cálculo en (17.13), ninguna es lógica para las variables explicativas discretas. En lugar de esto, es mejor usar la ecuación (17.9) para estimar directamente el cambio en la probabilidad. Para un cambio en x_k de c_k a $c_k + 1$, el análogo discreto del efecto parcial basado en (17.14) es

$$\begin{aligned} &G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k(c_k + 1)] \\ &\quad - G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k c_k), \end{aligned} \tag{17.16}$$

donde G es la fda normal estándar en el caso probit y $G(z) = \exp(z)/[1 + \exp(z)]$ en el caso logit. [Para x_k binaria, (17.16) suele calcularse mediante ciertos paquetes econométricos, como Stata®.] El efecto parcial promedio, que suele ser más comparable con las estimaciones del MPL, es

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\{&G[\hat{\beta}_0 + \hat{\beta}_1x_{i1} + \ldots + \hat{\beta}_{k-1}x_{ik-1} + \hat{\beta}_k(c_k + 1)] \\ &- G(\hat{\beta}_0 + \hat{\beta}_1x_{i1} + \ldots + \hat{\beta}_{k-1}x_{ik-1} + \hat{\beta}_k c_k)\}. \end{aligned} \tag{17.17}$$

Obtener la expresión (17.17) para probit o logit en realidad es muy fácil. Primero, para cada observación, se estima la probabilidad de éxito para los dos valores elegidos de x_k, insertando los resultados reales para las demás variables explicativas. (Así, se tendrían n diferencias estimadas.) Después, se promedian las diferencias en las probabilidades estimadas a través de todas las observaciones.

La expresión en (17.17) tiene una interpretación particularmente útil cuando x_k es una variable binaria. Para cada unidad i, se estima la diferencia predicha en la probabilidad de que $y_i = 1$ cuando $x_k = 1$ y $x_k = 0$, es decir,

$$G(\hat{\beta}_0 + \hat{\beta}_1x_{i1} + \ldots + \hat{\beta}_{k-1}x_{i,k-1} + \hat{\beta}_k) - G(\hat{\beta}_0 + \hat{\beta}_1x_{i1} + \ldots + \hat{\beta}_{k-1}x_{i,k-1}).$$

Para cada i, esta diferencia es el efecto estimado de cambiar x_k de cero a uno, ya sea que la unidad i tenga $x_{ik} = 1$ o $x_{ik} = 0$. Por ejemplo, si y es un indicador del empleo (igual a uno si la persona se contrata) después de la participación en un programa de capacitación, indicado por x_k, entonces se puede estimar la diferencia en las probabilidades de empleo para cada persona en ambas situaciones. Este *razonamiento contrafactual* es similar al indicado en el capítulo 16, el cual se usó para motivar los modelos de ecuaciones simultáneas. El efecto estimado del programa de capacitación laboral sobre la probabilidad de empleo es el promedio de las diferencias estimadas en las probabilidades. En otro ejemplo, suponga que y indica si la solicitud para hipoteca que hizo una familia se aprueba y que x_k es un indicador racial binario (por ejemplo, igual a uno para no blancos). Entonces, para cada familia se puede estimar la diferencia predicha en cuanto a que se apruebe la solicitud de hipoteca en función del ingreso, riqueza, calificaciones crediticias, etc., los cuales serían los elementos de $(x_{i1}, x_{i2}, ..., x_{i,k-1})$ en los dos escenarios en que el jefe de familia sea no blanco frente a que sea blanco. Se espera que se hayan contemplado los suficientes factores de manera que el promedio de las diferencias en las probabilidades genere una estimación convincente del efecto racial.

En las aplicaciones donde se aplica probit, logit y el MPL, es lógico calcular los factores escalares analizados antes para probit y logit al hacer comparaciones de efectos parciales. Sin embargo, algunas veces es necesaria una forma más rápida de comparar las magnitudes de las diferentes estimaciones. Como se mencionó antes, para probit $g(0) \approx .4$ y para logit, $g(0) = .25$. Por tanto, para hacer que las magnitudes de probit y logit sean aproximadamente comparables, se pueden multiplicar los coeficientes probit por $.4/.25 = 1.6$, o se pueden multiplicar las estimaciones logit por .625. En el MPL, $g(0)$ es efectivamente uno, así que las estimaciones de la pendiente logit se pueden dividir entre cuatro para hacerlas comparables a las estimaciones del MPL; las estimaciones de la pendiente probit se pueden dividir entre 2.5, para hacerlas comparables con las estimaciones MPL. Sin embargo, en la mayoría de los casos, se desean comparaciones más precisas obtenidas mediante el uso de factores escalares en (17.15) para logit y probit.

Ejemplo 17.1

[Participación en la fuerza laboral de las mujeres casadas]

Ahora se usarán los datos en MROZ.RAW para estimar el modelo de participación laboral del ejemplo 8.8 (vea también la sección 7.5), mediante logit y probit. También se reportan las estimaciones del modelo de probabilidad lineal del ejemplo 8.8, mediante los errores estándar robustos a la heterocedasticidad. Los resultados con los errores estándar entre paréntesis, se dan en la tabla 17.1.

Las estimaciones de los tres modelos revelan una historia consistente. Los signos de los coeficientes son los mismos entre todos los modelos y las mismas variables son estadísticamente significativas en cada modelo. La pseudo R-cuadrada para el MPL es la R-cuadrada usual de MCO; para logit y probit, la pseudo R-cuadrada es la medida basada en las log-verosimilitudes descritas antes.

Como ya se ha enfatizado, las *magnitudes* de las estimaciones de los coeficientes a través de los modelos no son comparables directamente. En lugar de ello, se calculan los factores escalares en las ecuaciones (17.14) y (17.15). Si se evalúa la función de densidad de probabilidad normal estándar $\phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k)$ en los promedios muestrales de las variables explicativas (incluido el promedio de *exper*^{2}, *kidslt6* y *kidsge6*), el resultado será aproximadamente .391. Cuando se calcula (17.14) para el caso logit, se obtiene alrededor de .243. La razón de éstos, $.391/.243 \approx 1.61$, es muy cercana a la simple regla general de aumentar las estimaciones probit para hacerlas comparables con las estimaciones

Pregunta 17.2

Mediante las estimaciones probit y la aproximación de cálculo, ¿cuál es el cambio aproximado en la probabilidad de respuesta cuando *exper* aumenta de 10 a 11?

TABLA 17.1

Estimaciones MPL, logit y probit de la participación en la fuerza laboral

Variable dependiente: *inlf*			
Variables independientes	**MPL (MCO)**	**Logit (EMV)**	**Probit (EMV)**
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
exper^2	−.00060 (.00018)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (.204)	−.868 (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constante	.586 (.151)	.425 (.860)	.270 (.509)
Porcentaje predicho correctamente Valor de log-verosimilitud Pseudo R-cuadrada	73.4 — .264	73.6 −401.77 .220	73.4 −401.30 .221

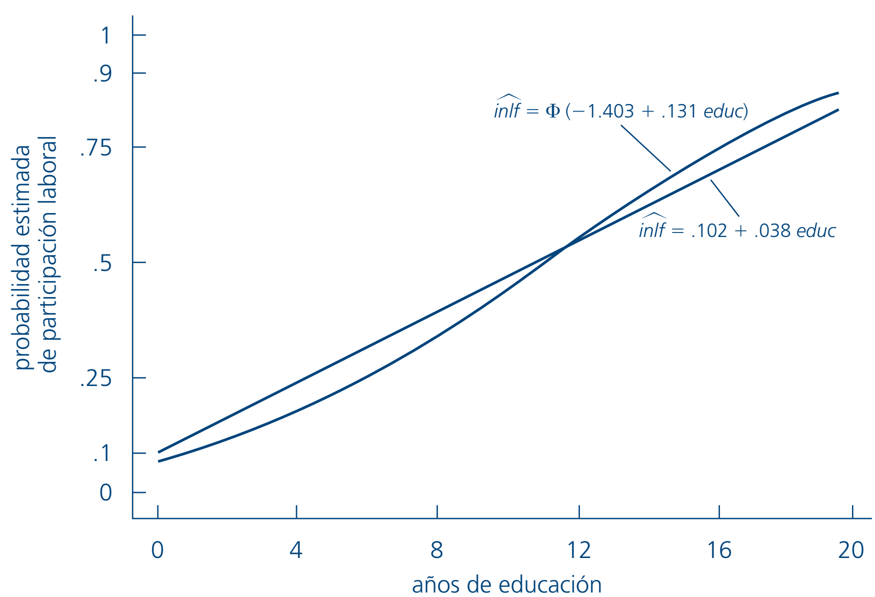
logit: multiplicar las estimaciones probit por 1.6. Sin embargo, para comparar las estimaciones logit y probit con las del MPL, es mejor usar (17.15). Estos factores escalares son de cerca de .301 (probit) y .179 (logit). Por ejemplo, el coeficiente escalado logit de *educ* es de cerca de .179(.221) ≈ .040, y el coeficiente escalado probit de *educ* es de alrededor de .301(.131) ≈ .039; ambos son extraordinariamente cercanos a la estimación del MPL de .038. Incluso en la variable discreta *kidslt6*, los coeficientes escalados logit y probit son similares al coeficiente del MPL de −.262. Éstos son .179(−1.443) ≈ −.258 (logit) y .301(−.868) ≈ −.261 (probit).

La mayor diferencia entre el modelo MPL y los modelos logit y probit es que el MPL supone efectos marginales *constantes* para *educ*, *kidslt6*, etc., mientras que los modelos logit y probit implican magnitudes decrecientes de los efectos parciales. En el MPL se estima que un niño pequeño adicional reduce la probabilidad de participación en la fuerza laboral por aproximadamente .262, sin importar cuántos niños pequeños tenga ya la mujer (y sin importar los niveles de otras variables explicativas). Se puede contrastar esto con el efecto marginal estimado de probit. Específicamente, tome el ejemplo de una mujer con *nwifeinc* = 20.13, *educ* = 12.3, *exper* = 10.6 y *age* = 42.5, que son aproximadamente los promedios muestrales, y *kidsge6* = 1. ¿Cuál es el decremento estimado en la probabilidad de trabajar al pasar de cero niños a un niño pequeño? Se evalúa la fda normal estándar, $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)$, con *kidslt6* = 1 y *kidslt6* = 0, y las otras variables independientes establecidas en los valores anteriores. Se obtiene aproximadamente .373 − .707 = −.334, lo cual significa que la probabilidad de participación en la fuerza laboral es aproximadamente .334 menor cuando una mujer tiene un niño pequeño. Si la mujer pasa de uno a dos niños pequeños, la probabilidad disminuye aún más, pero el efecto marginal no es tan grande: .117 − .373 = −.256. Es interesante observar que la estimación del modelo de probabilidad lineal, que se supone estima el efecto cercano al promedio, en realidad está entre estas dos estimaciones.

La figura 17.2 ilustra cómo pueden diferir las probabilidades de respuesta estimada de los modelos no lineales de respuesta binaria del modelo de probabilidad lineal. La probabilidad estimada de la participación en la fuerza laboral se grafica respecto a los años de educación para el modelo de probabilidad lineal y el modelo probit. (La gráfica para el modelo logit es muy similar a la del modelo probit.) En ambos casos, las variables explicativas, diferentes de *educ,* se establecen en sus promedios muestrales. En particular, las dos ecuaciones graficadas son $\widehat{inlf} = .102 + .038\, educ$ para el modelo lineal y $\widehat{inlf} = \Phi(-1.403 + .131\, educ)$. En los niveles más bajos de educación, el modelo de probabilidad lineal estima mayores probabilidades de participación laboral que el modelo probit. Por ejemplo, con ocho años de educación, el modelo de probabilidad lineal estima una probabilidad de participación laboral de .406 mientras que el modelo probit estima cerca de .361. Las estimaciones son las mismas en torno a los 11⅓ años de educación. A niveles más altos de educación, el modelo probit da mayores probabilidades de

FIGURA 17.2

Probabilidades estimadas de respuesta respecto a la educación para los modelos de probabilidad lineal y probit.



participación laboral. En esta muestra, la menor cantidad de años de educación es 5 y la mayor es 17, así que en realidad no se deben hacer comparaciones fuera de este rango.

Las mismas cuestiones concernientes a las variables explicativas endógenas en los modelos lineales también surgen en los modelos probit y logit, las cuales no se cubrirán, pero es posible probar y corregir las variables explicativas endógenas mediante métodos relacionados con los mínimos cuadrados en dos etapas. Evans y Schwab (1995) estimaron un modelo probit para si un estudiante asiste a la universidad, donde la variable explicativa clave es una variable binaria para si el estudiante asiste a una escuela católica. Evans y Schwab estimaron un modelo mediante máxima verosimilitud que permite que asistir a una escuela católica se considere endógeno. [Vea Wooldridge (2002, capítulo 15) para una explicación de estos métodos.]

En el contexto de los modelos probit, otras dos cuestiones han recibido atención. La primera es la no normalidad de e en el modelo de variable latente (17.6). Naturalmente, si e no tiene una distribución normal estándar, la probabilidad de respuesta no tendrá la forma probit. Algunos autores tienden a enfatizar la inconsistencia en estimar β_j, pero este es el enfoque equivocado a menos que sólo interese la dirección de los efectos. Debido a que la probabilidad de respuesta se ignora, no se podría estimar la magnitud de los efectos parciales aunque se tuvieran estimaciones consistentes de β_j.

Un segundo problema de especificación, también definido en términos del modelo de variable latente, es la heterocedasticidad en e. Si $\text{Var}(e|\mathbf{x})$ depende de $\mathbf{x}$, la probabilidad de respuesta ya no tiene la forma $G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$; en lugar de ello, depende de la forma de la varianza y requiere una estimación más general. Tales modelos suelen no usarse en la práctica, dado que logit y probit con las formas funcionales flexibles en las variables independientes tienden a funcionar bien.

Los modelos de respuesta binaria aplican con pocas modificaciones a la combinación independiente de cortes transversales o a otros conjuntos de datos donde las observaciones están distribuidas de manera independiente, pero no necesariamente idéntica. Con frecuencia, las variables binarias anuales o de otro periodo se incluyen para representar los efectos agregados del tiempo. Tal como en los modelos lineales, logit y probit se pueden usar para evaluar el impacto de ciertas políticas en el contexto de un experimento natural.

El modelo de probabilidad lineal se puede aplicar a los datos de panel; por lo general se estimaría mediante los efectos fijos (vea el capítulo 14). Los modelos logit y probit con efectos inobservables recién se han popularizado. Estos modelos se complican por la naturaleza no lineal de las probabilidades de respuesta y son difíciles de estimar e interpretar. [Vea Wooldridge (2002, capítulo 15).]

17.2 Modelo Tobit para respuestas de solución de esquina

Como se mencionó en la introducción del capítulo, otro tipo importante de variable dependiente limitada es una respuesta de solución de esquina. Tal variable es cero para una fracción no trivial de la población, pero tiene una distribución aproximadamente continua a través de valores positivos. Un ejemplo es la cantidad que un individuo gasta en alcohol en un mes determinado. En la población de personas de más de 21 años en Estados Unidos, esta variable asume un amplio rango de valores. Para alguna fracción importante, la cantidad gastada es de cero. El siguiente tratamiento omite la verificación de algunos detalles concernientes al modelo Tobit. [Estos se dan en Wooldridge (2002, capítulo 16).]

Sea y una variable esencialmente continua a través de valores estrictamente positivos pero que asume un cero con probabilidad positiva. Nada impide que se use un modelo lineal para y. De hecho, un modelo lineal podría ser una buena aproximación a $E(y|x_1, x_2, \ldots, x_k)$, en especial para x_j cerca de los valores promedio. Pero se podrían obtener valores ajustados negativos, lo cual generaría predicciones negativas para y; esto es análogo a los problemas del MPL para variables binarias. Por otra parte, el supuesto de que una variable explicativa que aparece en la forma de nivel tiene un efecto parcial constante sobre $\mathrm{E}(y|\mathbf{x})$ puede ser engañoso. Probablemente, $\mathrm{Var}(y|\mathbf{x})$ sería heterocedástica, aunque se puede tratar fácilmente la heterocedasticidad general mediante el cálculo de los errores estándar y los estadísticos de prueba robustos. Debido a que la distribución de y se acumula en cero, es claro que y no puede tener una distribución normal condicional. Así que toda la inferencia tendría sólo una justificación asintótica, como con el modelo de probabilidad lineal.

En algunos casos, es importante tener un modelo que implique valores predichos no negativos para y, y que tenga efectos parciales sensatos sobre un amplio rango de las variables explicativas. Además, algunas veces es necesario estimar las características de la distribucuión de y dadas $x_1, ..., x_k$ más allá de la expectativa condicional. El **modelo Tobit** es muy conveniente para estos propósitos. Por lo general, el modelo Tobit expresa la respuesta observada, y, en términos de una variable latente subyacente:

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \; u|\mathbf{x} \sim \mathrm{Normal}(0, \sigma^2) \tag{17.18}$$

$$y = \max(0, y^*). \tag{17.19}$$

La variable latente y^* satisface los supuestos del modelo lineal clásico; en particular, tiene una distribución normal, homocedástica con una media condicional lineal. La ecuación (17.19) implica que la variable observable, y, es igual a y^* cuando $y^* \geq 0$, pero $y = 0$ cuando $y^* < 0$. Debido a que y^* se distribuye normalmente, y tiene una distribución continua a través de valores estrictamente positivos. En particular, la densidad de y dada $\mathbf{x}$ es la misma que la densidad de y^* dada $\mathbf{x}$ para valores positivos. Además,

$$\begin{aligned} \mathrm{P}(y = 0|\mathbf{x}) &= \mathrm{P}(y^* < 0|\mathbf{x}) = \mathrm{P}(u < -\mathbf{x}\boldsymbol{\beta}|\mathbf{x}) \\ &= \mathrm{P}(u/\sigma < -\mathbf{x}\boldsymbol{\beta}/\sigma|\mathbf{x}) = \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \end{aligned}$$

debido a que u/σ tiene una distribución normal estándar y es independiente de $\mathbf{x}$; se ha absorbido el intercepto en $\mathbf{x}$ por simplicidad notacional. Por tanto, si $(\mathbf{x}_i, y_i)$ se extraen aleatoriamente de la población, la densidad de y_i dada $\mathbf{x}_i$ es

$$(2\pi\sigma^2)^{-1/2}\exp[-(y - \mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)] = (1/\sigma)\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \; y > 0 \tag{17.20}$$

$$\mathrm{P}(y_i = 0|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma), \tag{17.21}$$

donde ϕ es la función de densidad normal estándar.

De (17.20) y (17.21) se obtiene la función de log-verosimilitud para cada observación i:

$$\begin{aligned} \ell_i(\boldsymbol{\beta},\sigma) = {} & 1(y_i = 0)\log[1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)] \\ & + 1(y_i > 0)\log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}; \end{aligned} \tag{17.22}$$

observe cómo esto depende de σ, la desviación estándar de u, así como de las β_j. La log-verosimilitud para una muestra aleatoria de tamaño n se obtiene al sumar (17.22) a través de toda i. Las estimaciones de máxima verosimilitud de $\boldsymbol{\beta}$ y σ se obtienen al maximizar la log-verosimilitud; esto requiere métodos numéricos, aunque en la mayoría de los casos esto se realiza con facilidad mediante una rutina de software.

> **Pregunta 17.3**
>
> Sea y el número de aventuras extramaritales para una mujer casada de la población estadounidense; nos gustaría explicar esta variable en términos de otras características de la mujer (en particular, si trabaja fuera del hogar), de su esposo y su familia. ¿Es este caso un buen candidato para el modelo Tobit?

Como en el caso de logit y probit, cada estimación Tobit se acompaña con un error estándar, que se puede usar para construir estadísticos t para cada $\hat{\beta}_j$; la fórmula matricial que se usa para hallar los errores estándar es complicada y no se presentará aquí. [Vea, por ejemplo, Wooldridge (2002, capítulo 16).]

Probar restricciones de exclusión múltiples es fácil de hacer al usar la prueba de Wald o la prueba de razón de verosimilitudes. La prueba de Wald tiene una forma similar en el caso logit o probit; la prueba *RV* siempre la da (17.12), donde, por supuesto, se usan las funciones de log-verosimilitud Tobit para los modelos restringido y no restringido.

Interpretación de las estimaciones Tobit

Mediante las computadoras modernas, las estimaciones de máxima verosimilitud para los modelos Tobit suelen no ser mucho más difíciles de obtener que las estimaciones de MCO para un modelo lineal. Además, los resultados de Tobit y MCO casi siempre son similares. Esto hace que sea tentador interpretar las $\hat{\beta}_j$ de Tobit como si fueran estimaciones de una regresión lineal. Por desgracia, las cosas no son tan fáciles.

De la ecuación (17.18) se puede ver que las β_j miden los efectos parciales de las x_j sobre $\mathrm{E}(y^*|\mathbf{x})$, donde y^* es la variable latente. Algunas veces, y^* tiene un significado económico interesante, pero con más frecuencia no es así. La variable que se quiere explicar es y, puesto que es el resultado observado (como las horas trabajadas o la cantidad de contribuciones caritativas). Por ejemplo, en cuestión de políticas, nos interesa la sensibilidad de las horas trabajadas a los cambios en las tasas marginales de impuestos.

Se puede estimar $\mathrm{P}(y = 0|\mathbf{x})$ de (17.21), lo cual, por supuesto, nos permite estimar $\mathrm{P}(y > 0|\mathbf{x})$. ¿Qué sucede si se quiere estimar el valor esperado de y en función de $\mathbf{x}$? En los modelos Tobit, dos expectativas son de interés particular: $\mathrm{E}(y|y > 0,\mathbf{x})$, lo cual en ocasiones recibe el nombre de "expectativa condicional" debido a que es condicional sobre $y > 0$, y $\mathrm{E}(y|\mathbf{x})$, que es, por desgracia, llamada la "expectativa no condicional". (Ambas expectativas son condicionales sobre las variables explicativas.) La expectativa $\mathrm{E}(y|y > 0,\mathbf{x})$ nos dice, para los valores dados de $\mathbf{x}$, el valor esperado de y para la subpoblación donde y es positivo. Dada $\mathrm{E}(y|y > 0,\mathbf{x})$, se puede hallar con facilidad $\mathrm{E}(y|\mathbf{x})$:

$$\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y > 0|\mathbf{x})\cdot\mathrm{E}(y|y > 0,\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\cdot\mathrm{E}(y|y > 0,\mathbf{x}). \quad \boxed{\mathbf{17.23}}$$

Para obtener $\mathrm{E}(y|y > 0,\mathbf{x})$, se puede usar un resultado para las variables aleatorias con distribución normal: si $z \sim \mathrm{Normal}(0,1)$, entonces $\mathrm{E}(z|z > c) = \phi(c)/[1 - \Phi(c)]$ para cualquier c constante. Pero $\mathrm{E}(y|y > 0, \mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|u > -\mathbf{x}\boldsymbol{\beta}) = \mathbf{x}\boldsymbol{\beta} + \sigma\mathrm{E}[(u/\sigma)|(u/\sigma) > -\mathbf{x}\boldsymbol{\beta}/\sigma] = \mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$, debido a que $\phi(-c) = \phi(c)$, $1 - \Phi(-c) = \Phi(c)$, y u/σ tiene una distribución normal estándar independiente de $\mathbf{x}$.

Esto se puede resumir como

$$\mathrm{E}(y|y > 0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma), \quad \boxed{\mathbf{17.24}}$$

donde $\lambda(c) = \phi(c)/\Phi(c)$ recibe el nombre de **razón inversa de Mills**; esta es la razón entre la fdp normal estándar y fda normal estándar, cada una evaluada en c.

La ecuación (17.24) es importante. Muestra que el valor esperado de y condicional sobre $y > 0$ es igual a $\mathbf{x}\boldsymbol{\beta}$ más un término estrictamente positivo, el cual es σ veces la razón inversa de Mills evaluada en $\mathbf{x}\boldsymbol{\beta}/\sigma$. Esta ecuación también muestra por qué usar MCO sólo para observaciones donde $y_i > 0$ no siempre estima consistentemente $\boldsymbol{\beta}$; en esencia, la razón inversa de Mills es una variable omitida y por lo general está correlacionada con los elementos de $\mathbf{x}$.

Al combinar (17.23) y (17.24) se tiene

$$\mathrm{E}(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)] = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \tag{17.25}$$

donde la segunda igualdad se debe a que $\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma) = \phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$. Esta ecuación muestra que cuando y sigue un modelo Tobit, $\mathrm{E}(y|\mathbf{x})$ es una función no lineal de $\mathbf{x}$ *y también* de $\boldsymbol{\beta}$. Aunque esto no es evidente, el lado derecho de la ecuación (17.25) puede mostrarse positivo para cualquier valor de $\mathbf{x}$ y $\boldsymbol{\beta}$. Por tanto, una vez que se tienen las estimaciones de $\boldsymbol{\beta}$, se puede asegurar que los valores predichos para y, es decir, las estimaciones de $\mathrm{E}(y|\mathbf{x})$, son positivas. El costo de asegurar predicciones positivas para y es que la ecuación (17.25) es más complicada que un modelo lineal para $\mathrm{E}(y|\mathbf{x})$. Pero aun más importante es que los efectos parciales de (17.25) son más complicados que para un modelo lineal. Como se verá, los efectos parciales de x_j sobre $\mathrm{E}(y|y > 0,\mathbf{x})$ y $\mathrm{E}(y|\mathbf{x})$ tienen el mismo signo que el coeficiente, β_j, pero la magnitud de los efectos depende de los valores de *todas* las variables explicativas y parámetros. Debido a que σ aparece en (17.25), no es de sorprender que los efectos parciales dependan también de σ.

Si x_j es una variable continua, se pueden hallar los efectos parciales mediante el cálculo. Primero,

$$\partial\mathrm{E}(y|y > 0,\mathbf{x})/\partial x_j = \beta_j + \beta_j \cdot \frac{d\lambda}{dc}(\mathbf{x}\boldsymbol{\beta}/\sigma),$$

si se supone que x_j no tiene una relación funcional con otros regresores. Al diferenciar $\lambda(c) = \phi(c)/\Phi(c)$ y usar $d\Phi/dc = \phi(c)$ así como $d\phi/dc = -c\phi(c)$, se puede mostrar que $d\lambda/dc = -\lambda(c)[c + \lambda(c)]$. Por tanto,

$$\partial\mathrm{E}(y|y > 0,\mathbf{x})/\partial x_j = \beta_j\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}. \tag{17.26}$$

Esto muestra que el efecto parcial de x_j sobre $\mathrm{E}(y|y > 0,\mathbf{x})$ no está determinado sólo por β_j. El factor de ajuste está dado por el término entre los corchetes, $\{\cdot\}$, y depende de una función lineal de $\mathbf{x}$, $\mathbf{x}\boldsymbol{\beta}/\sigma = (\beta_0 + \beta_1x_1 + \ldots + \beta_kx_k)/\sigma$. Se puede mostrar que el factor de ajuste está estrictamente entre cero y uno. En la práctica se puede estimar (17.26) mediante la inserción en las EMV de β_j y σ. Como con los modelos logit y probit, se pueden insertar valores para x_j, por lo general los valores promedio u otros valores interesantes. La ecuación (17.26) revela un punto sutil que en ocasiones se pierde al aplicar el modelo Tobit a las respuestas de solución de esquina: el parámetro σ aparece directamente en los efectos parciales, así que tener una estimación de σ es crucial para estimar los efectos parciales. Algunas veces, σ recibe el nombre de parámetro "auxiliar" (lo cual significa secundario o sin importancia). Aunque es verdad que el valor de σ no afecta el signo de los efectos parciales, sí afecta las magnitudes, y la importancia económica de las variables explicativas suele ser interesante. Por tanto, caracterizar σ como auxiliar es engañoso, pues proviene de una confusión entre el modelo Tobit para aplicaciones de solución de esquina y aplicaciones a la censura de datos verdaderos. (Vea la sección 17.4.)

Todas las cantidades económicas usuales, como las elasticidades, se pueden calcular. Por ejemplo, la elasticidad de y respecto a x_1, condicional sobre $y > 0$, es

$$\frac{\partial\mathrm{E}(y|y > 0,\mathbf{x})}{\partial x_1} \cdot \frac{x_1}{\mathrm{E}(y|y > 0,\mathbf{x})}. \tag{17.27}$$

Esto se puede calcular cuando x_1 aparece en varias formas funcionales, incluidas las formas de nivel, logarítmicas y cuadráticas.

Si x_1 es una variable binaria, el efecto de interés se obtiene como la diferencia entre $\mathrm{E}(y|y > 0, \mathbf{x})$, con $x_1 = 1$ y $x_1 = 0$. Los efectos parciales que implican otras variables discretas (como el número de hijos) se pueden manejar de modo similar.

Se puede usar (17.25) para hallar la derivada parcial de $\mathrm{E}(y|\mathbf{x})$ respecto a x_j continuas. Esta derivada puede explicar el hecho de que la gente que comienza en $y = 0$ podría elegir $y > 0$ cuando x_j cambia:

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \frac{\partial \mathrm{P}(y > 0|\mathbf{x})}{\partial x_j} \cdot \mathrm{E}(y|y > 0, \mathbf{x}) + \mathrm{P}(y > 0|\mathbf{x}) \cdot \frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial x_j}. \tag{17.28}$$

Debido a que $\mathrm{P}(y > 0|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$,

$$\frac{\partial \mathrm{P}(y > 0|\mathbf{x})}{\partial x_j} = (\beta_j/\sigma)\phi(\mathbf{x}\beta/\sigma), \tag{17.29}$$

así que se puede estimar cada término en (17.28), una vez que se insertan las EMV de las β_j y σ y los valores particulares de las x_j.

Es de notar que cuando se inserta (17.26) y (17.29) en (17.28) y se usa el hecho de que $\Phi(c)\lambda(c) = \phi(c)$ para cualquier c, se obtiene

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \beta_j \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \tag{17.30}$$

La ecuación (17.30) nos permite hacer comparaciones aproximadas de las estimaciones de MCO y de Tobit. [La ecuación (17.30) también se puede derivar directamente de la ecuación (17.25) mediante el hecho de que $d\phi(z)/dz = -z\phi(z)$.] Los coeficientes de pendiente de MCO, por ejemplo, $\hat{\gamma}_j$, de la regresión de y_i sobre $x_{i1}, x_{i2}, ..., x_{ik}$, $i = 1, ..., n$ (es decir, usando todos los datos), son estimaciones directas de $\partial \mathrm{E}(y|\mathbf{x})/\partial x_j$. Para hacer que el coeficiente Tobit, $\hat{\beta}_j$, sea comparable con $\hat{\gamma}_j$, se debe multiplicar $\hat{\beta}_j$ por un factor de ajuste.

Como en los casos logit y probit, existen dos métodos comunes para calcular un factor de ajuste para obtener efectos parciales, al menos para variables explicativas continuas. Ambos se basan en la ecuación (17.30). Primero, el efecto parcial al promedio, EPA, se obtiene al evaluar $\Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma})$, que se denota $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Se puede usar este factor único para multiplicar los coeficientes de las variables explicativas continuas. El EPA tiene aquí las mismas desventajas que en los casos probit y logit: quizá no interese el efecto parcial para el "promedio" debido a que el promedio no es interesante o carece de significado. Además, se debe decidir si usar promedios de funciones no lineales o insertar los promedios en las funciones no lineales.

El efecto parcial promedio, EPP, se prefiere en la mayoría de los casos. Aquí se calcula el factor escalar como $n^{-1} \sum_{i=1}^{n} \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$. A diferencia de EPA, el EPP no requiere insertar una unidad ficticia o inexistente de la población, y no hay decisiones que tomar acerca de insertar promedios en funciones no lineales. Al igual que EPA, el factor escalar EPP siempre está entre cero y uno debido a que $0 < \Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma}) < 1$ para cualquier valor de las variables explicativas. De hecho, $\hat{\mathrm{P}}(y_i > 0|\mathbf{x}_i) = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$, y entonces el factor escalar EPP y el factor escalar EPA tienden a ser cercanos a uno cuando hay pocas observaciones con $y_i = 0$. En caso de que $y_i > 0$ para toda i, las estimaciones Tobit y de MCO de los parámetros son idénticas. [Por supuesto, si $y_i > 0$ para toda i, no se pude justificar el uso de un modelo Tobit. Usar log (y_i) en un modelo de regresión lineal tiene mucho más sentido.]

Por desgracia, para las variables explicativas discretas, comparar las estimaciones de MCO y de Tobit no es tan fácil (aunque usar el factor escalar para las variables explicativas continuas suele ser una aproximación útil). Para Tobit, el efecto parcial de una variable explicativa discreta, por ejemplo, una variable binaria, se debe obtener realmente al estimar $\mathrm{E}(y|\mathbf{x})$ de la ecuación (17.25). Por ejemplo, si x_1 es binaria, se debe insertar primero $x_1 = 1$ y después $x_1 = 0$. Si se establecen las demás variables explicativas en sus promedios muestrales, se obtiene una medida análoga a (17.16) para los casos logit y probit. Si se calcula la diferencia en los valores esperados para cada individuo y después se promedia la diferencia, se obtiene un EPP análogo a (17.17).

Ejemplo 17.2

[Fuerza de trabajo anual de las mujeres casadas]

El archivo MROZ.RAW incluye datos sobre las horas trabajadas de 753 mujeres casadas, 428 de las cuales trabajaron por un salario fuera del hogar durante el año; 325 de las mujeres trabajaron cero horas. Para las mujeres que trabajaron horas positivas, el rango es muy amplio, que va de 12 a 4,950. Por tanto, las horas anuales trabajadas es un buen candidato para un modelo Tobit. También se estima un modelo lineal (usando todas las 753 observaciones) mediante MCO. Los resultados se dan en la tabla 17.2.

Esta tabla tiene varias características notorias. Primero, las estimaciones Tobit de los coeficientes tienen el mismo signo que las estimaciones de MCO correspondientes y su significancia estadística es similar. (Las posibles excepciones son los coeficientes de *nwifeinc* y *kidsge6*, pero los estadísticos t tienen magnitudes similares.) Segundo, aunque es tentador comparar las magnitudes de las estimaciones de MCO y de Tobit, esto no es muy informativo. Es necesario tener cuidado en no pensar que, debido a que el coeficiente Tobit en *kidslt6* es aproximadamente el doble que el coeficiente de MCO, el modelo Tobit implica una respuesta mucho mayor de horas trabajadas respecto a los niños pequeños.

Se pueden multiplicar las estimaciones Tobit mediante los factores adecuados de ajuste para hacerlas aproximadamente comparables a las estimaciones de MCO. El factor escalar EPP $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ resulta ser de cerca de .589, el cual se puede usar para obtener los efectos parciales promedio para la estimación Tobit. Si, por ejemplo, se multiplica el coeficiente de *educ* por .589 se obtiene .589(80.65) $\approx$ 47.50 (es decir, 47.5 horas más), lo cual es un poco mayor que el efecto parcial de MCO, de aproximadamente 28.8 horas. Así que, incluso para estimar un efecto promedio, las estimaciones Tobit son notablemente mayores en magnitud que la estimación de MCO correspondiente. Si, en lugar de ello, se desea el efecto estimado de otro año de educación a partir de los valores promedio de todas las variables explicativas, entonces se calcula el factor escalar EPA $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Éste resulta ser de aproximadamente .645 [cuando se usa el promedio cuadrado de la experiencia, $(\overline{exper})^2$, en lugar del promedio de $exper^2$]. Este efecto parcial, que es de aproximadamente 52 horas, es casi el doble de la estimación de MCO. Con excepción de *kidsge6*, los coeficientes de pendiente escalados de Tobit son todos mayores en magnitud que el coeficiente correspondiente de MCO.

Se ha reportado una R-cuadrada tanto para el modelo Tobit como para el de regresión lineal. La R-cuadrada para MCO es la acostumbrada. Para Tobit, la R-cuadrada es el coeficiente de correlación entre y_i y $\hat{y}_i$, donde $\hat{y}_i = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})\mathbf{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}\phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ es la estimación de $\mathrm{E}(y|\mathbf{x} = \mathbf{x}_i)$. Esto está motivado por el hecho de que la R-cuadrada acostumbrada para MCO es igual a la correlación cuadrada entre y_i y los valores ajustados [vea la ecuación (3.29)]. En los modelos no lineales como el modelo Tobit, el coeficiente de correlación cuadrada no es idéntico a una R-cuadrada basado en una suma de residuales cuadrados como en (3.28). Esto es debido a que los valores ajustados, como se definieron antes, y los residuales, $y_i - \hat{y}_i$, están correlacionados en la muestra. Una R-cuadrada definida como el coeficiente de correlación cuadrado entre y_i y $\hat{y}_i$ tiene la ventaja de siempre estar entre cero y uno; una R-cuadrada basada en la suma de los residuales cuadrados no necesita tener esta característica.

TABLA 17.2

Estimación Tobit y de MCO de las horas anuales trabajadas

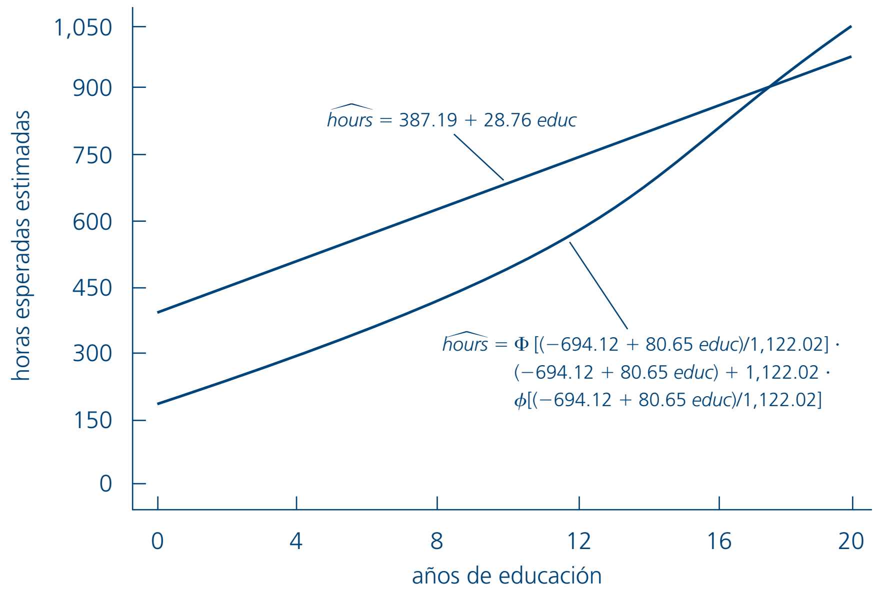
Variable dependiente: *hours*		
Variables independientes	**Lineal (MCO)**	**Tobit (EMV)**
nwifeinc	−3.45 (2.54)	−8.81 (4.46)
educ	28.76 (12.95)	80.65 (21.58)
exper	65.67 (9.96)	131.56 (17.28)
$exper^2$	−.700 (.325)	−1.86 (0.54)
age	−30.51 (4.36)	−54.41 (7.42)
kidslt6	−442.09 (58.85)	−894.02 (111.88)
kidsge6	−32.78 (23.18)	−16.22 (38.64)
constante	1,330.48 (270.78)	965.31 (446.44)
Valor de log-verosimilitud R-cuadrada $\hat{\sigma}$	— .266 750.18	−3,819.09 .274 1,122.02

Se puede ver que, con base en las medidas de R-cuadrada, la función de media condicional Tobit encaja un poco mejor con los datos de las horas, pero no es una mejora sustancial. No obstante, se debe recordar que las estimaciones Tobit no se eligen para maximizar una R-cuadrada (maximizan la función de log-verosimilitud), mientras que las estimaciones de MCO son los valores que producen la R-cuadrada más alta dada la forma funcional lineal.

Por construcción, todos los valores ajustados Tobit para *hours* son positivos. En contrate, 39 de los valores ajustados por MCO son negativos. Aunque las predicciones negativas son de algún interés, 39 de 753 sólo está por encima de 5% de las observaciones. No es completamente claro cómo se traducen los valores ajustados negativos de MCO a las diferencias en los efectos parciales estimados. La figura 17.3 presenta la gráfica de las estimaciones de $E(hours|\mathbf{x})$ en función de la educación; para el modelo Tobit, las demás variables explicativas se establecen en sus valores promedio. Para el modelo lineal, la ecuación

FIGURA 17.3

Valores esperados estimados de las horas respecto a la educación para los modelos lineal y Tobit.



graficada es $\widehat{hours} = 387.19 + 28.76\ educ$. Para el modelo Tobit, la ecuación graficada es $\widehat{hours} = \Phi[(-694.12 + 80.65\ educ)/1{,}122.02] \cdot (-694.12 + 80.65\ educ) + 1{,}122.02 \cdot \phi[(-694.12 + 80.65\ educ)/1{,}122.02]$. Como se puede ver en la figura, el modelo lineal da estimaciones notablemente superiores de las horas trabajadas esperadas incluso para niveles muy altos de educación. Por ejemplo, a los ocho años de educación, el valor predicho por MCO de las horas es de cerca de 617.5, mientras que la estimación Tobit es de aproximadamente 423.9. A los 12 años de educación, las horas predichas son de cerca de 732.7 y 598.3, respectivamente. Las dos líneas de predicción cruzan después de los 17 años de educación, pero ninguna mujer en la muestra tiene más de 17 años de educación. La pendiente creciente de la línea Tobit indica con claridad el creciente efecto marginal de la educación sobre las horas trabajadas esperadas.

Problemas de especificación en los modelos Tobit

El modelo Tobit y en particular las fórmulas para las expectativas en (17.24) y (17.25), dependen de forma crucial de la normalidad y la homocedasticidad en el modelo de la variable latente subyacente. Cuando $E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, se sabe del capítulo 5 que la normalidad condicional de y no desempeña una función en la insesgadez, consistencia o inferencia en muestras grandes. La heterocedasticidad no afecta el insesgamiento o consistencia de MCO, aunque se deben calcular errores estándar y estadísticos de prueba robustos para realizar la inferencia aproximada. En un modelo Tobit, si cualquiera de los supuestos en (17.18) falla, entonces es difícil saber qué está estimando la EMV Tobit. No obstante, para cambios moderados respecto a los supuestos, el modelo Tobit es probable que ofrezca buenas estimaciones de los efectos parciales sobre las medias condicionales. Es posible permitir supuestos más generales en (17.18), pero tales modelos son más difíciles de estimar e interpretar.

Una limitación potencialmente importante del modelo Tobit, al menos en ciertas aplicaciones, es que el valor esperado condicional en $y > 0$ está vinculado estrechamente con la probabilidad de que $y > 0$. Esto resulta claro en las ecuaciones (17.26) y (17.29). En particular, el efecto de x_j sobre $P(y > 0|\mathbf{x})$ es proporcional a β_j, así como el efecto sobre $E(y|y > 0,\mathbf{x})$, donde las dos funciones que multiplican β_j son positivas y dependen de $\mathbf{x}$ sólo a través de $\mathbf{x}\boldsymbol{\beta}/\sigma$. Esto descarta algunas posibilidades interesantes. Por ejemplo, considere la relación entre la cantidad de cobertura de un seguro de vida y la edad de una persona. Es menos probable que los jóvenes tengan un seguro de vida, así que la probabilidad de que $y > 0$ aumenta con la edad (al menos hasta cierto punto). Dado que se tiene un seguro de vida, el valor de las pólizas puede disminuir con la edad, dado que el seguro de vida se vuelve menos importante a medida que las personas se acercan al final de sus vidas. Esta posibilidad queda descartada en el modelo Tobit.

Una forma de evaluar de manera informal si el modelo Tobit es adecuado es estimar un modelo probit donde el resultado binario, por ejemplo, w, sea igual a uno si $y > 0$, y $w = 0$ si $y = 0$. Entonces, de (17.21), w sigue un modelo probit, donde el coeficiente de x_j es $\gamma_j = \beta_j/\sigma$. Esto significa que se puede estimar la razón de β_j sobre σ mediante probit, para cada j. Si el modelo Tobit es válido, la estimación probit, $\hat{\gamma}_j$, debe ser "cercana" a $\hat{\beta}_j/\hat{\sigma}$, donde $\hat{\beta}_j$ y $\hat{\sigma}$ son las estimaciones Tobit. Éstas nunca serán idénticas debido al error de muestreo. Pero se pueden buscar ciertos signos problemáticos. Por ejemplo, si $\hat{\gamma}_j$ es significativa y negativa, pero $\hat{\beta}_j$ es positiva, quizás el modelo Tobit no sea apropiado. O, si $\hat{\gamma}_j$ y $\hat{\beta}_j$ tienen el mismo signo, pero $|\hat{\beta}_j/\hat{\sigma}|$ es mucho mayor o menor que $|\hat{\gamma}_j|$, esto también podría indicar problemas. No es necesario preocuparse demasiado por los cambios de signo o las diferencias en la magnitud de las variables explicativas que sean insignificantes en ambos modelos.

En el ejemplo de las horas anuales trabajadas, $\hat{\sigma} = 1{,}122.02$. Cuando se dividió el coeficiente Tobit de *nwifeinc* entre $\hat{\sigma}$, se obtuvo $-8.81/1{,}122.02 \approx -.0079$; el coeficiente probit de *nwifeinc* es de cerca de $-.012$, lo cual es diferente, pero no de forma drástica. En *kidslt6*, el coeficiente estimado entre $\hat{\sigma}$ es de cerca de $-.797$, en comparación con la estimación probit de $-.868$. De nuevo, ésta no es una diferencia enorme, pero indica que tener niños pequeños tiene un efecto mayor sobre la decisión inicial de participar en la fuerza laboral que sobre cuántas horas elige trabajar una mujer una vez que está en dicha fuerza. (Tobit promedia de forma efectiva estos dos efectos.) No se sabe si los efectos son estadísticamente diferentes, pero son del mismo orden de magnitud.

¿Qué sucede si se concluye que el modelo Tobit es inadecuado? Existen modelos, que suelen conocerse como modelos de *dos partes o* de *obstáculos*, que se pueden usar cuando Tobit es inadecuado. Todos tienen la propiedad de que $P(y > 0|\mathbf{x})$ y $E(y|y > 0,\mathbf{x})$ depende de diferentes parámetros, así que x_j puede tener efectos disímiles sobre estas dos funciones. [Vea Wooldridge (2002, capítulo 16) para una descripción de estos modelos.]

17.3 El modelo de regresión de Poisson

Otra clase de variable dependiente no negativa es una **variable de conteo**, que puede asumir valores enteros no negativos: $\{0, 1, 2, \ldots\}$. Lo que más interesa aquí son los casos en los que y asume relativamente pocos valores, incluido el cero. Los ejemplos incluyen el número de niños que ha tenido una mujer, el número de veces que alguien es arrestado en el año o el número de patentes que una empresa registra al año. Por las mismas razones analizadas para las respuestas binarias y Tobit, un modelo lineal para $E(y|x_1, \ldots, x_k)$ quizá no proporcione el mejor ajuste a lo largo de todos los valores de las variables explicativas. (No obstante, siempre es informativo comenzar con un modelo lineal, como se hizo en el ejemplo 3.5.)

Como con un resultado Tobit no se puede obtener el logaritmo de una variable de conteo debido a que asume el valor cero. Un método útil es modelar el valor esperado como una función exponencial:

$$E(y|x_1, x_2, \ldots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k). \tag{17.31}$$

Debido a que $\exp(\cdot)$ siempre es positiva, (17.31) asegura que los valores predichos para y también serán positivos. La función exponencial se grafica en la figura A.5 del apéndice A.

Aunque (17.31) es más complicada que un modelo lineal, básicamente ya se sabe cómo interpretar los coeficientes. Al obtener el logaritmo de la ecuación (17.31) se muestra que

$$\log[E(y|x_1, x_2, \ldots, x_k)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{17.32}$$

así que el logaritmo del valor esperado es lineal. Por tanto, mediante las propiedades de aproximación de la función logaritmo que se han usado con frecuencia en los capítulos anteriores,

$$\%\Delta E(y|\mathbf{x}) \approx (100\beta_j)\Delta x_j.$$

En otras palabras, $100\beta_j$ es el cambio porcentual aproximado en $E(y|\mathbf{x})$, dado un incremento de una unidad en x_j. Algunas veces, es necesaria una estimación más precisa y es fácil encontrar una al observar los cambios discretos en el valor esperado. Mantenga todas las variables explicativas salvo x_k fijas y sea x_k^0 el valor inicial y x_k^1 el valor subsiguiente. Entonces, el cambio proporcional en el valor esperado es

$$[\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^1)/\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^0)] - 1 = \exp(\beta_k \Delta x_k) - 1,$$

donde $\mathbf{x}_{k-1}\beta_{k-1}$ es una abreviación de $\beta_1 x_1 + \ldots + \beta_{k-1}x_{k-1}$, y $\Delta x_k = x_k^1 - x_k^0$. Cuando $\Delta x_k = 1$ —por ejemplo, si x_k es una variable binaria que se cambia de cero a uno—, entonces el cambio es $\exp(\beta_k) - 1$. Dada $\hat{\beta}_k$, se obtiene $\exp(\hat{\beta}_k) - 1$ y se multiplica esto por 100 para transformar el cambio proporcional a un cambio porcentual.

Si, por ejemplo $x_j = \log(z_j)$ para alguna variable $z_j > 0$, entonces su coeficiente, β_j, se interpreta como una elasticidad respecto a z_j. Técnicamente es una elasticidad del *valor esperado* de y respecto a z_j debido a que no se puede calcular el cambio porcentual en casos donde $y = 0$. Para el propósito de este libro, la distinción es irrelevante. El resultado final es que, para fines prácticos, se pueden interpretar los coeficientes en la ecuación (17.31) como si se tuviera un modelo lineal, con $\log(y)$ como la variable dependiente. Existen algunas diferencias sutiles que no se deben estudiar aquí.

Debido a que (17.31) es no lineal en sus parámetros [recuerde que $\exp(\cdot)$ es una función no lineal], no se pueden usar los métodos de regresión lineal. Se podrían usar los *mínimos cuadrados no lineales*, que, igual que MCO, minimizan la suma de los residuales cuadrados. No obstante, resulta que todas las distribuciones estándar de datos de conteo exhiben heterocedasticidad y los mínimos cuadrados no lineales no explotan esto [vea Wooldridge (2002, capítulo 12)]. En lugar de esto, aquí se dependerá más de la máxima verosimilitud y del importante método relacionado de la *estimación de cuasi máxima* verosimilitud.

En el capítulo 4 se presentó la normalidad como el supuesto distribucional estándar para la regresión lineal. El supuesto de normalidad es razonable para variables dependientes (aproximadamente) continuas que asuman un amplio rango de valores. Una variable de conteo no puede tener una distribución normal (debido a que la distribución normal es para las variables continuas que puedan asumir todos los valores) y si asume sólo unos pocos valores, la distribución puede ser muy diferente de la normal. En lugar de ello, la distribución nominal para los datos de conteo es la **distribución de Poisson**.

Como lo que aquí interesa es el efecto de las variables explicativas sobre y, se debe observar la distribución de Poisson condicional sobre $\mathbf{x}$. La distribución de Poisson está determinada por completo por su media, así que sólo se necesita especificar $E(y|\mathbf{x})$. Se supone que ésta tiene la misma forma de (17.31), que se abrevia como $\exp(\mathbf{x}\boldsymbol{\beta})$. Entonces, la probabilidad de que y sea igual al valor h, condicional sobre $\mathbf{x}$, es

$$P(y = h|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})][\exp(\mathbf{x}\boldsymbol{\beta})]^h/h!,\ h = 0, 1, ...,$$

donde $h!$ denota un factorial (vea apéndice B). Esta distribución, que es la base del **modelo de regresión de Poisson**, permite hallar las probabilidades condicionales para cualquier valor de las variables explicativas. Por ejemplo, $P(y = 0|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$. Una vez que se tienen las estimaciones de las β_j, se pueden insertar en las probabilidades para diferentes valores de $\mathbf{x}$.

Dada una muestra aleatoria $\{(\mathbf{x}_i, y_i): i = 1, 2, ..., n\}$, se puede construir la función de log-verosimilitud:

$$\mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i\mathbf{x}_i\boldsymbol{\beta} - \exp(\mathbf{x}_i\boldsymbol{\beta})\}, \quad \boxed{17.33}$$

donde se desecha el término $-\log(y_i!)$ debido a que no depende de $\boldsymbol{\beta}$. Esta función de log-verosimilitud es fácil de maximizar, aunque las EMV de Poisson no se obtienen de forma cerrada.

Los errores estándar de las estimaciones de Poisson $\hat{\beta}_j$ son fáciles de obtener después de que se ha maximizado la función de log-verosimilitud; la fórmula está en el apéndice 17B. Éstas se reportan junto con las $\hat{\beta}_j$ mediante cualquier software.

Como con los modelos probit, logit y Tobit, no se pueden comparar directamente las magnitudes de las estimaciones de Poisson de una función exponencial con las estimaciones de MCO de una función lineal. Sin embargo, es posible una comparación aproximada, al menos para las variables explicativas continuas. Si (17.31) aplica, entonces el efecto parcial de x_j respecto a $E(y|x_1, x_2, ..., x_k)$ es $\partial E(y|x_1, x_2, ..., x_k)/\partial x_j = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) \cdot \beta_j$. Esta expresión se desprende de la regla de la cadena en cálculo debido a que la derivada de la función exponencial es sólo la función exponencial. Si $\hat{\gamma}_j$ denota un coeficiente de pendiente MCO de la regresión de y sobre $x_1, x_2, ..., x_k$, entonces se puede comparar aproximadamente la magnitud de las $\hat{\gamma}_j$ y el efecto parcial promedio de una función de regresión exponencial. Es interesante ver que el factor escalar EPP en este caso, $n^{-1}\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}) = n^{-1}\sum_{i=1}^{n}\hat{y}_i$, es simplemente el promedio muestral $\bar{y}$ de y_i, donde se definen los valores ajustados como $\hat{y}_i = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$. En otras palabras, para la regresión de Poisson con una función media exponencial, el promedio de los valores ajustados es el mismo que el promedio de los resultados originales de y_i, tal como en el caso de regresión lineal. Esto hace que sea sencillo escalar las estimaciones Poisson, $\hat{\beta}_j$, para hacerlas comparables a las estimaciones de MCO correspondientes, $\hat{\gamma}_j$: para una variable explicativa continua, se puede comparar $\hat{\gamma}_j$ con $\bar{y} \cdot \hat{\beta}_j$.

Aunque el análisis de EMV de Poisson es un primer paso natural para los datos de conteo, suele ser muy restrictivo. Todas las probabilidades y momentos mayores de la distribución de Poisson se determinan por completo por la media. En particular, la varianza es igual a la media:

$$\mathrm{Var}(y|\mathbf{x}) = E(y|\mathbf{x}). \quad \boxed{17.34}$$

Esto es restrictivo y se ha mostrado que se viola en muchas aplicaciones. Por fortuna, la distribución de Poisson tiene una propiedad de robustez muy buena: manténgase o no la distribución de Poisson, aún se obtendrán estimadores asintóticamente normales y consistentes de las β_j. [Vea

Wooldridge (2002, capítulo 19) para ver más detalles.] Esto es análogo al estimador de MCO, que es consistente y asintóticamente normal manténgase o no el supuesto de normalidad; aun así, MCO es la EMV bajo normalidad.

Cuando se usa la EMV de Poisson, pero no se supone que la distribución de Poisson sea completamente correcta, este análisis recibe el nombre de **estimación de cuasi máxima verosimilitud (ECMV)**. La ECMV de Poisson es muy útil debido a que está programada en muchos paquetes de econometría. No obstante, a menos que el supuesto de varianza de Poisson (17.34) se mantenga, se deben ajustar los errores estándar.

Un simple ajuste a los errores estándar está disponible cuando se supone que la varianza es proporcional a la media:

$$\mathrm{Var}(y|\mathbf{x}) = \sigma^2 \mathrm{E}(y|\mathbf{x}), \tag{17.35}$$

donde $\sigma^2 > 0$ es un parámetro desconocido. Cuando $\sigma^2 = 1$, se obtiene el supuesto de la varianza de Poisson. Cuando $\sigma^2 > 1$, la varianza es mayor que la media para toda $\mathbf{x}$; a esto se le llama **sobredispersión**, debido a que la varianza es mayor que en el caso de Poisson y se observa en muchas aplicaciones de regresiones de conteo. El caso de $\sigma^2 < 1$, llamado *subdispersión*, es menos común pero se permite en (17.35).

Bajo (17.35) es fácil ajustar los errores estándar de la EMV de Poisson. Si $\hat{\beta}_j$ denota la ECMV de Poisson y se definen los residuales como $\hat{u}_i = y_i - \hat{y}_i$, donde $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$ es el valor ajustado. Como es usual, el residual para la observación i es la diferencia entre y_i y su valor ajustado. Un estimador consistente de σ^2 es $(n - k - 1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2/\hat{y}_i$, donde la división entre $\hat{y}_i$ es el ajuste apropiado de heterocedasticidad y $n - k - 1$ son los *gl* dadas n observaciones y $k + 1$ estimadores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Dejando que $\hat{\sigma}$ sea la raíz cuadrada positiva de $\hat{\sigma}^2$, se multiplican los errores estándar de Poisson usuales por $\hat{\sigma}$. Si $\hat{\sigma}$ es notablemente mayor que uno, los errores estándar corregidos pueden ser mucho mayores que los errores estándar nominales, generalmente incorrectos, de la EMV de Poisson.

Incluso (17.35) no es completamente general. Tal como en el modelo lineal, se pueden obtener los errores estándar para la ECMV de Poisson que no restrinjan en absoluto la varianza. [Vea Wooldridge (2002, capítulo 19) para una mayor explicación.]

Bajo el supuesto distribucional de Poisson, se puede usar el estadístico de la razón de verosimilitudes para probar las restricciones de exclusión que, como siempre, tienen la forma de (17.12). Si se tienen q restricciones de exclusión, el estadístico se distribuye aproximadamente como χ_q^2 bajo la hipótesis nula. Bajo el supuesto menos restrictivo (17.35), un simple ajuste está disponible (y entonces se denomina al estadístico **estadístico de razón de cuasi verosimilitudes**): se divide (17.12) entre $\hat{\sigma}^2$, donde $\hat{\sigma}^2$ se obtiene del modelo no restringido.

Pregunta 17.4

Suponga que se obtiene $\hat{\sigma}^2 = 2$. ¿Cómo se comparan los errores estándar ajustados con los errores estándar usuales de la EMV de Poisson? ¿Cómo se compara el estadístico razón de cuasi de verosimilitudes con el estadístico usual de razón de verosimilitudes?

Ejemplo 17.3

[Regresión de Poisson para el número de arrestos]

Ahora se aplicará el modelo de regresión de Poisson a los datos de arrestos en CRIME1.RAW, que se usó, entre otros casos, en el ejemplo 9.1. La variable dependiente, *narr86*, es el número de veces que un hombre fue arrestado en 1986. Esta variable es cero para 1,970 de 2,725 hombres de la muestra y sólo ocho valores de *narr86* son mayores que cinco. Por tanto, el modelo de regresión de Poisson es más apropiado que el modelo de regresión lineal. La tabla 17.3 también presenta los resultados de la estimación de MCO de un modelo de regresión lineal.

TABLA 17.3

Determinantes del número de arrestos de hombres jóvenes

Variable dependiente: *narr86*		
Variables independientes	**Lineal (MCO)**	**Exponencial (ECMV de Poisson)**
pcnv	−.132 (.040)	−.402 (.085)
avgsen	−.011 (.012)	−.024 (.020)
tottime	.012 (.009)	.024 (.015)
ptime86	−.041 (.009)	−.099 (.021)
qemp86	−.051 (.014)	−.038 (.029)
inc86	−.0015 (.0003)	−.0081 (.0010)
black	.327 (.045)	.661 (.074)
hispan	.194 (.040)	.500 (.074)
born60	−.022 (.033)	−.051 (.064)
constante	.577 (.038)	−.600 (.067)
Valor de log-verosimilitud R-cuadrada $\hat{\sigma}$	— .073 .829	−2,248.76 .077 1.232

Los errores estándar de MCO son los acostumbrados; sin lugar a dudas se podrían hacer robustos para la heterocedasticidad. Los errores estándar para la regresión de Poisson son los errores estándar usuales de máxima verosimilitud. Debido a que $\hat{\sigma} = 1.232$, los errores estándar de la regresión de Poisson se deben inflar por este factor (de manera que cada error estándar corregido es aproximadamente 23% más grande). Por ejemplo, un error estándar más confiable para *tottime* es $1.23(.015) \approx .0185$, lo cual da un estadístico t de aproximadamente 1.3. El ajuste a los errores estándar reduce la significancia de todas las variables, pero varias de ellas aún son estadísticamente significativas.

Los coeficientes de MCO y de Poisson no son comparables directamente y tienen significados muy diferentes. Por ejemplo, el coeficiente de *pcnv* implica que, si $\Delta pcnv = .10$, el número esperado de arrestos desciende .013 (*pcnv* es la proporción de arrestos previos que desembocaron en una condena). El coeficiente de Poisson implica que $\Delta pcnv = .10$ reduce los arrestos esperados cerca de 4% [$.402(.10) = .0402$, y se multiplica esto por 100 para obtener el efecto porcentual]. Como cuestión de políticas, esto sugiere que se pueden reducir los arrestos generales cerca de 4% si se puede incrementar la probabilidad de condena .1.

El coeficiente de Poisson de *black* implica que, cuando los demás factores se mantienen iguales, el número esperado de arrestos para los hombres negros se estima como de cerca de $100 \cdot [\exp(.661) - 1] \approx 93.7\%$ mayor que para los hombres blancos con los mismos valores para las demás variables explicativas.

Como con la aplicación Tobit en la tabla 17.2, se reporta una R-cuadrada para la regresión de Poisson: el coeficiente de correlación cuadrado entre y_i y $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$. La motivación de esta medición de la bondad de ajuste es la misma que para el modelo Tobit. Se puede ver que el modelo de regresión exponencial, estimado por la ECMV de Poisson, se ajusta un poco mejor. Recuerde que las estimaciones de MCO se eligen para maximizar la R-cuadrada, pero las estimaciones de Poisson no. (Se eligen para maximizar la función de log-verosimilitud.)

Se han propuesto otros modelos de regresión de conteo de datos y se usan en aplicaciones, los cuales generalizan la distribución de Poisson de varias formas. Si lo que interesa son los efectos de las x_j sobre la media de la respuesta, hay poca razón para ir más allá de la regresión de Poisson: es simple, suele dar buenos resultados y tienen una propiedad de robustez de la que se habló antes. De hecho, se puede aplicar la regresión de Poisson a una y que sea un resultado parecido a Tobit, siempre y cuando (17.31) sea válida. Esto podría dar buenas estimaciones de los efectos sobre la media. Las extensiones de la regresión de Poisson son más útiles cuando lo que interesa es estimar probabilidades, como $P(y > 1|\mathbf{x})$. [Vea por ejemplo, Cameron y Trivedi (1998).]

17.4 Modelos de regresión censurada y truncada

Los modelos en las secciones 17.1, 17.2 y 17.3 aplican a varias clases de variables dependientes limitadas que con frecuencia surgen en el trabajo econométrico. Al usar estos métodos es importante recordar que se usa un modelo probit o logit para una respuesta binaria, un modelo Tobit para un resultado de solución de esquina o un modelo de regresión de Poisson para una respuesta de conteo debido a que deseamos modelos que den cuenta de las características importantes de la distribución de y. No hay problemas de observabilidad de los datos. Por ejemplo, en la aplicación Tobit al trabajo femenino en el ejemplo 17.2, no hay problema con observar las horas trabajadas: es simplemente el caso de que una fracción no trivial de mujeres casadas en la población elige no trabajar a cambio de un salario. En la aplicación de regresión de Poisson a los arrestos anuales, se observa la variable dependiente para todo hombre joven en una muestra aleatoria de la población, pero la variable dependiente puede tomar el valor de cero, así como otros valores enteros pequeños.

Por desgracia, la distinción entre la aglutinación de una variable de resultado (como tomar el valor de cero para una fracción no trivial de la población) y los problemas de censura de datos puede ser confusa. Esto es particularmente cierto cuando se aplica el modelo Tobit. En este libro, el modelo Tobit estándar descrito en la sección 17.2 es sólo para resultados de solución de esquina. Pero la literatura sobre los modelos Tobit suele tratar otra situación dentro del mismo marco: la variable de respuesta ha sido censurada por encima o debajo de algún límite. Por lo general, la censura se debe al diseño de la encuesta y, en algunos casos, a las restricciones institucionales. En lugar de tratar los problemas de censura de datos junto con los resultados de solución de esquina, se resuelve la censura de datos al aplicar el **modelo de regresión censurada**. En esencia, el problema resuelto por un modelo de regresión censurada es el del dato faltante en la variable de respuesta, y. Aunque es posible extraer aleatoriamente las unidades de la población y obtener información sobre las variables explicativas de todas las unidades, el resultado sobre y_i falta para cierta i. Aún así, se sabe que los valores faltantes están sobre o debajo de un límite determinado y este conocimiento nos da información útil para estimar los parámetros.

Un **modelo de regresión truncada** surge cuando se excluye, con base en y, un subconjunto de la población en el esquema de muestreo. En otras palabras, no se tiene una muestra aleatoria de la población subyacente, pero se conoce la regla que se usó para incluir unidades en la muestra. Esta regla se determina por si y está por encima o debajo de cierto límite. Se explicará con mayor detalle la diferencia entre los modelos de regresión truncada y censurada más tarde.

Modelos de regresión censurada

Si bien los modelos de regresión censurada se pueden definir sin supuestos distribucionales, en esta subsección se estudia el **modelo de regresión censurada normal**. La variable que se quisiera explicar, y, sigue el modelo lineal clásico. Para enfatizar, se coloca un subíndice i en una extracción aleatoria de la población:

$$y_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i,\ u_i|\mathbf{x}_i, c_i \sim \text{Normal}(0, \sigma^2) \quad \textbf{17.36}$$

$$w_i = \min(y_i,c_i). \quad \textbf{17.37}$$

En lugar de observar y_i, se observa sólo si es menor que un valor censurado, c_i. Observe que (17.36) incluye el supuesto de que u_i es independiente de c_i. (Para especificar, se considera explícitamente la censura desde arriba, o *censura a la derecha*; el problema de censura desde abajo, o *censura a la izquierda*, se maneja de manera similar.)

Un ejemplo de censura de datos a la derecha es la **codificación superior**. Cuando una variable se codifica de manera superior, se sabe que su valor sólo llega a cierto umbral o límite. Para respuestas mayores que el umbral, sólo se sabe que la variable es al menos tan grande como el umbral. Por ejemplo, en algunas encuestas la riqueza familiar se codifica de manera superior. Suponga que se les pregunta a los encuestados acerca de su riqueza, pero se les permite que respondan con "más de \$500,000". Entonces, se observa la riqueza real de aquellos encuestados cuya riqueza es menor que \$500,000, pero no de los encuestados cuya riqueza es mayor que \$500,000. En este caso el umbral de censura, c_i, es el mismo para toda i. En muchas situaciones, el umbral de censura cambia con las características individuales o familiares.

Pregunta 17.5

Sea mvp_i el valor del producto marginal para el trabajador i; este es el precio del bien de una empresa multiplicado por el producto marginal del trabajador. Suponga que mvp_i es una función lineal de las variables exógenas, como la educación, experiencia, etc., y un error no observable. En condiciones de competencia perfecta y sin restricciones institucionales, cada trabajador recibe como pago el valor de su producto marginal. Sea $minwage_i$ la variable que denota el salario mínimo para el trabajador i, el cual varía por estado. Se observa $wage_i$, el cual es el máximo entre mvp_i y $minwage_i$. Escriba el modelo apropiado del salario observado.

Si se observara una muestra aleatoria para ($\mathbf{x}$, y), simplemente se estimaría $\boldsymbol{\beta}$ por MCO, y la inferencia estadística sería la estándar. (Nuevamente, por simplicidad, se absorbe el intercepto en $\mathbf{x}$.) La censura causa problemas. Mediante argumentos similares a los del modelo Tobit, una regresión de MCO que sólo usa observaciones no censuradas, es decir, aquéllas con $y_i < c_i$, produce estimadores inconsistentes de β_j. Una regresión por MCO de w_i sobre $\mathbf{x}_i$, usando todas las observaciones, no estima de forma consistente β_j, a menos que no haya censura. Esto es similar al caso Tobit, pero el problema es muy diferente. En el modelo Tobit, se está modelando el comportamiento económico, que suele producir resultados de cero; el modelo Tobit se supone que refleja esto. Con la regresión censurada, se tiene un problema de recolección de datos debido a que, por alguna razón, los datos están censurados.

Bajo los supuestos en (17.36) y (17.37), se puede estimar $\boldsymbol{\beta}$ (así como σ^2) mediante máxima verosimilitud, dada una muestra aleatoria para ($\mathbf{x}_i$, w_i). Para esto, se necesita la densidad de w_i, dada ($\mathbf{x}_i$, c_i). Para las observaciones no censuradas, $w_i = y_i$, y la densidad de w_i es la misma que para y_i: Normal($\mathbf{x}_i\boldsymbol{\beta}$,$\sigma^2$). Para observaciones censuradas, se necesita la probabilidad de que w_i sea igual al valor de censura, c_i, dada $\mathbf{x}_i$:

$$\mathrm{P}(w_i = c_i|\mathbf{x}_i) = \mathrm{P}(y_i \geq c_i|\mathbf{x}_i) = \mathrm{P}(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma].$$

Se pueden combinar estas dos partes para obtener la densidad de w_i, dada $\mathbf{x}_i$ y c_i:

$$f(w|\mathbf{x}_i,c_i) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \quad w = c_i, \tag{17.38}$$

$$= (1/\sigma)\phi\,[(w - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \quad w < c_i. \tag{17.39}$$

La log-verosimilitud para la observación i se obtiene al obtener el logaritmo natural de la densidad para cada i. Se puede maximizar la suma de éstas a través de i, respecto a las β_j y σ, para obtener estimaciones de máxima verosimilitud.

Es importante saber que se pueden interpretar las β_j tal como en un modelo de regresión lineal bajo muestreo aleatorio. Esto es muy diferente a las aplicaciones Tobit con respuestas de solución de esquina, donde las expectativas de interés son funciones no lineales de las β_j.

Una aplicación importante de los modelos de regresión censurada es el **análisis de duración**. Una *duración* es una variable que mide el tiempo antes de que cierto evento ocurra. Por ejemplo, quizá se quiera explicar el número de días antes de que un criminal liberado de prisión vuelva a ser arrestado. Para algunos delincuentes, esto nunca pasará, o puede suceder después de tanto tiempo que es necesario censurar la duración con el fin de analizar los datos.

En las aplicaciones de duración en una regresión normal censurada, así como en la codificación superior, se suele utilizar el logaritmo natural como la variable dependiente, lo cual significa que también se obtiene el logaritmo del umbral de censura en (17.37). Como se ha visto a lo largo del libro, usar la transformación logaritmo para la variable dependiente ocasiona que los parámetros se interpreten como cambios porcentuales. Además, como con muchas variables positivas, el logaritmo de una duración por lo general tiene una distribución más cercana a la normal (condicional) que la duración misma.

Ejemplo 17.4

[Duración de la reincidencia]

El archivo RECID.RAW contiene datos sobre el tiempo en meses hasta que un recluso en una prisión de Carolina del Norte es arrestado después de haber sido liberado de prisión: llamemos a esto *durat*. Algunos reclusos participaron en un programa laboral mientras estaban en prisión. También se controla una variedad de variables demográficas, así como medidas de historia criminal y de prisión.

De 1,445 reclusos, 893 no habían sido arrestados durante el periodo de seguimiento; por tanto, estas observaciones están censuradas. Las duraciones censuradas difirieron entre los reclusos, con un rango de 70 a 81 meses.

La tabla 17.4 da los resultados de la regresión censurada normal para log(*durat*). Cada uno de los coeficientes, multiplicados por 100, dan el cambio porcentual estimado en la duración esperada dado un incremento *ceteris paribus* de una unidad en la variable explicativa correspondiente.

Varios coeficientes en la tabla 17.4 son interesantes. Las variables *priors* (número de condenas anteriores) y *tserved* (meses totales pasados en prisión) tienen efectos negativos sobre el periodo que transcurre hasta que ocurra un nuevo arresto. Esto sugiere que estas variables miden la tendencia de la actividad criminal más que representar un efecto disuasivo. Por ejemplo, un recluso con una condena previa adicional

TABLA 17.4

Estimación de regresión censurada de la reincidencia criminal

Variable dependiente: log(*durat*)	
Variables independientes	**Coeficiente (error estándar)**
workprg	−.063 (.120)
priors	−.137 (.021)
tserved	−.019 (.003)
felon	.444 (.145)
alcohol	−.635 (.144)
drugs	−.298 (.133)
black	−.543 (.117)
married	.341 (.140)
educ	.023 (.025)
age	.0039 (.0006)
constante	4.099 (.348)
Valor de log-verosimilitud $\hat{\sigma}$	−1,597.06 1.810

tiene un periodo hasta el siguiente arresto que es casi 14% menor. Un año adicional de tiempo en prisión reduce la duración por casi $100 \cdot 12(.019) = 22.8\%$. Un hallazgo un tanto sorprendente es que un hombre que purga una condena por un delito grave (*felon*) tiene un periodo esperado estimado que es casi 56% [$\exp(.444) - 1 \approx .56$] *mayor* que un hombre que purga una condena por un delito no grave.

Aquellos que tienen una historia de abuso de drogas y alcohol tienen duraciones esperadas sustancialmente menores hasta el siguiente arresto. (Las variables *alcohol* y *drugs* son variables binarias.) Los hombres de mayor edad (*age*) y los hombres que estaban casados (*married*) al momento de su condena en prisión, tienen periodos esperados más largos hasta su siguiente arresto. Los hombres negros (*black*) tienen periodos sustancialmente menores del orden de 42% [$\exp(-.543) - 1 \approx -.42$].

La variable de política clave, *workprg*, no tiene el efecto deseado. La estimación puntual es que, todo lo demás constante, los hombres que participaron en el programa laboral tienen periodos de reincidencia que son aproximadamente 6.3% menores que los hombres que no participaron. El coeficiente tiene un estadístico *t* pequeño, así que probablemente concluiríamos que el programa de trabajo no tiene efecto. Esto se puede deber a un problema de autoselección, o podría ser producto de la forma en que los hombres se reclutaron en el programa. Por supuesto, simplemente puede deberse a que el programa fue ineficaz.

En este ejemplo es crucial tener en cuenta la censura, en especial debido a que casi 62% de las duraciones eran censuradas. Si se aplica MCO de forma directa a toda la muestra y se tratan los periodos censurados como si fueran no censurados, los coeficientes estimados serán marcadamente diferentes. De hecho, todos se reducen hacia cero. Por ejemplo, el coeficiente de *priors* se convierte en $-.059$ (ee $=$ .009) y el de *alcohol* se convierte en $-.262$ (ee $=$ .060). Aunque las direcciones de los efectos son las mismas, la importancia de estas variables se ve disminuida en gran medida. Las estimaciones de regresión censurada son mucho más confiables.

Existen otras formas de medir los efectos de cada una de las variables explicativas de la tabla 17.4 sobre la duración, en lugar de enfocarse sólo en la duración esperada. Un análisis del tratamiento de la duración moderna está más allá del alcance de este libro. [Para una introducción sobre este tema, vea Wooldridge (2002, capítulo 20).]

Si cualquiera de los supuestos del modelo de regresión censurada normal se viola, en particular, si existe heterocedasticidad y no normalidad en u_i, las EMV por lo general serán inconsistentes. Esto muestra que la censura es potencialmente muy costosa, puesto que MCO que use una muestra no censurada no requiere normalidad ni homocedasticidad para ser consistente. Existen métodos que no requieren que se suponga una distribución, pero que son más avanzados [Vea Wooldridge (2002, capítulo 16).]

Modelos de regresión truncada

El modelo de regresión truncada difiere de manera importante del modelo de regresión censurada. En el caso de censura de datos, se *hace* una muestra aleatoria de la población. El problema de censura es que, mientras siempre se observan las variables explicativas de cada unidad extraída aleatoriamente, se observa el resultado en *y* sólo cuando no está censurada por encima o debajo de un umbral determinado. Con el truncamiento de datos, se restringe la atención a un subconjunto de la población antes del muestreo; así que hay una parte de la población para la cual no se observa información. En particular, no se tiene información sobre las variables explicativas. El escenario de muestreo truncado por lo general surge cuando una encuesta se enfoca en un subconjunto particular de la población y, quizá debido a consideraciones de costo, ignora por completo la otra parte de la población. Después, los investigadores podrían querer usar la muestra truncada para responder preguntas acerca de la población total, pero se debe reconocer que el esquema muestral no generó una muestra aleatoria de toda la población.

Como ejemplo, Hausman y Wise (1977) usaron datos de un experimento de impuestos negativos sobre la renta para estudiar los diferentes determinantes de ingresos. Para incluirse en el estudio, una familia debía tener un ingreso menor a 1.5 veces la línea de pobreza de 1967, donde ésta dependía del tamaño de la familia. Hausman y Wise quisieron usar los datos para estimar una ecuación de ingresos para toda la población.

El **modelo de regresión truncada normal** comienza con un modelo poblacional subyacente que satisface los supuestos del modelo lineal clásico:

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \; u|\mathbf{x} \sim \text{Normal}(0,\sigma^2). \qquad \boxed{17.40}$$

Recuerde que este es un fuerte conjunto de supuestos, debido a que u debe no sólo ser independiente de $\mathbf{x}$, sino distribuirse normalmente. El enfoque se centra en este modelo debido a que relajar los supuestos es difícil.

Bajo (17.40) se sabe que, dada una muestra aleatoria de la población, MCO es el procedimiento de estimación más eficiente. El problema surge debido a que no se observa una muestra aleatoria de la población: el supuesto RLM.2 se viola. En particular, una extracción aleatoria $(\mathbf{x}_i, y_i)$ se observa sólo si $y_i \leq c_i$, donde c_i es el umbral de truncamiento que puede depender de las variables exógenas, en particular, la $\mathbf{x}_i$. (En el ejemplo de Hausman y Wise, c_i depende del tamaño de la familia.) Esto significa que, si $\{(\mathbf{x}_i, y_i): i = 1, \ldots, n\}$ es nuestra muestra *observada*, entonces y_i es necesariamente menor o igual a c_i. Esto difiere del modelo de regresión censurada: en un modelo de regresión censurada, se observa $\mathbf{x}_i$ para cualquier observación extraída aleatoriamente de la población; en el modelo truncado, se observa sólo $\mathbf{x}_i$ si $y_i \leq c_i$.

Para estimar las β_j (junto con σ), se necesita la distribución de y_i, dado que $y_i \leq c_i$ y dada $\mathbf{x}_i$. Esto se escribe como

$$g(y|\mathbf{x}_i,c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)}, \quad y \leq c_i, \tag{17.41}$$

donde $f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ denota la densidad normal con media $\beta_0 + \mathbf{x}_i\boldsymbol{\beta}$ y varianza σ^2 y $F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ es la fda normal con la misma media y varianza, evaluadas en c_i. Esta expresión para la densidad, condicional en $y_i \leq c_i$, es lógica de manera intuitiva: es la densidad poblacional y, dada $\mathbf{x}$, dividida entre la probabilidad de que y_i sea menor que o igual a c_i (dada $\mathbf{x}_i$), $\mathrm{P}(y_i \leq c_i|\mathbf{x}_i)$. En efecto, se renormaliza la densidad al dividir entre el área bajo $f(\cdot|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ que es a la izquierda de c_i.

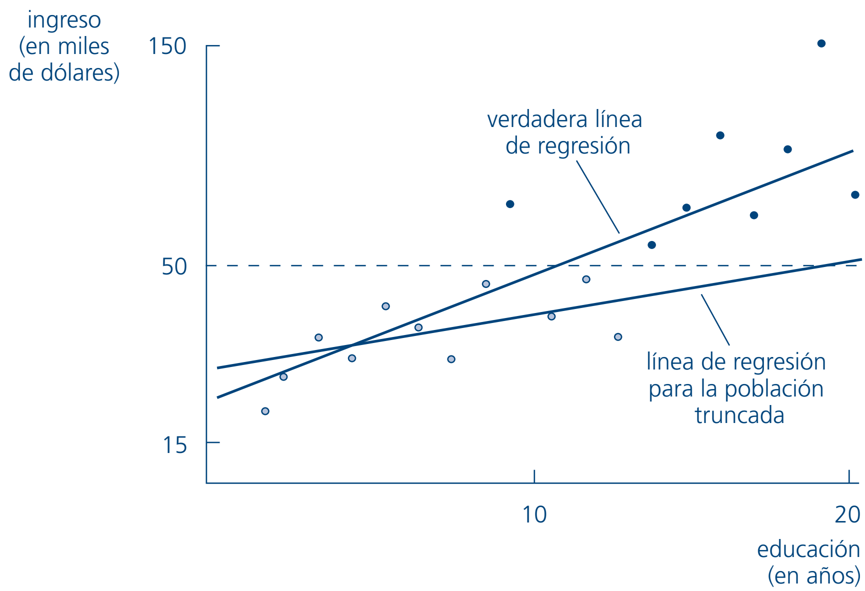
Si se obtiene el logaritmo de (17.41), se suma a través de toda i, y se maximiza el resultado respecto a las β_j y σ^2, se obtienen los estimadores de máxima verosimilitud. Esto produce estimadores consistentes y aproximadamente normales. La inferencia, incluidos los errores estándar y los estadísticos de log-verosimilitud, es la estándar.

Se pueden analizar los datos del ejemplo 17.4 como una muestra truncada si se descartan todos los datos de una observación si ésta está censurada. Esto daría 552 observaciones de una distribución normal truncada, donde el punto de truncamiento difiere a través de i. No obstante, nunca se analizarían los datos de duración (o datos de codificación superior) de esta forma, puesto que esto elimina información útil. El hecho que se conozca un límite inferior para 893 duraciones, junto con las variables explicativas, es información útil; la regresión censurada usa esta información, mientras que la regresión truncada no.

Un mejor ejemplo de regresión truncada se da en Hausman y Wise (1977), donde se enfatiza que MCO aplicados a una muestra truncada por arriba por lo general produce estimadores sesgados hacia cero. Intuitivamente, esto tiene sentido. Suponga que la relación de interés está entre los niveles de ingreso y educación. Si sólo se observan personas cuyo ingreso está por debajo de cierto umbral, se estará ignorando el extremo superior. Esto tiende a aplanar la línea estimada en relación con la verdadera línea de regresión de toda la población. La figura 17.4 ilustra el problema cuando el ingreso se trunca desde arriba en $50,000. Aunque se observan los puntos de datos representados por círculos abiertos, no se observan los conjuntos de datos representados por círculos oscurecidos. Un análisis de regresión que usa la muestra truncada no genera estimadores consistentes. De manera incidental, si la muestra en la figura 17.4 fuese censurada en vez de truncada, es decir, si se tuvieran datos de codificación superior, se observarían los niveles educativos en todos los puntos de la figura 17.4, pero para individuos con ingresos superiores a $50,000 no sabríamos la cantidad exacta de ingreso. Sólo se sabría que el ingreso fue de al menos $50,000. En efecto, todas las observaciones representadas por los círculos más oscuros se reducirían a la línea horizontal en *ingreso* = 50.

FIGURA 17.4

Una línea de regresión verdadera, o poblacional, y una línea de regresión incorrecta para la población truncada con ingresos inferiores a $50,000.



Como con la regresión censurada, si el supuesto de normalidad y homocedasticidad en (17.40) se viola, la EMV normal truncada es sesgada e inconsistente. Existen métodos que no requieren estos supuestos; vea Wooldridge (2002, capítulo 17) para análisis y referencias.

17.5 Correcciones de la selección muestral

La regresión truncada es un caso especial de un problema general conocido como **selección muestral no aleatoria**. Pero el diseño de la encuesta no es la única causa de la selección muestral no aleatoria. Con frecuencia, los encuestados no responden ciertas preguntas, lo cual lleva a la falta de datos para las variables dependientes o independientes. Debido a que no se pueden usar estas observaciones en la estimación, nos debemos preguntar si desecharlas ocasionará sesgo en nuestras estimaciones.

Otro ejemplo general suele denominarse **truncamiento incidental**. Aquí no se observa y debido al resultado de otra variable. El ejemplo principal es estimar la llamada *función de oferta salarial* de la economía laboral. El interés yace en cómo los diferentes factores, como la educación, afectan el salario que un individuo podría ganar en la fuerza laboral. Para las personas que trabajan, se observa la oferta salarial como el salario actual. Pero para aquellos que están actualmente fuera de la fuerza laboral, no se observa la oferta salarial. Debido a que trabajar puede estar sistemáticamente correlacionado con inobservables que afectan la oferta de trabajo, sólo usar gente trabajando, como hasta ahora se ha visto en los ejemplos relativos al salario, puede producir estimadores sesgados de los parámetros en la ecuación de la oferta salarial.

La selección no aleatoria también puede surgir cuando se tiene un panel de datos. En el caso más simple, se tienen dos años de datos, pero debido a la desaparición, algunas personas abandonan la muestra. Esto es un problema particularmente en el análisis de políticas, donde la desaparición puede estar relacionada con la efectividad de un programa.

¿Cuándo es consistente MCO sobre la muestra seleccionada?

En la sección 9.4 se ofreció una breve discusión de los tipos de selección aleatoria que se pueden ignorar. La distinción clave es entre la selección muestral *endógena* y *exógena*. En el caso de Tobit truncado, claramente se tiene una selección muestral endógena, y MCO es sesgado e inconsistente. Por otra parte, si la muestra está determinada sólo por una variable explicativa exógena, se tiene una selección muestral exógena. Los casos entre estos extremos son poco claros y ahora se ofrecerán definiciones detalladas y supuestos para ellas. El modelo poblacional es

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \ \mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0. \tag{17.42}$$

Es útil escribir el modelo poblacional para una extracción *aleatoria* como

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{17.43}$$

donde se utiliza $\mathbf{x}_i\boldsymbol{\beta}$ como abreviatura de $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. Ahora, sea n el tamaño de la *muestra aleatoria* de la población. Si pudiéramos observar y_i y cada x_{ij} para toda i, simplemente se usaría MCO. Suponga que, por alguna razón, ya sea y_i o algunas variables independientes no se observan para cierta i. En al menos algunas observaciones se observa el conjunto completo de variables. Defina un *indicador de selección* s_i para cada i por $s_i = 1$ si se observa todo $(y_i, \mathbf{x}_i)$ y $s_i = 0$ en otro caso. Por tanto, $s_i = 1$ indica que se utilizará la observación en el análisis; $s_i = 0$ significa que la observación no se utilizará. Estamos interesados en las propiedades estadísticas de los estimadores de MCO usando la **muestra seleccionada**, es decir, usando las observaciones para las cuales $s_i = 1$. Por tanto, se utilizan menos que n observaciones, por ejemplo, n_1.

Resulta fácil obtener las condiciones bajo las cuales MCO es consistente (e incluso insesgado). En efecto, en lugar de estimar (17.43) se puede estimar sólo la ecuación

$$s_i y_i = s_i\mathbf{x}_i\boldsymbol{\beta} + s_i u_i. \tag{17.44}$$

Cuando $s_i = 1$, simplemente se tiene (17.43); cuando $s_i = 0$, simplemente se tiene $0 = 0 + 0$, lo cual por supuesto no dice nada acerca de $\boldsymbol{\beta}$. Realizar la regresión de $s_i y_i$ sobre $s_i\mathbf{x}_i$ para $i = 1, 2, \ldots, n$ es lo mismo que regresar y_i sobre $\mathbf{x}_i$ usando las observaciones para las cuales $s_i = 1$. Por tanto, se puede saber de la consistencia de las $\hat{\beta}_j$ al estudiar (17.44) para una muestra aleatoria.

A partir del análisis del capítulo 5, los estimadores de MCO de (17.44) son consistentes si el término de error tiene una media cero y no está correlacionado con la variable explicativa. En la población, el supuesto de la media cero es $\mathrm{E}(su) = 0$, y los supuestos de cero correlación se pueden expresar como

$$\mathrm{E}[(sx_j)(su)] = \mathrm{E}(sx_j u) = 0, \tag{17.45}$$

donde s, x_j y u son variables aleatorias que representan la población; se ha usado el hecho de que $s^2 = s$ debido a que s es una variable binaria. La condición (17.45) es diferente de la que se necesita si se observan todas las variables de una muestra aleatoria: $\mathrm{E}(x_j u) = 0$. Por tanto, en la población, se necesita que u no esté correlacionada con sx_j.

La condición clave para la insesgadez es $E(su|sx_1, \ldots, sx_k) = 0$. Como de costumbre, es un supuesto más fuerte que el que se necesita para la consistencia.

Si s está en función sólo de las variables explicativas, entonces sx_j es sólo una función de x_1, x_2, …, x_k; por el supuesto de la media condicional en (17.42), sx_j tampoco está correlacionado con u. De hecho, $E(su|sx_1, \ldots, sx_k) = sE(u|sx_1, \ldots, sx_k) = 0$, debido a que $E(u|x_1, \ldots, x_k) = 0$. Este es el caso de la **selección muestral exógena**, donde $s_i = 1$ está determinada completamente por $x_{i1}, \ldots, x_{ik}$. Como ejemplo, si se está interesado en una ecuación salarial donde las variables explicativas sean la educación, la experiencia, la antigüedad, el sexo, el estado civil, etc. (que se suponen exógenas) se podrá seleccionar la muestra con base en cualquiera o en todas las variables explicativas.

Si la selección muestral es enteramente aleatoria en el sentido de que s_i es *independiente* de $(\mathbf{x}_i, u_i)$, entonces $E(sx_ju) = E(s)E(x_ju) = 0$, debido a que $E(x_ju) = 0$ bajo (17.42). Por tanto, si se comienza con una muestra aleatoria y se descartan observaciones aleatoriamente, MCO siguen siendo consistentes. De hecho, MCO nuevamente es insesgado en este caso, siempre y cuando no haya multicolinealidad perfecta en la muestra seleccionada.

Si s depende de las variables explicativas y de términos aleatorios adicionales independientes de $\mathbf{x}$ y u, MCO también son consistentes e insesgados. Por ejemplo, suponga que el puntaje de IQ es una variable explicativa en la ecuación salarial, pero el dato IQ falta en algunas personas. Suponga que se cree que la selección se puede describir mediante $s = 1$ si $IQ \geq v$, y $s = 0$ si $IQ < v$, donde v es una variable aleatoria inobservable que es independiente de IQ, u y las otras variables explicativas. Esto significa que es más probable que se observen IQ altos, pero siempre habrá alguna probabilidad de no observar ninguno. De manera condicional sobre las variables explicativas, s es independiente de u, lo cual significa que $E(u|x_1, \ldots, x_k, s) = E(u|x_1, \ldots, x_k)$, y la última expectativa es cero mediante el supuesto en el modelo poblacional. Si se agrega el supuesto de homocedasticidad $E(u^2|\mathbf{x},s) = E(u^2) = \sigma^2$, entonces los errores estándar usuales de MCO y los estadísticos de prueba son válidos.

Hasta ahora se han mostrado varias situaciones en donde MCO sobre la muestra seleccionada son insesgados, o al menos consistentes. ¿Cuándo es inconsistente MCO sobre la muestra seleccionada? Ya se vio un ejemplo: la regresión con una muestra trunca. Cuando el truncamiento proviene de arriba, $s_i = 1$ si $y_i \leq c_i$, donde c_i es el umbral de truncamiento. De forma equivalente, $s_i = 1$ si $u_i \leq c_i - \mathbf{x}_i\boldsymbol{\beta}$. Debido a que s_i depende directamente de u_i, ocurre que s_i y u_i estarán correlacionados, incluso de manera condicional sobre $\mathbf{x}_i$. Esta es la razón de que MCO sobre la muestra seleccionada no estime de forma consistente las β_j. Existen formas menos evidentes de que s y u puedan correlacionarse, las cuales se considerarán en la siguiente subsección.

Los resultados sobre la consistencia de MCO se extienden a la estimación de variables instrumentales. Si las VI se denotan como z_h en la población, la condición clave para la consistencia de MC2E es $E(sz_hu) = 0$, la cual aplica si $E(u|\mathbf{z},s) = 0$. Por tanto, si la selección se determina por completo por las variables exógenas $\mathbf{z}$, o si s depende de otros factores independientes de u y $\mathbf{z}$, entonces MC2E sobre la muestra seleccionada por lo general es consistente. Es necesario suponer que las variables explicativas e instrumentales están correlacionadas de manera adecuada en la parte seleccionada de la población. Wooldridge (2002, capítulo 17) contiene afirmaciones precisas de estos supuestos.

Se puede mostrar también que, cuando la selección es completamente una función de las variables exógenas, la estimación de máxima verosimilitud de un modelo no ineal, como el modelo logit o probit, produce estimadores consistentes y asintóticamente normales y los errores estándar y estadísticos de prueba usuales son válidos. [Nuevamente vea Wooldridge (2002, capítulo 17).]

Truncamiento incidental

Como se mencionó antes, una forma común de selección muestral se llama truncamiento incidental. De nuevo, se empieza con el modelo poblacional en (17.42). No obstante, se supone que siempre se observan las variables explicativas x_j. El problema es que sólo se observa y para un subconjunto de la población. La regla que determina si se observa y *no* depende directamente del resultado de y. Un ejemplo guía es cuando $y = \log(wage^o)$, donde $wage^o$ es la *oferta salarial*, o el salario por hora que un individuo podría recibir en el mercado laboral. Si la persona está trabajando en el momento de la encuesta, entonces se observa la oferta salarial debido a que se supone que es el salario observado. Pero para las personas ajenas a la fuerza laboral, no se puede observar $wage^o$. Por tanto, el truncamiento de la oferta salarial es *incidental* debido a que depende de otra variable, a saber, la participación en la fuerza laboral. Es importante anotar que por lo general se observaría todo el resto de la información acerca de un individuo, como la educación, experiencia previa, sexo, estado civil, etcétera.

El método acostumbrado para el truncamiento incidental es agregar una ecuación explícita de selección al modelo poblacional de interés:

$$y = \mathbf{x}\boldsymbol{\beta} + u, \ \mathrm{E}(u|\mathbf{x}) = 0 \tag{17.46}$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0], \tag{17.47}$$

donde $s = 1$ si se observa y, y es cero de otra forma. Se supone que los elementos de $\mathbf{x}$ y $\mathbf{z}$ siempre se observan y se escribe $\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ y $\mathbf{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_m z_m$.

La ecuación que más interesa es (17.46) y se podría estimar $\boldsymbol{\beta}$ mediante MCO dada una muestra aleatoria. La ecuación de selección, (17.47), depende de las variables observables, z_h, y de un error inobservable, v. Un supuesto estándar que se hará es que $\mathbf{z}$ es exógena en (17.46):

$$\mathrm{E}(u|\mathbf{x},\mathbf{z}) = 0.$$

De hecho, para que los siguientes métodos propuestos funcionen bien, se requerirá que $\mathbf{x}$ sea un subconjunto estricto de $\mathbf{z}$: cualquier x_j es también un elemento de $\mathbf{z}$, y se tienen algunos elementos de $\mathbf{z}$ que no están en $\mathbf{x}$. Se verá más adelante por qué esto es crucial.

Se supone que el término de error v en la ecuación de selección muestral es independiente de $\mathbf{z}$ (y por tanto de $\mathbf{x}$). También se supone que v tiene una distribución normal estándar. Se puede ver fácilmente que la correlación entre u y v por lo general ocasiona un problema de selección muestral. Para ver por qué, se supone que (u, v) es independiente de $\mathbf{z}$. Después, tomando la expectativa de (17.46), condicional sobre $\mathbf{z}$ y v, y al usar el hecho de que $\mathbf{x}$ es un subconjunto de $\mathbf{z}$ se tiene

$$\mathrm{E}(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|v),$$

donde $\mathrm{E}(u|\mathbf{z},v) = \mathrm{E}(u|v)$ debido a que (u, v) es independiente de $\mathbf{z}$. Ahora, si u y v son conjuntamente normales (con media cero), entonces $\mathrm{E}(u|v) = \rho v$ para algún parámetro ρ. Por tanto,

$$\mathrm{E}(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \rho v.$$

No se observa v, pero se puede usar esta ecuación para calcular $\mathrm{E}(y|\mathbf{z},s)$ y después especializar esto en $s = 1$. Ahora se tiene:

$$\mathrm{E}(y|\mathbf{z},s) = \mathbf{x}\boldsymbol{\beta} + \rho\mathrm{E}(v|\mathbf{z},s).$$

Debido a que s y v están relacionadas mediante (17.47), y v tiene una distribución normal estándar, se puede mostrar que $\mathrm{E}(v|\mathbf{z},s)$ es simplemente la razón inversa de Mills, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, cuando $s = 1$. Esto genera la importante ecuación

$$\mathrm{E}(y|\mathbf{z},s = 1) = \mathbf{x}\beta + \rho\lambda(\mathbf{z}\boldsymbol{\gamma}). \tag{17.48}$$

La ecuación (17.48) muestra que el valor esperado de y, dado $\mathbf{z}$ *así como* la observabilidad de y, es igual a $\mathbf{x}\boldsymbol{\beta}$, más un término adicional que depende de la razón inversa de Mills evaluada en

$\mathbf{z}\boldsymbol{\gamma}$. Recuerde que se espera estimar $\boldsymbol{\beta}$. Esta ecuación muestra que se puede lograr usando sólo la muestra seleccionada, siempre y cuando se incluya el término $\lambda(\mathbf{z}\boldsymbol{\gamma})$ como regresor adicional.

Si $\rho = 0$, $\lambda(\mathbf{z}\boldsymbol{\gamma})$ no aparece, y MCO de y en $\mathbf{x}$ usando la muestra seleccionada, estima consistentemente $\boldsymbol{\beta}$. De otra forma, se habrá omitido efectivamente una variable, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, que por lo general está correlacionada con $\mathbf{x}$. ¿Cuándo ocurre que $\rho = 0$? La respuesta es cuando u y v no estén correlacionadas.

Debido a que $\boldsymbol{\gamma}$ se desconoce, no se puede evaluar $\lambda(\mathbf{z}_i\boldsymbol{\gamma})$ para cada i. No obstante, a partir de los supuestos formulados, s dada z sigue un modelo probit:

$$P(s = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\gamma}). \tag{17.49}$$

Por tanto, se puede estimar $\boldsymbol{\gamma}$ mediante un probit de s_i en $\mathbf{z}_i$, usando la muestra *entera*. En un segundo paso se puede estimar $\boldsymbol{\beta}$. Se resume el procedimiento, que recién se ha dado en llamar el **método Heckit** en la literatura econométrica en honor al trabajo de Heckman (1976).

Corrección de la selección muestral:

i) Usando todas las n observaciones se estima un modelo probit de s_i sobre $\mathbf{z}_i$ y se obtienen las estimaciones $\hat{\gamma}_h$. Se calcula la razón inversa de Mills, $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$ para cada i. (En realidad, se necesitan éstas sólo para la i con $s_i = 1$.)

ii) Se usa la muestra seleccionada, es decir, las observaciones para las cuales $s_i = 1$ (por ejemplo, n_1 de ellas), se ejecuta la regresión de

$$y_i \text{ sobre } \mathbf{x}_i, \hat{\lambda}_i. \tag{17.50}$$

Las $\hat{\beta}_j$ son consistentes y se distribuyen de manera aproximadamente normal.

Una prueba simple para el sesgo de selección está disponible de la regresión (17.50). A saber, se puede usar el estadístico t usual de $\hat{\lambda}_i$ como prueba de H_0: $\rho = 0$. Bajo H_0, no existe un problema de selección muestral.

Cuando $\rho \neq 0$, los errores estándar MCO usuales reportados de (17.50) no son exactamente correctos. Esto es porque no dan cuenta de la estimación de $\boldsymbol{\gamma}$, la cual usa las mismas observaciones en la regresión (17.50) y más. Algunos paquetes econométricos calculan los errores estándar corregidos. [Por desgracia, eso no es tan simple como un ajuste de heterocedasticidad. Vea Wooldridge (2002, capítulo 6) para un análisis más detallado.] En muchos casos, los ajustes no generan diferencias importantes, pero es difícil saberlo con anticipación (a menos que $\hat{\rho}$ sea pequeña e insignificante).

Recién se mencionó que $\mathbf{x}$ debe ser un subconjutno estricto de $\mathbf{z}$. Esto tiene dos implicciones. Primero, cualquier elemento que aparezca como variable explicativa en (17.46) debe ser una variable explicativa en la ecuación de selección. Aunque en casos raros es lógico excluir elementos de la ecuación de selección, incluir todos los elementos de $\mathbf{x}$ en $\mathbf{z}$ no es muy costoso; excluirlos puede ocasionar una inconsistencia si son erróneamente excluidos.

Una segunda implicación importante es que se tiene al menos un elemento de $\mathbf{z}$ que no está en $\mathbf{x}$. Esto significa que se necesita una variable que afecte la selección, pero que *no* tenga un efecto parcial en y. Esto no es absolutamente necesario para aplicar el procedimiento (de hecho, se pueden realizar mecánicamente los dos pasos cuando $\mathbf{z} = \mathbf{x}$) pero los resultados por lo general son menos que convicentes a menos que se tenga una *restricción de exclusión* en (17.46). La razón de esto es que mientras la razón inversa de Mills es una función no lineal de $\mathbf{z}$, suele aproximarse bien mediante una función lineal. Si $\mathbf{z} = \mathbf{x}$, $\hat{\lambda}_i$ puede estar altamente correlacionada con los elementos de $\mathbf{x}_i$. Como se sabe, tal multicolinealidad puede generar errores estándar muy altos para las $\hat{\beta}_j$. Intuitivamente, si no se cuenta con una variable que afecte la selección, pero no a y, es en extremo difícil, si no imposible, distinguir la selección muestral de una forma funcional mal especificada en (17.46).

Ejemplo 17.5

[Ecuación de la oferta salarial para mujeres casadas]

Se aplica la corrección de selección muestral a los datos sobre las mujeres casadas en MROZ.RAW. Recuerde que de las 753 mujeres en la muestra, 428 trabajan por un salario durante el año. La ecuación de oferta salarial es la estándar, con log(*wage*) como la variable dependiente y *educ*, *exper* y $exper^2$ como las variables explicativas. Con el fin de probar y corregir el sesgo de selección muestral, debido a la inobservabilidad de la oferta salarial para mujeres que no están trabajando, se necesita estimar un modelo probit para la participación en la fuerza laboral. Además de las variables de educación y experiencia, se incluyen los factores en la tabla 17.1: otros ingresos, edad, número de hijos pequeños y número de hijos mayores. El hecho de que estas cuatro variables estén excluidas de la ecuación de oferta salarial es un *supuesto*: se supone que, dados los factores de productividad, *nwifeinc*, *age*, *kidslt6* y *kidsge6* no tienen efecto sobre la oferta salarial. Resulta claro de los resultados probit en la tabla 17.1 que al menos las variables *age* y *kidslt6* tienen un fuerte efecto sobre la participación en la fuerza laboral.

La tabla 17.5 contiene los resultados de MCO y Heckit. [Los errores estándar reportados para los resultados Heckit son sólo los errores estándar de MCO usuales de la regresión 17.50).] No hay evidencia de un problema de selección muestral en la estimación de la ecuación de oferta salarial. El coeficiente de $\hat{\lambda}$ tiene un estadístico t muy pequeño (.239), así que no se puede rechazar H_0: $\rho = 0$. Tan importante como lo anterior es que prácticamente no hay grandes diferencias en los coeficientes asociados a las pendientes estimadas en la tabla 17.5. Los rendimientos estimados de la educación difieren por sólo una décima parte de un punto porcentual.

TABLA 17.5

Ecuación de oferta salarial para mujeres casadas

Variable dependiente: log(*wage*)		
Variables independientes	**MCO**	**Heckit**
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
$exper^2$	−.00081 (.00039)	−.00086 (.00044)
constante	−.522 (.199)	−.578 (.307)
$\hat{\lambda}$	—	.032 (.134)
Tamaño de muestra *R*-cuadrada	428 .157	428 .157

Como una alternativa al método anterior de estimación de dos pasos está la estimación completa de máxima verosimilitud. Es más complicado puesto que requiere obtener la distribución conjunta de y y s. Suele ser lógico probar la selección muestral mediante el procedimiento anterior; si no hay evidencia de selección muestral, no hay razón para continuar. Si se detecta un sesgo en la selección muestral, se pueden usar las estimaciones de dos pasos o estimar las ecuaciones de regresión y de selección de manera conjunta mediante EMV. [Vea Wooldridge (2002, capítulo 17).]

En el ejemplo 17.5 se sabe más que sólo si una mujer trabajó durante el año: se sabe cuántas horas trabajó cada mujer. Resulta que se puede usar esta información como un procedimiento de selección muestral alternativo. En lugar de la razón inversa de Mills, $\hat{\lambda}_i$, se usan los residuales Tobit, por ejemplo, $\hat{v}_i$, los cuales se calculan como $\hat{v}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ siempre que $y_i > 0$. Se puede mostrar que la regresión en (17.50) con $\hat{v}_i$ en lugar de $\hat{\lambda}_i$ también produce estimaciones consistentes de las β_j, y el estadístico t estándar de $\hat{v}_i$ es una prueba válida para el sesgo de la selección muestral. Este enfoque tiene la ventaja de usar más información, pero es menos aplicable en general. [Vea Wooldridge (2002, capítulo 17).]

Existen muchos más temas concernientes a la selección muestral. Uno que vale la pena mencionar es el modelo con variables explicativas endógenas *en adición* al posible sesgo de selección muestral. Escriba un modelo con una sola variable explicativa como

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1, \tag{17.51}$$

donde y_1 sólo se observa cuando $s = 1$, y y_2 puede observarse sólo junto con y_1. Un ejemplo es cuando y_1 es el porcentaje de votos recibidos por un funcionario que busca la reelección y y_2 es el porcentaje del total de gastos de campaña que realiza. Para los funcionarios que no participan en la elección, no se puede observar y_1 o y_2. Si se tienen factores exógenos que afecten la decisión de participar y que estén correlacionados con los gastos de campaña, se puede estimar consistemente α_1 y los elementos de $\boldsymbol{\beta}_1$ por variables instrumentales. Para ser convincentes, son necesarias *dos* variables exógenas que no aparezcan en (17.51). Ciertamente, una afectará la decisión de selección y otra estará correlacionada con y_2 [el requisito acostumbrado para estimar (17.51) por MC2E]. En breve, el método es para estimar la ecuación de selección mediante probit, donde *todas* las variables exógenas aparecen en la ecuación probit. Entonces, se agrega la razón inversa de Mills a (17.51) y se estima la ecuación mediante MC2E. La razón inversa de Mills actúa como su propio instrumento, puesto que depende sólo de las variables exógenas. Se usan todas las variables exógenas como los demás instrumentos. Como antes, se puede usar el estadístico t de $\hat{\lambda}_i$ como una prueba del sesgo de selección. [Vea Wooldridge (2002, capítulo 17) para mayor información.]

RESUMEN

En este capítulo se han cubierto varios métodos avanzados que suelen usarse en aplicaciones, en especial en microeconomía. Los modelos logit y probit se usan para las variables de respuesta binaria. Estos modelos tienen algunas ventajas sobre el modelo de probabilidad lineal: las probabilidades ajustadas están entre cero y uno, y los efectos parciales son decrecientes. La principal desventaja de logit y probit es que son más difíciles de interpretar.

El modelo Tobit es aplicable a resultados no negativos que se acumulan en cero pero que también asumen un amplio rango de valores positivos. Muchas variables de elección individual, como la oferta laboral, cantidad del seguro de vida y cantidad del fondo de pensiones invertida en acciones, tienen esta característica. Como con logit y probit, los valores esperados de y dada $\mathbf{x}$, ya sea condicional en $y > 0$ o no condicional, dependen de $\mathbf{x}$ y $\boldsymbol{\beta}$ en formas no lineales. Se dieron las expresiones para estas expectativas así como fórmulas para los efectos parciales de cada x_j sobre las expectativas. Éstas se pueden estimar después de que se ha estimado el modelo Tobit mediante máxima verosimilitud.

Cuando la variable dependiente es una variable de conteo, es decir, asume valores enteros no negativos, un modelo de regresión Poisson es adecuado. El valor esperado de y dada x_j tiene una forma exponencial. Esto da las interpretaciones de parámetro como semielasticidades o elasticidades, dependiendo de si x_j está en nivel o en forma logarítmica. En resumen, se pueden interpretar los parámetros *como si* estuvieran en un modelo lineal con log(y) como la variable dependiente. Los parámetros se pueden estimar mediante EMV. No obstante, debido a que la distribución de Poisson impone igualdad de la varianza y la media, suele ser necesario calcular los errores estándar y estadísticos de prueba que permitan una sobre o subdispersión. Éstos son simples ajustes a los estadísticos y errores estándar de EMV.

Los modelos de regresión censurada y truncada manejan clases específicas de problemas de datos faltantes. En la regresión censurada, la variable dependiente se censura por encima o debajo de un umbral. Se puede usar la información de los resultados censurados debido a que siempre se observan las variables explicativas, como en las aplicaciones de duración o la codificación superior de las observaciones. Un modelo de regresión truncada surge cuando parte de la población se excluye por completo, no se observa información sobre las unidades que no estén cubiertas por el esquema de muestreo. Este es un caso especial de problema de selección muestral.

La sección 17.5 ofrece un tratamiento sistemático de la selección muestral no aleatoria. Se mostró que la selección muestral exógena no afecta la consistencia de MCO cuando se aplican a la submuestra, pero la selección muestral endógena sí lo hace. Se mostró cómo probar y corregir el sesgo de selección muestral para el problema general de truncamiento incidental, donde las observaciones faltan en y debido al resultado de otra variable (como la participación en la fuerza laboral). El método de Heckman es relativamente fácil de implementar en estas situaciones.

TÉRMINOS CLAVE

Análisis de duración
Codificación superior
Distribución de Poisson
Efecto parcial al promedio (EPA)
Efecto parcial al promedio (EPP)
Estadístico de razón de cuasi verosimilitudes
Estadístico de razón de verosimilitudes
Estadístico de Wald
Estimación de cuasi máxima verosimilitud (ECMV)
Estimación de máxima verosimilitud (EMV)
Función de log-verosimilitud
Método Heckit
Modelo de regresión censurada
Modelo de regresión de Poisson
Modelo de regresión normal censurada
Modelo de regresión normal truncada
Modelo de regresión truncada
Modelo de variable latente
Modelo logit
Modelo probit
Modelo Tobit
Modelos de respuesta binaria
Muestra seleccionada
Porcentaje predicho correctamente
Probabilidad de respuesta
Pseudo R-cuadrada
Razón inversa de Mills
Respuesta de solución de esquina
Selección muestral exógena
Selección muestral no aleatoria
Sobredispersión
Truncamiento incidental
Variable de conteo
Variable dependiente limitada (VDL)

PROBLEMAS

17.1 i) Para una respuesta binaria y, sea $\bar{y}$ la proporción de unos en la muestra (que es igual al promedio muestral de y_i). Sea $\hat{q}_0$ el porcentaje predicho correctamente para el resultado $y = 0$ y sea $\hat{q}_1$ el porcentaje correctamente predicho para el resultado $y = 1$. Si $\hat{p}$ es el porcentaje general predicho correctamente, muestre que $\hat{p}$ es un promedio ponderado de $\hat{q}_0$ y $\hat{q}_1$:

$$\hat{p} = (1 - \bar{y})\,\hat{q}_0 + \bar{y}\hat{q}_1.$$

ii) En una muestra de 300, suponga que $\bar{y} = .70$, de manera que hay 210 resultados con $y_i = 1$ y 90 con $y_i = 0$. Suponga que el porcentaje correctamente predicho cuando $y = 0$ es 80, y el porcentaje correctamente predicho cuando $y = 1$ es 40. Determine el porcentaje correctamente predicho.

17.2 Sea *grad* una variable binaria para si un atleta colegial en una universidad grande se graduará en cinco años. Sean *hsGPA* y *SAT* el promedio de calificaciones de bachillerato y las puntuaciones del SAT de admisión a la universidad, respectivamente. Sea *study* el número de horas por semana que pasa un estudiante en un aula de estudio. Suponga que, usando los datos sobre 420 atletas colegiales se obtiene el siguiente modelo logit:

$$\hat{\mathrm{P}}(grad = 1|hsGPA,SAT,study) = \Lambda(-1.17 + .24\; hsGPA + .00058\; SAT + .073\; study),$$

donde $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ es la función logit. Si se mantiene *hsGPA* en 3.0 y el *SAT* fijo en 1,200, calcule la diferencia estimada en la probabilidad de graduación para alguien que pasa 10 horas a la semana en el aula de estudio y alguien que pasa 5 horas por semana.

17.3 (Requiere cálculo)

i) Suponga en el modelo Tobit que $x_1 = \log(z_1)$ y este es el único lugar en el que z_1 aparece en $\mathbf{x}$. Muestre que

$$\frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial z_1} = (\beta_1/z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}, \tag{17.52}$$

donde β_1 es el coeficiente de $\log(z_1)$.

ii) Si $x_1 = z_1$ y $x_2 = z_1^2$, demuestre que

$$\frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial z_1} = (\beta_1 + 2\beta_2 z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\},$$

donde β_1 es el coeficiente de z_1 y β_2 es el coeficiente de z_1^2.

17.4 Sea mvp_i el valor del producto marginal para el trabajador i, que es el precio del bien de una empresa multiplicado por el producto marginal del trabajador. Suponga que

$$\log(mvp_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$
$$wage_i = \max(mvp_i, minwage_i),$$

donde las variables explicativas incluyen educación, experiencia, etc., y $minwage_i$ es el salario mínimo relevante para la persona i. Escriba $\log(wage_i)$ en términos de $\log(mvp_i)$ y $\log(minwage_i)$.

17.5 (Requiere cálculo). Sea *patents* el número de patentes solicitadas por una empresa durante un año determinado. Suponga que la expectativa condicional de *patents* dadas *sales* y *RD* es

$$\mathrm{E}(patents|sales,RD) = \exp[\beta_0 + \beta_1 \log(sales) + \beta_2 RD + \beta_3 RD^2],$$

donde *sales* son las ventas anuales de la empresa y *RD* es el gasto total en investigación y desarrollo durante los pasados 10 años.

i) ¿Cómo estimaría las β_j? Justifique su respuesta mediante un análisis de la naturaleza de *patents*.

ii) ¿Cómo interpreta usted β_1?

iii) Encuentre el efecto parcial de *RD* en $\mathrm{E}(patents|sales,RD)$.

17.6 Considere la función de ahorro de una familia para la población de todas las familias en Estados Unidos:

$$sav = \beta_0 + \beta_1 inc + \beta_2 hhsize + \beta_3 educ + \beta_4 age + u,$$

donde *hhsize* es el tamaño de la familia, *educ* son los años de educación del jefe de familia y *age* es la edad del jefe de familia. Suponga que $\mathrm{E}(u|inc,hhsize,educ,age) = 0$.

i) Suponga que la muestra incluye sólo a familias cuyo jefe tiene más de 25 años de edad. Si se usa MCO en tal muestra, ¿se obtienen estimadores insesgados de las β_j? Explique.

ii) Ahora, suponga que nuestra muestra incluye sólo a parejas casadas sin hijos. ¿Se pueden estimar todos los parámetros en la ecuación de ahorro? ¿Cuáles se pueden estimar?

iii) Suponga que se excluyen las familias de la muestra que ahorran más de $25,000 al año. ¿MCO produce estimadores consistentes de las β_j?

17.7 Suponga que lo contrata una universidad para estudiar los factores que determinan si un estudiante admitido a la universidad en realidad asiste a ella. Se le da una muestra aleatoria grande de estudiantes que fueron admitidos el año anterior. Se tiene información sobre si cada estudiante eligió asistir, el desempeño en el bachillerato, el ingreso familiar, la asistencia financiera ofrecida, la raza y las variables geográficas. Alguien le dice: "Cualquier análisis de los datos generará resultados sesgados debido a que ésta no es una muestra aleatoria de todos los que solicitaron su ingreso a la universidad, sino sólo de aquellos que lo solicitaron para esta universidad". ¿Qué piensa usted de tal crítica?

EJERCICIOS EN COMPUTADORA

C17.1 Use los datos en PNTSPRD.RAW para este ejercicio.

i) La variable *favwin* es una variable binaria si el equipo favorecido por la diferencia de puntos (*spread*) predicha por Las Vegas gana. Un modelo de probabilidad lineal para estimar si el equipo favorecido gana es

$$\mathrm{P}(favwin = 1|spread) = \beta_0 + \beta_1 spread.$$

Explique por qué, si la diferencia de puntos (*spread*) incorpora toda la información relevante, se espera $\beta_0 = .5$.

ii) Estime el modelo de la parte i) mediante MCO. Pruebe H_0: $\beta_0 = .5$ frente a una alternativa de dos colas. Use los errores estándar usuales y los robustos a la heterocedasticidad.

iii) ¿La variable *spread* es estadísticamente significativa? ¿Cuál es la probabilidad estimadas de que el equipo favorito gane cuando $spread = 10$?

iv) Ahora estime un modelo probit para P(*favwin* = 1|*spread*). Interprete y pruebe la hipótesis nula de que el intercepto es cero. [*Sugerencia*: recuerde que $\Phi(0) = .5$.]

v) Use el modelo probit para estimar la probabilidad de que el equipo favorito gane cuando *spread* = 10. Compare esto con la estimación del MPL de la parte iii).

vi) Agregue las variables *favhome*, *fav25* y *und25* al modelo probit y pruebe la significancia conjunta de estas variables mediante la prueba de razón de verosimilitudes. (¿Cuántos *gl* hay en la distribución ji-cuadrada?) Interprete este resultado enfocándose en la pregunta de si la diferencia de puntos incorpora toda la información observable antes de un juego.

C17.2 Use los datos en LOANAPP.RAW para este ejercicio; vea también el ejercicio en computadora C7.8.

i) Estime un modelo probit de *approve* sobre *white*. Encuentre la probabilidad de que se apruebe un préstamo tanto para blancos como para no blancos. ¿Cómo se compara esto con las estimaciones de probabilidad lineal?

ii) Ahora agregue las variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2* y *vr* al modelo probit. ¿Hay alguna evidencia estadísticamente significativa de discriminación contra los no blancos?

iii) Estime el modelo de la parte ii) mediante logit. Compare el coeficiente de *white* respecto a la estimación probit.

iv) Use la ecuación (17.17) para estimar las dimensiones de los efectos discriminativos para probit y logit.

C17.3 Use los datos en FRINGE.RAW para este ejercicio.

i) ¿Para qué porcentaje de trabajadores en la muestra *pension* es igual a cero? ¿Cuál es el rango de *pension* para trabajadores con beneficios de pensión no iguales a cero? ¿Por qué un modelo Tobit es apropiado para modelar *pension*?

ii) Estime un modelo Tobit que explique *pension* en términos de *exper* (experiencia), *age* (edad), *tenure* (antigüedad), *educ* (educación), *depends* (dependientes), *married* (estado civil casado), *white* (blanco) y *male* (hombre). ¿Los hombres y los blancos tienen prestaciones de pensión esperadas significativamente más altas?

iii) Use los resultados de la parte ii) para estimar la diferencia en las prestaciones de pensión esperadas para los hombres blancos y para las mujeres no blancas, ambos con 35 años de edad, solteros sin dependientes, con 16 años de educación y con 10 de experiencia.

iv) Agregue *union* (variable binaria igual a uno si pertenece a un sindicato) al modelo Tobit y comente sobre su importancia.

v) Aplique el modelo Tobit de la parte iv) pero con *peratio*, la razón pensión-ingreso, como la variable dependiente. (Observe que esta es una fracción entre cero y uno, pero aunque suele asumir el valor de cero, nunca se acerca a la unidad. Por tanto, un modelo Tobit está bien como aproximación.) ¿El género o la raza tienen un efecto sobre la relación pensión-ingreso?

C17.4 En el ejemplo 9.1 se agregaron los términos cuadráticos $pcnv^2$, $ptime86^2$ e $inc86^2$ al modelo lineal para *narr86*.

i) Use los datos en CRIME1.RAW para agregar estos mismos términos a la regresión de Poisson en el ejemplo 17.3.

ii) Calcule la estimación de σ^2 dada por $\hat{\sigma}^2 = (n - k - 1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2/\hat{y}_i$. ¿Hay alguna evidencia de sobredispersión? ¿Cómo se deben ajustar los errores estándar de la EMV de Poisson?

iii) Use los resultados de las partes i) y ii) y la tabla 17.3 para calcular el estadístico de razón de cuasi verosimilitudes para la significancia conjunta de los tres términos cuadráticos. ¿Qué concluye usted?

C17.5 Consulte la tabla 13.1 en el capítulo 13. Ahí, se usaron los datos en FERTIL1.RAW para estimar un modelo lineal para *kids*, el número de niños que ha tenido una mujer.

i) Estime un modelo de regresión Poisson para $\widehat{kids}$, empleando las mismas variables de la tabla 13.1. Interprete el coeficiente de *y82*.

ii) ¿Cuál es la diferencia porcentual estimada en la fertilidad entre una mujer negra y una mujer no negra, manteniendo fijos los demás factores?

iii) Obtenga $\hat{\sigma}$. ¿Hay evidencia de sobre o subdispersión?

iv) Compare los valores ajustados de la regresión de Poisson y obtenga la *R*-cuadrada como la correlación cuadrada entre $kids_i$ y $kids_i$. Compare esto con la *R*-cuadrada del modelo de regresión lineal.

C17.6 Use los datos en RECID.RAW para estimar el modelo del ejemplo 17.4 por MCO, usando únicamente las 552 duraciones no censuradas. Comente en términos generales cómo estas estimaciones se comparan con las de la tabla 17.4.

C17.7 Use los datos MROZ.RAW para este ejercicio.

i) Tomando las 428 mujeres que estaban en la fuerza laboral, estime el rendimiento de la educación mediante MCO incluyendo las variables explicativas *exper*, $exper^2$, *nwife-inc*, *age*, *kidslt6* y *kidsge6*. Reporte su estimación para *educ* y su error estándar.

ii) Ahora estime el rendimiento de la educación mediante Heckit, donde todas las variables exógenas aparecen en la regresión de la segunda etapa. En otras palabras, la regresión es log(*wage*) sobre *educ*, *exper*, $exper^2$, *nwifeinc*, *age*, *kidslt6*, *kidsge6* y $\hat{\lambda}$. Compare el rendimiento estimado de la educación y su error estándar con el de la parte i).

iii) Usando sólo las 428 observaciones para las mujeres que trabajan, efectúe la regresión de $\hat{\lambda}$ sobre *educ*, *exper*, $exper^2$, *nwifeinc*, *age*, *kidslt6* y *kidsge6*. ¿Qué tan grande es la *R*-cuadrada? ¿Cómo ayuda esto a explicar sus hallazgos en la parte ii)? (*Sugerencia:* piense en la multicolinealidad.)

C17.8 El archivo JTRAIN2.RAW contiene datos sobre un experimento de capacitación laboral para un grupo de hombres. Los hombres podrían ingresar al programa a partir de enero de 1976 hasta mediados de 1977. El programa terminó en diciembre de 1977. La idea es probar si la participación en el programa de capacitación laboral tuvo un efecto sobre las probabilidades de desempleo y los ingresos en 1978.

i) La variable *train* es el indicador de capacitación laboral. ¿Cuántos hombres en la muestra participaron en el programa de capacitación laboral? ¿Cuál es el mayor número de meses en el que un hombre participó realmente en el programa (*mostrn*)?

ii) Ejecute una regresión lineal de *train* sobre varias variables demográficas y anteriores a la capacitación: *unem74, unem75, age, educ, black, hisp* y *married*. ¿Estas variables en conjunto son significativas al nivel de 5%?

iii) Estime una versión probit del modelo lineal en la parte ii). Calcule la prueba de razón de verosimilitudes para la significancia conjunta de todas las variables. ¿Qué concluye?

iv) Con base en sus respuestas a las partes ii) y iii), ¿parece que la participación en la capacitación laboral puede tratarse como exógena para explicar el estatus de desempleo de 1978? Explique.

v) Ejecute una regresión simple de *unem78* en *train* y reporte los resultados en forma de ecuación. ¿Cuál es el efecto estimado de participar en el programa de capacitación laboral sobre la probabilidad de ser desempleado en 1978? ¿Es estadísticamente significativo?

vi) Ejecute un probit de *unem78* sobre *train*. ¿Es lógico comparar el coeficiente probit de *train* con el coeficiente obtenido del modelo lineal en la parte v)?

vii) Determine las probabilidades ajustadas de las partes v) y vi). Explique por qué son idénticas. ¿Qué método usaría para medir el efecto y la significancia estadística del programa de capacitación laboral?

viii) Agregue todas las variables de la parte ii) como controles adicionales a los modelos de las partes v) y vi). ¿Ahora son idénticas las probabilidades ajustadas? ¿Cuál es la correlación entre ellas?

ix) Mediante el modelo de la parte viii), estime el efecto parcial promedio de *train* sobre la probabilidad de desempleo de 1978. Use (17.17) con $c_k = 0$. ¿Cómo se compara esta estimación con la de MCO de la parte viii)?

C17.9 Use los datos en APPLE.RAW para este ejercicio. Éstos son los datos de una encuesta telefónica que intentaba provocar la demanda de una manzana (ficticia) "ecológicamente amigable". A cada familia se le presentó un conjunto de precios para manzanas regulares (*regprc*) y manzanas ecológicas (*ecoprc*). Se les preguntó cuántas libras de cada tipo de manzanas comprarían (*reglbs* y *ecolbs*, respectivamente).

i) De las 660 familias en la muestra, ¿cuántas reportaron no querer ninguna de las manzanas ecológicas al precio fijado ($ecolbs = 0$)?

ii) ¿La variable *ecolbs* parece tener una distribución continua sobre valores estrictamente positivos? ¿Qué implicaciones tiene su respuesta para la idoneidad de un modelo Tobit para *ecolbs*?

iii) Estime un modelo Tobit para *ecolbs* con *ecoprc, regprc, faminc* y *hhsize* como las variables explicativas. ¿Qué variables son significativas al nivel de 1%?

iv) ¿Son *faminc* y *hhsize* en conjunto significativas?

v) ¿Los signos de los coeficientes de las variables de precio de la parte iii) fueron lo que usted esperaba? Explique.

vi) Sea β_1 el coeficiente de *ecoprc* y β_2 el coeficiente en *regprc*. Pruebe la hipótesis H_0: $-\beta_1 = \beta_2$ frente a la alternativa de dos colas. Reporte el valor-*p* de la prueba. (Quizá quiera consultar la sección 4.4 si su paquete de regresión no calcula fácilmente tales pruebas.)

vii) Obtenga las estimaciones de $E(ecolbs|\mathbf{x})$ para todas las observaciones en la muestra. [Vea la ecuación (17.25).] Llame a éstas $\widehat{ecolbs}_i$. ¿Cuáles son los valores ajustados mayores y menores?

viii) Calcule la correlación cuadrada entre $ecolbs_i$ y $\widehat{ecolbs}_i$.

ix) Ahora, estime un modelo lineal para *ecolbs* mediante las mismas variables explicativas de la parte iii). ¿Por qué las estimaciones de MCO son mucho menores que las estimaciones Tobit? En términos de bondad de ajuste, ¿el modelo Tobit es mejor que el modelo lineal?

x) Evalúe la siguiente afirmación: "Debido a que la *R*-cuadrada del modelo Tobit es tan pequeña, los efectos estimados del precio probablemente sean inconsistentes".

C17.10 Use los datos en SMOKE.RAW para este ejercicio.

i) La variable *cigs* es el número de cigarrillos fumados por día. ¿Cuántas personas en la muestra no fuman en absoluto? ¿Qué fracción de personas afirman fumar 20 cigarrillos al día? ¿Por qué piensa que hay una acumulación de personas en 20 cigarrillos?

ii) Dadas sus respuestas en la parte i), ¿*cigs* parece ser un buen candidato para tener una distribución de Poisson condicional?

iii) Estime un modelo de regresión de Poisson para *cigs*, incluyendo log(*cigpric*), log(*income*), *white, educ, age* y age^2 como variables explicativas. ¿Cuáles son las elasticidades estimadas de precio e ingreso?

iv) Usando los errores estándar de máxima verosimilitud, ¿las variables de precio e ingreso son estadísticamente significativas al nivel de 5%?

v) Obtenga la estimación de σ^2 descrita después de la ecuación (17.35). ¿Cuánto es $\hat{\sigma}$? ¿Cómo se deben ajustar los errores estándar de la parte iv)?

vi) Empleando los errores estándar ajustados de la parte v), ¿las elasticidades de precio e ingreso ahora difieren estadísticamente de cero? Explique.

vii) ¿Las variables de educación y edad son significativas empleando errores estándar más robustos? ¿Cómo interpreta el coeficiente de *educ*?

viii) Obtenga los valores ajustados, $\hat{y}_i$, del modelo de regresión de Poisson. Determine los valores máximos y mínimos, y analice qué tan bien predice el modelo exponencial el consumo intensivo de cigarrillos.

ix) Utilizando los valores ajustados de la parte viii), obtenga el coeficiente de correlación cuadrada entre $\hat{y}_i$ y y_i.

x) Estime un modelo lineal para *cigs* mediante MCO, usando las variables explicativas (y las mismas formas funcionales) de la parte iii). ¿El modelo lineal o el modelo exponencial ofrecen un mejor ajuste? ¿Alguna R-cuadrada es muy grande?

C17.11 Use los datos en CPS91.RAW en este ejercicio. Estos datos son de mujeres casadas, que también tienen información demográfica y sobre el ingreso de cada marido.

i) ¿Qué fracción de mujeres reportó estar en la fuerza laboral ($inlf = 1$)?

ii) Sólo empleando los datos para mujeres que trabajan (no tiene opción), estime la ecuación salarial

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 black + \beta_5 hispanic + u$$

mediante mínimos cuadrados ordinarios. Reporte los resultados en la forma acostumbrada. ¿Parece haber diferencias salariales significativas por raza y etnicidad?

iii) Estime un modelo probit para *inlf* que incluya las variables explicativas en la ecuación salarial de la parte ii), así como *nwifeinc* (ingreso familiar no debido a la esposa) y *kidlt6* (binaria igual a uno si tiene al menos un hijo menor de 6 años de edad). ¿Las últimas dos variables tienen coeficientes con el signo esperado? ¿Son estadísticamente significativas?

iv) Explique por qué, para fines de probar y posiblemente corregir en la ecuación salarial el problema de la selección en la fuerza laboral, es importante que *nwifeinc* y *kidlt6* ayuden a explicar *inlf*. ¿Qué debe suponer acerca de *nwifeinc* y *kidlt6* en la ecuación salarial?

v) Calcule la razón inversa de Mills (para cada observación) y agréguela como un regresor adicional a la ecuación salarial de la parte ii). ¿Cuál es el valor-p de dos colas? ¿Piensa que es particularmente pequeño con 3,286 observaciones?

vi) ¿Agregar la razón inversa de Mills cambia los coeficientes en la regresión salarial de forma importante? Explique.

C17.12 Use los datos en CHARITY.RAW para contestar estas preguntas.

i) La variable *respond* es una variable binaria igual a uno si un individuo responde con una donación a la solicitud más reciente. La base de datos consiste sólo de personas que han contestado al menos una vez en el pasado. ¿Qué fracción de personas respondió más recientemente ($resplast = 1$)?

ii) Estime un modelo probit para *respond*, utilizando *resplast*, *weekslast*, *propresp*, *mailsyear* y *avggift* como las variables explicativas. ¿Cuál de las variables explicativas es estadísticamente significativa?

iii) Encuentre el efecto parcial promedio de *mailsyear* y compárelo con el coeficiente de un modelo de probabilidad lineal.

iv) Empleando las mismas variables explicativas, estime un modelo Tobit para *gift*, la cantidad de la donación más reciente (en florines holandeses). Ahora, ¿qué variable explicativa es estadísticamente significativa?

v) Compare el EPP Tobit para *mailsyear* con el de la regresión lineal. ¿Son similares?

vi) ¿Las estimaciones de las partes ii) y iv) son enteramente compatibles con un modelo Tobit? Explique.

C17.13 Use los datos en HTV.RAW para responder esta pregunta.

i) Mediante MCO en la muestra total, estime un modelo para log(*wage*) mediante las variables explicativas *educ*, *abil*, *exper*, *nc*, *west*, *south* y *urban*. Reporte el rendimiento estimado de la educación y su error estándar.

ii) Ahora, estime la ecuación de la parte i) usando sólo personas con $educ < 16$. ¿Qué porcentaje de la muestra se pierde? Ahora, ¿cuál es el rendimiento estimado de un año de escolaridad? ¿Cómo se compara esto con la parte i)?

iii) Ahora deseche todas las observaciones con $wage \geq 20$, de manera que todos los que queden en la muestra ganan menos de $20 por hora. Ejecute la regresión de la parte i) y comente acerca del coeficiente de *educ*. (Debido a que el modelo de regresión truncada normal supone que *y* es continua, no importa en teoría si se desechan las observaciones con $wage \geq 20$ o $wage > 20$. En la práctica, incluida esta aplicación, esto puede importar un poco debido a que hay algunas personas que ganan exactamente $20 por hora.)

iv) Utilizando la muestra en la parte iii), realice la regresión truncada [tomando log(20) como punto de truncamiento superior]. ¿La regresión trunca parece recuperar el rendimiento de la educación en toda la población, en el supuesto de que la estimación de i) sea consistente? Explique.

Apéndice 17A

Estimación de máxima verosimilitud con variables explicativas

El apéndice C ofrece un repaso de la estimación de máxima verosimilitud (EMV) en el caso más simple de estimar los parámetros en una distribución no condicional. Pero la mayoría de los modelos en econometría tienen variables explicativas, ya sea que se estimen esos modelos mediante MCO o EMV. Esta última es indispensable para los modelos no lineales y aquí se ofrece una breve descripción del enfoque general.

Todos los modelos que se cubren en este capítulo se pueden poner en la siguiente forma. Sea $f(y|\mathbf{x},\boldsymbol{\beta})$ la función de densidad para una extracción aleatoria y_i de la población, condicional en $\mathbf{x}_i = \mathbf{x}$. El estimador de máxima verosimilitud (EMV) de $\boldsymbol{\beta}$ maximiza la función de log-verosimilitud,

$$\max_{\mathbf{b}} \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \mathbf{b}), \tag{17.53}$$

donde el vector $\mathbf{b}$ es el argumento binario en el problema de maximización. En la mayoría de los casos, la EMV, que se escribe como $\hat{\boldsymbol{\beta}}$, es consistente y tiene una distribución normal aproximada en muestras grandes. Esto es verdad aun cuando no se pueda escribir una fórmula para $\hat{\boldsymbol{\beta}}$ salvo en circunstancias muy especiales.

Para el caso de respuesta binaria (logit y probit), la densidad condicional se determina por dos valores, $f(1|\mathbf{x},\boldsymbol{\beta}) = \mathrm{P}(y_i = 1|\mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\beta})$ y $f(0|\mathbf{x},\boldsymbol{\beta}) = \mathrm{P}(y_i = 0|\mathbf{x}_i) = 1 - G(\mathbf{x}_i\boldsymbol{\beta})$. De hecho, una forma sucinta de escribir la densidad es $f(y|\mathbf{x},\boldsymbol{\beta}) = [1 - G(\mathbf{x}\boldsymbol{\beta})]^{(1-y)}[G(\mathbf{x}\boldsymbol{\beta})]^y$ para $y = 0, 1$. Por tanto, (17.53) se puede escribir como

$$\max_{\mathbf{b}} \sum_{i=1}^{n} \{(1 - y_i)\log[1 - G(\mathbf{x}_i\mathbf{b})] + y_i \log[G(\mathbf{x}_i\mathbf{b})]\}. \tag{17.54}$$

En general, las soluciones a (17.54) se encuentran rápidamente mediante computadoras modernas y métodos iterativos para maximizar una función. El tiempo total de cálculo incluso para los conjuntos de datos muy grandes es muy rápido.

La función de log-verosimilitud para el modelo Tobit y para la regresión censurada y truncada son sólo ligeramente más complicados, dependiendo de un parámetro de varianza adicional además de $\boldsymbol{\beta}$. Se derivan fácilmente de las densidades obtenidas en el libro. Vea Wooldridge (2002) para mayores detalles.

Apéndice 17B

Errores estándar asintóticos en modelos de variable dependiente limitada

Las derivaciones de los errores estándar asintóticos para los modelos y métodos presentados en este capítulo están más allá del alcance de este libro. No sólo las derivaciones requieren álgebra matricial, sino que también requieren una teoría asintótica avanzada sobre estimación no lineal. Los antecedentes necesarios para un análisis cuidadoso de estos métodos y varias derivaciones se dan en Wooldridge (2002).

Es instructivo ver las fórmulas para obtener los errores estándar asintóticos para al menos algunos de los métodos. Dado el modelo de respuesta binaria $\mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})$, donde $G(\cdot)$ es la función logit o probit y $\boldsymbol{\beta}$ es el vector $k \times 1$ de parámetros, la matriz de varianza asintótica de $\hat{\boldsymbol{\beta}}$ se estima como

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) \equiv \left(\sum_{i=1}^{n} \frac{[g(\mathbf{x}_i\hat{\boldsymbol{\beta}})]^2\mathbf{x}_i'\mathbf{x}_i}{G(\mathbf{x}_i\hat{\boldsymbol{\beta}})[1 - G(\mathbf{x}_i\hat{\boldsymbol{\beta}})]}\right)^{-1}, \tag{17.55}$$

que es una matriz $k \times k$ (Vea el apéndice D para un resumen de álgebra matricial.) Sin los términos que implican $g(\cdot)$ y $G(\cdot)$, esta fórmula se parece mucho a la matriz de varianza estimada para el estimador de MCO, menos el término $\hat{\sigma}^2$. La expresión en (17.55) representa la naturaleza no lineal de la probabilidad de respuesta, es decir, la naturaleza no lineal de $G(\cdot)$, así como la forma particular de heterocedasticidad en un modelo de respuesta binaria: $\mathrm{Var}(y|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})[1 - G(\mathbf{x}\boldsymbol{\beta})]$.

Las raíces cuadradas de los elementos de la diagonal de (17.55) son los errores estándar asintóticos de las $\hat{\beta}_j$, y suelen reportarse mediante el software de econometría que soporta el análisis logit y probit. Una vez que esto se tiene, los estadísticos t y los intervalos de confianza (asintóticos) se obtienen de la forma acostumbrada.

La matriz en (17.55) también es la base de las pruebas de Wald de restricciones múltiples sobre $\boldsymbol{\beta}$ [vea Wooldridge (2002, capítulo 15)].

La matriz de varianza asintótica para Tobit es más complicada pero tiene una estructura similar. Observe que se puede obtener un error estándar para $\hat{\sigma}$ también. La varianza asintótica para la regresión de Poisson, que permite $\sigma^2 \neq 1$ en (17.35), tiene una forma muy similar a (17.55):

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\sum_{i=1}^{n} \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\mathbf{x}_i\right)^{-1}. \qquad \boxed{\textbf{17.56}}$$

Las raíces cuadradas de los elementos de la diagonal de esta matriz son los errores estándar asintóticos. Si el supuesto de Poisson es válido, se puede descartar $\hat{\sigma}^2$ de la fórmula (debido a que $\sigma^2 = 1$).

Los errores estándar asintóticos para la regresión censurada, regresión trunca y la corrección de la selección muestral Heckit son más complicados, aunque comparten características con las fórmulas anteriores. [Vea Wooldridge (2002) para mayores detalles.]

CAPÍTULO 18

Temas avanzados de series de tiempo

En este capítulo se analizarán algunos temas más avanzados de series de tiempo en la econometría. En los capítulos 10, 11 y 12 se hizo énfasis en distintos lugares que utilizan datos de series de tiempo en el análisis de regresión y que requieren de ciertos cuidados debido a la naturaleza persistente de muchas series de tiempo económicas. Además de estudiar temas como los modelos de rezagos distribuidos infinitos y pronóstico, también se analizan algunos avances recientes en el análisis de procesos de series de tiempo con raíces unitarias.

En la sección 18.1 se describen los modelos de rezagos distribuidos infinitos, los cuales permiten un cambio en una variable explicativa que afectan todos los valores futuros de la variable dependiente. Conceptualmente, estos modelos son extensiones directas de los modelos de rezagos distribuidos finitos del capítulo 10, pero la estimación de estos modelos presenta algunos interesantes desafíos.

En la sección 18.2 se muestra cómo probar formalmente las raíces unitarias en un proceso de series de tiempo. Recuerde del capítulo 11 que se excluyeron los procesos de raíces unitarias para aplicar la teoría asintótica usual. Debido a que la presencia de una raíz unitaria implica que un shock hoy tiene un impacto persistente, determinar si un proceso tiene una raíz unitaria es interesante por sí mismo.

En la sección 18.3 se aborda la noción de regresión espuria entre dos procesos de series de tiempo, cada uno de los cuales tiene una raíz unitaria. El principal resultado es que aunque dos series con raíces unitarias sean *independientes*, es muy probable que la regresión de una sobre la otra produzca un estadístico t estadísticamente significativo. Esto enfatiza las consecuencias potencialmente serias de utilizar la inferencia estándar cuando las variables dependientes e independientes sean procesos integrados.

La noción de cointegración aplica cuando dos series son I(1), pero una combinación lineal de ellas es I(0); en este caso, la regresión de una sobre la otra no es espuria, pero en cambio, expresa algo acerca de la relación de largo plazo entre ellas. La cointegración entre dos series también implica un tipo particular de modelo, llamado modelo de corrección del error, para la dinámica de corto plazo. Se cubrirán estos modelos en la sección 18.4.

En la sección 18.5 se ofrece un panorama general del pronóstico y se conjuntarán todas las herramientas de éste y de capítulos previos para mostrar cómo se pueden usar los métodos de regresión para pronosticar los valores futuros de una serie de tiempo. La literatura sobre los pronósticos es vasta, así que este capítulo sólo se enfocará en los métodos más comunes basados en la regresión. Asimismo, se analizará el tema relacionado con la causalidad de Granger.

18.1 Modelos de rezagos distribuidos infinitos

Sea $\{(y_t, z_t): t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ un proceso bivariado de series de tiempo (que se observa sólo parcialmente). Un **modelo de rezago distribuido infinito (RDI)** que relaciona y_t con todos los valores actuales y pasados de z es

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \ldots + u_t, \qquad \textbf{18.1}$$

donde la suma de las z rezagadas se extiende a un pasado indefinido. Este modelo sólo es una aproximación a la realidad, pues ningún proceso económico comienza en un pasado tan lejano. En comparación con un modelo de rezagos distribuidos finitos, un modelo de RDI no requiere que se trunque el rezago en un valor particular.

Para que el modelo (18.1) tenga sentido, los coeficientes de las variables rezagadas, δ_j, deben tender a cero a medida que $j \to \infty$. Esto no significa que δ_2 sea menor en magnitud que δ_1; esto sólo significa que el impacto de z_{t-j} sobre y_t tiende a ser menor a medida que j aumenta. En la mayoría de las aplicaciones, esto también tiene sentido económico: valores lejanos de z deben ser menos importantes para explicar y que los valores recientes de z.

Aun si se decide que (18.1) es un modelo útil, es obvio que no será posible estimarlo sin algunas restricciones. En primer lugar, sólo se observa una historia finita de datos. La ecuación (18.1) implica un número infinito de parámetros, $\delta_0, \delta_1, \delta_2, \ldots$, que no es posible estimar sin restricciones. Más adelante se imponen restricciones sobre δ_j que permitirán estimar (18.1).

Como en el caso de los modelos de rezagos distribuidos finitos (RDF), el impacto de la propensión en (18.1) es simplemente δ_0 (vea el capítulo 10). Por lo general, las δ_h tienen la misma interpretación como en un RDF. Suponga que $z_s = 0$ para toda $s < 0$ y que $z_0 = 1$ y $z_s = 0$ para toda $s > 1$; en otras palabras, en el tiempo $t = 0$, z se incrementa en una unidad en el primer periodo y después regresa a su valor inicial de cero. Para toda $h \geq 0$, se tiene $y_h = \alpha + \delta_h + u_h$ para toda $h \geq 0$, y por tanto

$$\mathrm{E}(y_h) = \alpha + \delta_h, \qquad \textbf{18.2}$$

donde se emplea el supuesto estándar de que u_h tiene media cero. De lo anterior se desprende que δ_h es el cambio en $\mathrm{E}(y_h)$, dado un cambio temporal de una unidad en z al tiempo cero. Se ha mencionado que δ_h debe tender a cero a medida que h crece, para que el modelo RDI tenga sentido. Esto significa que un cambio temporal en z no tiene *efecto a largo plazo* en el valor esperado de y: $\mathrm{E}(y_h) = \alpha + \delta_h \to \alpha$ conforme $h \to \infty$.

Se hace el supuesto que el proceso z comienza en $z_s = 0$ y que el incremento de una unidad ocurrió en $t = 0$. Esto fue sólo para fines de ilustración. En términos más generales, si z tiene un aumento unitario temporal (a partir de un nivel inicial) en el tiempo t, entonces δ_h mide el cambio en el valor esperado de y después de h periodos. La distribución rezagada, que es δ_h graficada en función de h, muestra la trayectoria esperada que sigue y en el futuro dado un cambio en z una vez transcurrido un periodo.

La propensión de largo plazo en el modelo (18.1) es la suma de todos los coeficientes de los rezagos:

$$PLP = \delta_0 + \delta_1 + \delta_2 + \delta_3 + \ldots, \qquad \textbf{18.3}$$

donde se ha supuesto que la suma infinita está bien definida. Debido a que δ_j debe converger a cero, la PLP suele aproximarse bien por una suma finita de la forma $\delta_0 + \delta_1 + \ldots + \delta_p$ para una p suficientemente grande. Para interpretar la PLP, suponga que el proceso z_t es constante en $z_s = 0$ para $s < 0$. En $t = 0$, el proceso se incrementa permanentemente en una unidad. Por ejemplo, si z_t es el cambio porcentual en la oferta monetaria y y_t es la tasa inflacionaria, entonces, lo que interesa son los efectos de un incremento permanente de un punto porcentual en el crecimiento de la oferta monetaria. Después, al sustituir $z_s = 0$ para $s < 0$ y $z_t = 1$ para $t \geq 0$, se tiene

$$y_h = \alpha + \delta_0 + \delta_1 + \ldots + \delta_h + u_h,$$

donde $h \geq 0$ es cualquier horizonte. Debido a que u_t tiene media cero para toda t, se tiene

$$\mathrm{E}(y_h) = \alpha + \delta_0 + \delta_1 + \ldots + \delta_h. \qquad \boxed{18.4}$$

[Es útil comparar (18.4) y (18.2).] A medida que el horizonte aumenta, es decir, cuando $h \to \infty$, el lado derecho de (18.4) es, por definición, el multiplicador de largo plazo, más α. Por tanto, la PLP mide el cambio a largo plazo en el valor esperado de y dado un incremento unitario *permanente* en z.

Pregunta 18.1

Suponga que $z_s = 0$ para $s < 0$ y que $z_0 = 1$, $z_1 = 1$ y $z_s = 0$ para $s > 1$. Determine $\mathrm{E}(y_{-1})$, $\mathrm{E}(y_0)$ y $\mathrm{E}(y_h)$ para $h \geq 1$. ¿Qué sucede cuando $h \to \infty$?

La derivación previa de la PLP y la interpretación de δ_j utilizan el hecho de que los errores tienen media cero, esto no es un supuesto, siempre y cuando se incluya un intercepto en el modelo. Un análisis minucioso de este razonamiento muestra que la suposición que se hizo del cambio en z durante cualquier periodo, no tuvo efecto en el valor esperado de u_t. Esta es la versión de rezagos distribuidos infinitos del supuesto de *exogeneidad estricta* que se presentó en el capítulo 10 (en particular, el supuesto TS.3). Formalmente,

$$\mathrm{E}(u_t|\ldots, z_{t-2}, z_{t-1}, z_t, z_{t+1}, \ldots) = 0, \qquad \boxed{18.5}$$

así que el valor esperado de u_t no depende de z en *ningún* periodo. A pesar de que (18.5) es natural para algunas aplicaciones, descarta otras importantes posibilidades. De hecho, (18.5) no permite retroalimentación de y_t a z en el futuro, pues z_{t+h} no debe correlacionarse con u_t para $h > 0$. En el ejemplo de la inflación y crecimiento en la oferta monetaria, donde y_t es la inflación y z_t es el crecimiento de la oferta monetaria, (18.5) descarta los cambios futuros en el crecimiento de la oferta monetaria que estén vinculados a los cambios en la tasa inflacionaria de hoy. Dado que la política de oferta monetaria suele mantener dentro de ciertos límites las tasas de inflación y de interés, esto podría ser poco realista.

Un método para estimar la δ_j, que se cubrirá en la siguiente subsección, requiere un supuesto de exogeneidad estricta con el fin de obtener estimadores consistentes de la δ_j. Un supuesto más débil es

$$\mathrm{E}(u_t|z_t, z_{t-1}, \ldots) = 0. \qquad \boxed{18.6}$$

De acuerdo con (18.6), el error no está relacionado con la z, actual ni con las *anteriores*, pero puede estar correlacionado con z futuras; esto permite que z_t sea una variable que sigue reglas de políticas que dependen de las y pasadas. En ocasiones, (18.6) es suficiente para estimar la δ_j; esto se explicará en la siguiente subsección.

Es importante recordar que (18.5) ni (18.6) revelan nada acerca de las propiedades de correlación serial de $\{u_t\}$. (Esto es tal como en los modelos de rezagos distribuidos finitos.) En cualquier caso, es de esperarse que $\{u_t\}$ esté serialmente correlacionada debido a que (18.1) por lo general, no es dinámicamente completo en el sentido que se discutió en la sección 11.4. Se estudiará más adelante el problema de la correlación serial.

¿Cómo se interpretan los coeficientes de los rezagos y la PLP si (18.6) es válida pero (18.5) no? La respuesta es: de la misma forma que antes. Aún se puede hacer el experimento mental anterior, aunque los datos que se observaron los haya generado alguna retroalimentación entre y_t y z futura. Por ejemplo, se puede preguntar acerca del efecto a largo plazo de un incremento permanente en el crecimiento de la oferta monetaria sobre la inflación, aunque los datos acerca del crecimiento de la oferta monetaria no se puedan caracterizar como estrictamente exógenos.

Rezagos distribuidos geométricos (o de Koyck)

Debido a que, por lo general, existe un número infinito de δ_j, no es posible estimarlas de manera consistente sin algunas restricciones. La versión más simple de (18.1), que aún hace que el modelo dependa de un número infinito de rezagos, es el **rezago distribuido geométrico (o de Koyck**). En este modelo, la δ_j depende sólo de dos parámetros:

$$\delta_j = \gamma\rho^j, |\rho| < 1, \quad j = 0, 1, 2, \ldots. \tag{18.7}$$

Los parámetros γ y ρ pueden ser positivos o negativos, pero ρ debe ser menor que la unidad en valor absoluto. Esto asegura que $\delta_j \rightarrow 0$ cuando $j \rightarrow \infty$. De hecho, esta convergencia sucede a una gran velocidad. (Por ejemplo, con $\rho = .5$ y $j = 10$, $\rho^j = 1/1024 < .001$.)

El multiplicador de impacto (PI) en el RDG es simplemente $\delta_0 = \gamma$, de manera que el signo de PI está determinado por el signo de γ. Si $\gamma > 0$ y $\rho > 0$, entonces todos los coeficientes rezagados son positivos. Si $\rho < 0$, los coeficientes de los rezagos alternan en signos (ρ^j es negativa si j es non). El multiplicador de largo plazo es más difícil de obtener, pero se puede utilizar un resultado estándar en la suma de una serie geométrica: para $|\rho| < 1$, $1 + \rho + \rho^2 + \ldots + \rho^j + \ldots = 1/(1 - \rho)$, y por tanto

$$PLP = \gamma/(1 - \rho).$$

La PLP tiene el mismo signo que γ.

Si se inserta (18.7) en (18.1), aún se tiene un modelo que depende de las z del pasado indefinido. Sin embargo, una sencilla resta genera un modelo estimable. Se escribe el modelo RDI en los tiempos t y $t - 1$ como:

$$y_t = \alpha + \gamma z_t + \gamma\rho z_{t-1} + \gamma\rho^2 z_{t-2} + \ldots + u_t \tag{18.8}$$

y

$$y_{t-1} = \alpha + \gamma z_{t-1} + \gamma\rho z_{t-2} + \gamma\rho^2 z_{t-3} + \ldots + u_{t-1}. \tag{18.9}$$

Si se multiplica la segunda ecuación por ρ y se resta de la primera, casi la totalidad de los términos se cancelan:

$$y_t - \rho y_{t-1} = (1 - \rho)\alpha + \gamma z_t + u_t - \rho u_{t-1},$$

que puede escribirse como

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + u_t - \rho u_{t-1}, \tag{18.10}$$

donde $\alpha_0 = (1 - \rho)\alpha$. Esta ecuación se parece a un modelo estándar con una variable dependiente rezagada, donde z_t aparece en forma contemporánea. Debido a que γ es el coeficiente en z_t y ρ es el coeficiente en y_{t-1}, parece que se pueden estimar estos parámetros. [Si, por alguna razón, se tiene interés en α, siempre se puede obtener $\hat{\alpha} = \hat{\alpha}_0/(1 - \hat{\rho})$ después de estimar ρ y α_0.]

La simplicidad de (18.10) es un tanto confusa. El término de error en esta ecuación, $u_t - \rho u_{t-1}$, generalmente está correlacionado con y_{t-1}. A partir de (18.9), resulta evidente que u_{t-1} y y_{t-1} están correlacionadas. Por tanto, si se escribe (18.10) como

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + v_t, \qquad \boxed{18.11}$$

donde $v_t \equiv u_t - \rho u_{t-1}$, entonces por lo general, se tiene una correlación entre v_t y y_{t-1}. Sin más supuestos, el estimador de MCO de (18.11) produce estimaciones inconsistentes de γ y ρ.

Un caso donde v_t *debe* estar correlacionado con y_{t-1} ocurre cuando u_t es independiente de z_t y de *todos* los valores pasados de z y y. Entonces, (18.8) es dinámicamente completa, así que u_t no está correlacionado con y_{t-1}. Con base en (18.9), la covarianza entre v_t y y_{t-1} es $-\rho\text{Var}(u_{t-1}) = -\rho\sigma_u^2$, la cual es cero sólo si $\rho = 0$. Se observa con facilidad que v_t está serialmente correlacionado: debido a que $\{u_t\}$ no está serialmente correlacionado, $\text{E}(v_t v_{t-1}) = \text{E}(u_t u_{t-1}) - \rho\text{E}(u_{t-1}^2) - \rho\text{E}(u_t u_{t-2}) + \rho^2\text{E}(u_{t-1}u_{t-2}) = -\rho\sigma_u^2$. Para $j > 1$, $\text{E}(v_t v_{t-j}) = 0$. Así, $\{v_t\}$ es un proceso de promedio móvil de orden uno (vea la sección 11.1). Ésta, y la ecuación (18.11), dan un ejemplo de un modelo que se deriva del modelo original de interés y que tiene una variable dependiente rezagada *además* de una clase particular de correlación serial.

Si se hace el supuesto de exogeneidad estricta (18.5), entonces z_t no está correlacionada con u_t y u_{t-1}, y por tanto, con v_t. Así, si se puede encontrar una variable instrumental idónea para y_{t-1}, a continuación se puede estimar (18.11) por VI. ¿Cuál es un buen candidato de VI para y_{t-1}? Por suposición, u_t y u_{t-1} no están correlacionados con z_{t-1}, así que v_t no está correlacionada con z_{t-1}. Si $\gamma \neq 0$, z_{t-1} y y_{t-1} están correlacionados, incluso después de parcializar z_t. Por consiguiente, se pueden utilizar los instrumentos (z_t, z_{t-1}) para estimar (18.11). En general, es necesario ajustar los errores estándar para la correlación serial en el $\{v_t\}$, como se analizó en la sección 15.7.

Una alternativa al estimador VI explota el hecho que $\{u_t\}$ puede contener un tipo específico de correlación serial. En particular, además de (18.6), suponga que $\{u_t\}$ sigue el modelo AR(1)

$$u_t = \rho u_{t-1} + e_t \qquad \boxed{18.12}$$

$$\text{E}(e_t|z_t, y_{t-1}, z_{t-1}, \ldots) = 0. \qquad \boxed{18.13}$$

Es importante observar que ρ que aparece en (18.12) es el mismo parámetro multiplicador y_{t-1} en (18.11). Si (18.12) y (18.13) son válidas, se puede escribir la ecuación (18.10) como

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + e_t, \qquad \boxed{18.14}$$

la cual es un modelo dinámicamente completo de (18.13). A partir del capítulo 11 se pueden obtener estimadores consistentes, asintóticamente normales de los parámetros por MCO. Esto es muy conveniente, pues no hay necesidad de trabajar con correlaciones seriales en los errores. Si e_t satisface el supuesto de homocedasticidad $\text{Var}(e_t|z_t, y_{t-1}) = \sigma_e^2$, la inferencia acostumbrada aplica. Una vez que se estima γ y ρ, se puede estimar con facilidad la PLP: $\widehat{PLP} = \hat{\gamma}/(1 - \hat{\rho})$.

La simplicidad de este procedimiento depende del supuesto potencialmente sólido de que $\{u_t\}$ sigue un proceso AR(1) con la *misma* ρ que aparece en (18.7). Esto, por lo general no es peor que suponer que $\{u_t\}$ está serialmente correlacionada. Sin embargo, debido a que la consistencia de los estimadores depende de este supuesto, es buena idea probarlo. Una prueba simple comienza por especificar $\{u_t\}$ como un proceso AR(1) con un parámetro *diferente*, es decir, $u_t = \lambda u_{t-1} + e_t$. McClain y Wooldridge (1995) diseñaron una prueba simple de multiplicador de Lagrange de H_0: $\lambda = \rho$ que puede calcularse después de la estimación MCO de (18.14).

El modelo de rezagos distribuidos geométricos se extiende a variables explicativas múltiples, así que se tiene un RD infinito en cada variable explicativa, pero después se debe poder escribir el coeficiente en $z_{t-j,h}$ como $\gamma_h \rho^j$. En otras palabras, aunque γ_h es diferente de cada variable explicativa, ρ es la misma. Por tanto, se puede escribir

$$y_t = \alpha_0 + \gamma_1 z_{t1} + \ldots + \gamma_k z_{tk} + \rho y_{t-1} + v_t. \tag{18.15}$$

Las mismas cuestiones que surgen en el caso con una z lo hacen también en el caso de varias z. Con base en la extensión natural de (18.12) y (18.13), tan sólo reemplazar z_t con $\mathbf{z}_t = (z_{t1}, \ldots, z_{tk})$ MCO es consistente y asintóticamente normal. O se puede usar un método VI.

Modelos de rezagos distribuidos racionales

Los rezagos distribuidos geométricos implican una distribución de rezagos muy restrictiva. Cuando $\gamma > 0$ y $\rho > 0$, las δ_j son positivas y declinan de manera monótona a cero. Es posible tener modelos de rezagos distribuidos infinitos más generales. El RDG es un caso especial de lo que comúnmente recibe el nombre de **modelos de rezagos distribuidos racionales (RDR)**. Un tratamiento general está más allá del alcance de este libro [Harvey (1990) es una buena referencia], aunque se hará un análisis simple pero útil de este tema.

Tal modelo RDR es más fácil de describir al sumar un rezago de z a la ecuación (18.11):

$$y_t = \alpha_0 + \gamma_0 z_t + \rho y_{t-1} + \gamma_1 z_{t-1} + v_t, \tag{18.16}$$

donde $v_t = u_t - \rho u_{t-1}$, como antes. Mediante la sustitución repetida, se puede mostrar que (18.16) es equivalente al modelo de rezago distribuido infinito

$$\begin{aligned} y_t &= \alpha + \gamma_0(z_t + \rho z_{t-1} + \rho^2 z_{t-2} + \ldots) \\ &\quad + \gamma_1(z_{t-1} + \rho z_{t-2} + \rho^2 z_{t-3} + \ldots) + u_t \\ &= \alpha + \gamma_0 z_t + (\rho\gamma_0 + \gamma_1)z_{t-1} + \rho(\rho\gamma_0 + \gamma_1)z_{t-2} \\ &\quad + \rho^2(\rho\gamma_0 + \gamma_1)z_{t-3} + \ldots + u_t, \end{aligned}$$

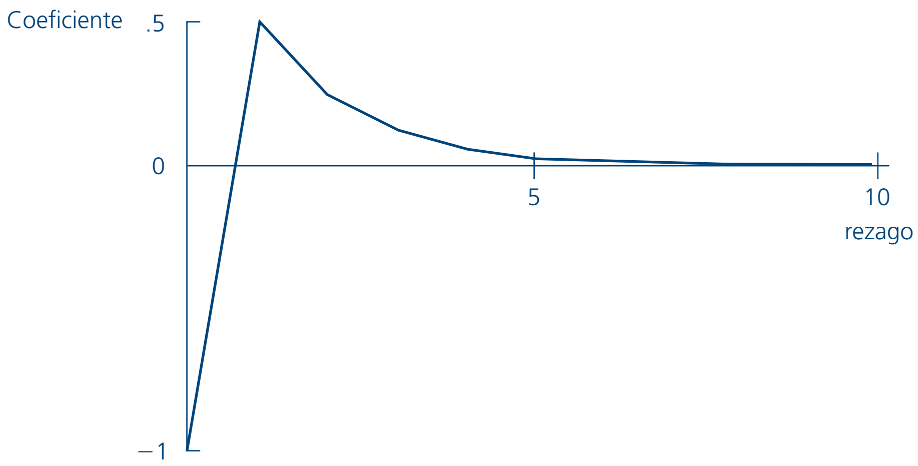
donde nuevamente se necesita el supuesto de $|\rho| < 1$. De esta última ecuación se puede leer la distribución de rezagos. En particular, el multiplicador de impacto es γ_0, mientras que el coeficiente en z_{t-h} es $\rho^{h-1}(\rho\gamma_0 + \gamma_1)$ para $h \geq 1$. Por tanto, este modelo permite que el multiplicador de impacto difiera de signo de los demás coeficientes de los rezagos, aun si $\rho > 0$. No obstante, si $\rho > 0$, the δ_h tiene el mismo signo que $(\rho\gamma_0 + \gamma_1)$ para toda $h \geq 1$. La distribución de rezagos se grafica en la figura 18.1 para $\rho = .5$, $\gamma_0 = -1$ y $\gamma_1 = 1$.

La forma más fácil de calcular la propensión de largo plazo es colocar a y y z en sus valores a largo plazo para toda t, por decir, y^* y z^*, después determinar el cambio y^* respecto a z^* (vea también el problema 10.3). Se tiene que $y^* = \alpha_0 + \gamma_0 z^* + \rho y^* + \gamma_1 z^*$, y resolver dado $y^* = \alpha_0/(1 - \rho) + (\gamma_0 + \gamma_1)/(1 - \rho)z^*$. Ahora, se puede usar el hecho de que la $PLP = \Delta y^*/\Delta z^*$:

$$PLP = (\gamma_0 + \gamma_1)/(1 - \rho).$$

FIGURA 18.1

La distribución de rezagos para el rezago distribuido racional (18.16) con $\rho = .5$, $\gamma_0 = -1$ y $\gamma_1 = 1$.



Ya que $|\rho| < 1$, el PLP tiene el mismo signo que $\gamma_0 + \gamma_1$, y el PLP es cero si, y sólo si, $\gamma_0 + \gamma_1 = 0$, como en la figura 18.1.

Ejemplo 18.1

[Inversión en vivienda e inflación en los precios residenciales]

Al aplicar el MCO a (18.14) y (18.16) respectivamente, se estiman tanto los modelos de rezagos distribuidos básicos geométricos como los racionales. La variable dependiente es log(*invpc*) *después* que se ha eliminado una tendencia de tiempo lineal [es decir, se eliminó la tendencia lineal de log(*invpc*)]. Para z_t, se usa el crecimiento en el índice de precios. Esto permite estimar cómo afecta el precio de la inflación residencial los movimientos en la inversión en vivienda en torno a su tendencia. Los resultados de la estimación, mediante los datos en HSEINV.RAW, se muestran en la tabla 18.1.

Los datos rechazan claramente los modelos de rezagos distribuidos geométricos, puesto que $gprice_{-1}$ es muy significativo. Las R-cuadradas ajustadas también muestran que el modelo RDR se ajusta mejor.

Los dos modelos producen estimaciones muy diferentes del multiplicador de largo plazo. Si se utiliza el RDG de manera incorrecta, la PLP estimada es casi cinco: un incremento de un punto porcentual permanente en la inflación del precio residencial aumenta la inversión a largo plazo en la vivienda por 4.7% (por encima de su valor de tendencia). En términos económicos, esto parece poco factible. La PLP estimada del modelo de rezagos distribuidos está por debajo de uno. De hecho, no es posible rechazar la hipótesis nula H_0: $\gamma_0 + \gamma_1 = 0$ a cualquier nivel razonable de significancia (valor $p = .83$), así que no hay evidencia de que la PLP sea diferente de cero. Este es un buen ejemplo de cómo un error en la especificación de la dinámica de un modelo por omitir los rezagos relevantes puede generar conclusiones erróneas.

TABLA 18.1

Modelos de rezagos distribuidos para inversión en vivienda

Variable dependiente: log(*invpc*), sin tendencia		
Variables independientes	**RD geométrico**	**RD racional**
gprice	3.095 (.933)	3.256 (.970)
y_{-1}	.340 (.132)	.547 (.152)
$gprice_{-1}$	—	−2.936 (.973)
constante	−.010 (.018)	.006 (.017)
Multiplicador de largo plazo	4.689	.706
Tamaño de la muestra *R*-cuadrada ajustada	41 .375	40 .504

18.2 Prueba de raíces unitarias

A continuación se analizará el importante problema de probar si una serie de tiempo sigue un **proceso de raíz unitaria**. En el capítulo 11 se dieron algunos lineamientos generales e informales para decidir si una serie es I(1) o no. En varios casos, es útil realizar una prueba formal para una raíz unitaria. Como se verá, tales pruebas deben aplicarse con precaución.

El método más simple de probar si existe una raíz unitaria comienza con un modelo AR(1):

$$y_t = \alpha + \rho y_{t-1} + e_t, t = 1, 2, \ldots, \tag{18.17}$$

donde y_0 es el valor inicial observado. En esta sección $\{e_t\}$ denotará un proceso que tiene media cero, dadas las y observadas en el pasado:

$$\mathrm{E}(e_t|y_{t-1}, y_{t-2}, \ldots, y_0) = 0. \tag{18.18}$$

[De acuerdo con (18.18), se dice que $\{e_t\}$ es una **secuencia de una martingala en diferencia** respecto a $\{y_{t-1}, y_{t-2}, \ldots\}$. Si se supone que $\{e_t\}$ es i.i.d. con media cero e independiente de y_0, entonces también satisface la ecuación (18.18).]

Si $\{y_t\}$ sigue (18.17), tiene una raíz unitaria si, y sólo si, $\rho = 1$. Si $\alpha = 0$ y $\rho = 1$, $\{y_t\}$ es una caminata aleatoria sin tendencia (estocástica) [con las innovaciones e_t que satisfacen la ecuación (18.18)]. Si $\alpha \neq 0$ y $\rho = 1$, $\{y_t\}$ es una caminata aleatoria con tendencia (estocástica), lo cual significa que $E(y_t)$ es una función lineal de t. Un proceso de raíz unitaria con tendencia (estocástica) se comporta de forma muy diferente de uno sin tendencia (estocástica). Sin embargo, es común dejar α sin especificar bajo la hipótesis nula, y este es el enfoque que aquí se adopta. Por tanto, la hipótesis nula es que $\{y_t\}$ tiene una raíz unitaria:

$$H_0: \rho = 1. \tag{18.19}$$

En casi todos los casos, se tiene interés en la alternativa de una cola

$$H_1: \rho < 1. \tag{18.20}$$

(En la práctica, esto significa que $0 < \rho < 1$, puesto que $\rho < 0$ sería muy poco común para una serie que se sospecha tiene una raíz unitaria.) La alternativa H_1: $\rho > 1$ por lo general no se considera, dado que implica que y_t es explosiva. De hecho, si $\alpha > 0$, y_t tiene una tendencia exponencial en su media cuando $\rho > 1$.

Cuando $|\rho| < 1$, $\{y_t\}$ es un proceso estable AR(1), lo cual significa que es débilmente dependiente o que no se correlaciona asintóticamente. Recuerde del capítulo 11 que $\text{Corr}(y_t, y_{t+h}) = \rho^h \rightarrow 0$ cuando $|\rho| < 1$. Por tanto, probar (18.19) en el modelo (18.17), con la alternativa dada por (18.20), es en realidad una prueba de si $\{y_t\}$ es I(1) contra la alternativa de que $\{y_t\}$ sea I(0). [No se toma la hipótesis nula como I(0) en esta composición debido a que $\{y_t\}$ es I(0) para cualquier valor de ρ estrictamente entre -1 y 1, algo que las pruebas de hipótesis clásicas no manejan fácilmente. Existen pruebas donde la hipótesis nula es I(0) contra la alternativa de I(1), pero adoptan un método diferente. Vea, por ejemplo, Kwiatkowski, Phillips, Schmidt y Shin (1992).]

Una ecuación conveniente para realizar la prueba de raíz unitaria es restar y_{t-1} de ambos lados de (18.17) y definir $\theta = \rho - 1$:

$$\Delta y_t = \alpha + \theta y_{t-1} + e_t. \tag{18.21}$$

Con base en (18.18), este es un modelo dinámicamente completo y también parece sencillo para probar H_0: $\theta = 0$ contra H_1: $\theta < 0$. El problema es que bajo H_0, y_{t-1} es I(1) y, por tanto, el teorema del límite central en el que se basa la distribución normal estándar asintótica para el estadístico t no aplica: el estadístico t no tiene una distribución normal estándar aproximada inclusive en grandes tamaños de muestras. La distribución asintótica del estadístico t bajo H_0 se conoce como **distribución de Dickey-Fuller** en honor de Dickey y Fuller (1979).

Aunque no se pueden usar los valores críticos usuales, es *posible* utilizar el estadístico t para $\hat{\theta}$ en (18.21), al menos una vez que los valores críticos apropiados se han tabulado. La prueba resultante es conocida como **prueba Dickey-Fuller (DF)** para una raíz unitaria. La teoría empleada para determinar los valores críticos asintóticos es muy complicada y se cubre en libros avanzados de econometría de series de tiempo. [Vea, por ejemplo, Banerjee, Dolado, Galbraith y Hendry (1993), o BDGH para abreviar.] Por el contrario, utilizar estos resultados es muy fácil. Los valores críticos para el estadístico t han sido tabulados por varios autores, comenzando por el trabajo original de Dickey y Fuller (1979). La tabla 18.2 contiene los valores críticos de una muestra grande para varios niveles de significancia, tomados de BDGH (1993, tabla 4.2). (Los valores críticos ajustados para tamaños pequeños de muestra se cubren en BDGH.)

TABLA 18.2

Valores críticos asintóticos para prueba *t* de raíz unitaria: sin tendencia de tiempo

Nivel de significancia	1%	2.5%	5%	10%
Valor crítico	−3.43	−3.12	−2.86	−2.57

Se rechaza la hipótesis nula H_0: $\theta = 0$ contra H_1: $\theta < 0$ si $t_{\hat{\theta}} < c$, donde c es uno de los valores negativos en la tabla 18.2. Por ejemplo, para realizar la prueba a un nivel de significancia de 5%, se rechaza si $t_{\hat{\theta}} < -2.86$. Esto requiere un estadístico t con una magnitud mucho mayor que si se usara el valor crítico normal estándar, el cual sería -1.65. Si se usa el valor crítico normal estándar para probar una raíz unitaria, se rechazaría H_0 mucho más de 5% de las veces cuando H_0 es verdadero.

Ejemplo 18.2

[Prueba de raíz unitaria para tasas de interés sobre certificados del Tesoro a tres meses]

Los datos trimestrales en INTQRT.RAW se utilizan para probar la presencia de una raíz unitaria sobre las tasas de interés de los certificados del Tesoro a tres meses. Cuando se estima (18.20), se obtiene

$$\begin{aligned} \widehat{\Delta r3_t} &= .625 - .091\ r3_{t-1} \\ &\quad\ (.261)\quad (.037) \\ n &= 123,\ R^2 = .048, \end{aligned} \qquad \boxed{18.22}$$

donde se mantiene la convención de reportar los errores estándar entre paréntesis debajo de las estimaciones. Recuerde que estos errores estándar no se pueden utilizar para construir intervalos de confianza tradicionales o para realizar las pruebas t acostumbradas debido a que éstos no se comportan de la forma en que suelen hacerlo cuando existe una raíz unitaria. El coeficiente en $r3_{t-1}$ muestra que el estimador de ρ es $\hat{\rho} = 1 + \hat{\theta} = .909$. Si bien este es menor que la unidad, no se sabe si *estadísticamente* es menor que uno. El estadístico t sobre $r3_{t-1}$ es $-.091/.037 = -2.46$. De la tabla 18.2, 10% del valor crítico es -2.57; por tanto, no se puede rechazar H_0: $\rho = 1$ contra H_1: $\rho < 1$ a un nivel de significancia de 10%.

Como con otras pruebas de hipótesis, cuando no se puede rechazar H_0, *no* se dice que se acepta H_0. ¿Por qué? Suponga que se prueba H_0: $\rho = .9$ en el ejemplo anterior mediante una prueba t estándar, que es asintóticamente válida, debido a que y_t es I(0) bajo H_0. A continuación, se obtiene $t = .001/.037$, que es muy pequeño y no ofrece evidencia contra $\rho = .9$. Sin embargo, no tiene sentido aceptar $\rho = 1$ y $\rho = .9$.

Cuando no se puede rechazar una raíz unitaria, como en el ejemplo anterior, sólo se debe concluir que los datos no ofrecen evidencia sólida contra H_0. En este ejemplo, la prueba ofrece *cierta* evidencia contra H_0 debido a que el estadístico t es cercano al valor crítico de 10%. (Idealmente, se calcularía un valor-p, pero esto requiere software especial debido a que la distribución no es normal). Además, aunque $\hat{\rho} \approx .91$ implica un alto grado de persistencia en $\{r3_t\}$, la correlación entre las observaciones que a 10 periodos de distancia de un modelo AR(1) con $\rho = .9$ es cerca de .35, más que casi uno si $\rho = 1$.

¿Qué sucede ahora si se quiere utilizar $r3_t$ como variable explicativa en un análisis de regresión? El resultado de la prueba de raíz unitaria implica que se debe tener mucho cuidado: si $r3_t$ tiene una raíz unitaria, las aproximaciones asintóticas tradicionales no necesitan ser válidas (como se analizó en el capítulo 11). Una solución es usar la primera diferencia de $r3_t$ en cualquier análisis. Como se verá en la sección 18.4, esa no es la única posibilidad.

Es necesario también probar para raíces unitarias en modelos con dinámicas más complicadas. Si $\{y_t\}$ sigue (18.17) con $\rho = 1$, entonces Δy_t no está serialmente correlacionada. Es fácil permitir a $\{\Delta y_t\}$ que siga un modelo AR al aumentar a la ecuación (18.21) con rezagos adicionales. Por ejemplo,

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t, \tag{18.23}$$

donde $|\gamma_1| < 1$. Esto asegura que, bajo H_0: $\theta = 0$, $\{\Delta y_t\}$ sigue un modelo AR(1) estable. En la alternativa H_1: $\theta < 0$, se puede mostrar que $\{y_t\}$ sigue un modelo AR(2) estable.

En términos más generales, es posible agregar p rezagos de Δy_t a la ecuación para dar cuenta de la dinámica del proceso. La forma de probar la hipótesis nula de una raíz unitaria es muy similar: realizar la regresión de

$$\Delta y_t \text{ sobre } y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-p} \tag{18.24}$$

y se realiza la prueba t en $\hat{\theta}$, el coeficiente en y_{t-1}, tal como antes. Esta versión ampliada de la prueba de Dickey-Fuller suele conocerse como la **prueba de Dickey-Fuller aumentada** debido a que la regresión se aumentó con los cambios rezagados, Δy_{t-h}. Los valores críticos y la regla de rechazo son los mismos que antes. La inclusión de los cambios rezagados en (18.24) tiene el fin de eliminar cualquier correlación serial en Δy_t. Cuantos más rezagos se incluyan en (18.24), más observaciones iniciales se pierden. Si se incluyen demasiados rezagos, el poco poder de la prueba, por lo general, se ve afectado. Pero si se incluyen muy pocos rezagos, el tamaño de la prueba será incorrecto, incluso asintóticamente, debido a que la validez de los valores críticos en la tabla 18.2 depende de la dinámica que aparezca completamente en el modelo. A menudo, la longitud de los rezagos está determinada por la frecuencia de los datos (así como por el tamaño de la muestra). Para datos anuales, con uno o dos rezagos suele bastar. Para datos mensuales, se deben incluir 12 rezagos. Pero no se tienen que seguir reglas estrictas en un caso particular.

Es interesante observar que los estadísticos t en los cambios rezagados tienen distribuciones t aproximadas. El estadístico F para significancia conjunta de cualquier grupo de términos Δy_{t-h} también es válido asintóticamente. (Este sostiene el supuesto de homocedasticidad que se analizó en la sección 11.5.) Por tanto, se pueden utilizar pruebas estándar para determinar si se tienen suficientes rezagos de las diferencias en (18.24).

Ejemplo 18.3

[Prueba de raíz unitaria de la inflación estadounidense anual]

Se utilizan datos anuales sobre la inflación estadounidense, con base en el índice de precios al consumidor, para probar una raíz unitaria en la inflación (vea PHILLIPS.RAW), lo cual abarca de 1948 a 1996. Si se permite un rezago de Δinf_t en la regresión Dickey-Fuller aumentada se tiene

$$\begin{aligned} \widehat{\Delta inf_t} &= \underset{(.517)}{1.36} - \underset{(.103)}{.310}\, inf_{t-1} + \underset{(.126)}{.138}\, \Delta inf_{t-1} \\ n &= 47,\ R^2 = .172. \end{aligned}$$

El estadístico t para la prueba de la raíz unitaria es $-.310/.103 = -3.01$. Debido a que el valor crítico de 5% es -2.86, se rechaza la hipótesis de la raíz unitaria al nivel de 5%. La estimación de ρ es de aproximadamente .690. En conjunto, forman una evidencia razonablemente sólida contra una raíz unitaria en la inflación. El rezago Δinf_{t-1} tiene un estadístico t de aproximadamente 1.10, así que no es necesario incluirlo, pero no se puede saber esto antes de tiempo. Si se elimina Δinf_{t-1}, la evidencia contra una raíz unitaria es ligeramente mas robusta: $\hat{\theta} = -.335$ ($\hat{\rho} = .665$) y $t_{\hat{\theta}} = -3.13$.

Para series que tienen tendencias temporales claras, se necesita modificar la prueba para raíces unitarias. Un proceso de tendencia estacionaria, que tiene una tendencia lineal en su media pero es I(0) alrededor de su tendencia, puede confundirse con un proceso de raíz unitaria si no se controla la tendencia temporal en la regresión de Dickey-Fuller. En otras palabras, si se realiza la prueba DF tradicional o la aumentada sobre una serie con tendencia pero que es I(0), probablemente se pueda hacer poco para rechazar una raíz unitaria.

Para permitir una serie con tendencias temporales, se cambia la ecuación básica a

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + e_t, \tag{18.25}$$

donde nuevamente la hipótesis nula es H_0: $\theta = 0$ y la alternativa es H_1: $\theta < 0$. Bajo la alternativa, $\{y_t\}$ es un proceso de tendencia estacionaria. Si y_t tiene una raíz unitaria, entonces $\Delta y_t = \alpha + \delta t + e_t$, y también el *cambio* en y_t tiene una media lineal en t a menos que $\delta = 0$. [Se puede mostrar que $E(y_t)$ en realidad es *cuadrática* en t.] Es poco frecuente que la primera diferencia en una serie económica tenga una tendencia lineal, así que una hipótesis nula más apropiada es probablemente H_0: $\theta = 0$, $\delta = 0$. Aunque es posible probar esta hipótesis conjunta mediante una prueba F, pero con valores críticos modificados, es común sólo probar H_0: $\theta = 0$ mediante una prueba t. Aquí se sigue este método. [Vea BDGH (1993, sección 4.4) para más detalles sobre la prueba conjunta.]

Cuando se incluye una tendencia temporal en la regresión, los valores críticos de la prueba cambian. Intuitivamente, esto ocurre porque la eliminación de la tendencia de un proceso de raíz unitaria tiende a asemejarlo a un proceso I(0). Por tanto, se requiere una magnitud mayor para el estadístico t con el fin de rechazar H_0. Los valores críticos de Dickey-Fuller para la prueba t que incluye una tendencia temporal se presentan en la tabla 18.3; se tomaron de BDGH (1993, tabla 4.2).

TABLA 18.3

Valores críticos asintóticos para la prueba *t* de raíz unitaria: tendencia de tiempo lineal

Nivel de significancia	1%	2.5%	5%	10%
Valor crítico	−3.96	−3.66	−3.41	−3.12

Por ejemplo, para rechazar una raíz unitaria al nivel de 5%, es necesario que el estadístico t sobre $\hat{\theta}$ sea menor que -3.41, como se compara con -2.86 sin una tendencia temporal.

Se puede aumentar la ecuación (18.25) con rezagos de Δy_t para dar cuenta de la correlación serial, tal como en el caso sin una tendencia.

Ejemplo 18.4

[Raíz unitaria en el logaritmo del producto interno bruto real estadounidense]

Se puede aplicar la prueba de la raíz unitaria con una tendencia temporal a los datos del producto interno bruto (*GDP*) estadounidense en INVEN.RAW. Estos datos anuales cubren los años de 1959 a 1995. Se prueba si $\log(GDP_t)$ tiene una raíz unitaria. Esta serie tiene una tendencia pronunciada que se ve como aproximadamente lineal. Se incluye un solo rezago de $\Delta\log(GDP_t)$, que es simplemente el crecimiento en el *GDP* (en forma decimal) para dar cuenta de la dinámica:

$$
\begin{aligned}
\widehat{gGDP}_t &= \underset{(.67)}{1.65} + \underset{(.0027)}{.0059}\, t - \underset{(.087)}{.210} \log(GDP_{t-1}) + \underset{(.165)}{.264}\, gGDP_{t-1} \\
n &= 35,\ R^2 = .268.
\end{aligned}
\tag{18.26}
$$

A partir de esta ecuación se obtiene $\hat{\rho} = 1 - .21 = .79$, lo cual es claramente menor que uno. Pero *no* se puede rechazar una raíz unitaria en el logaritmo de GDP: el estadístico t en $\log(GDP_{t-1})$ es $-.210/.087 = -2.41$, lo cual está muy por arriba del valor crítico de 10% de -3.12. El estadístico t en $gGDP_{t-1}$ es 1.60, que es casi significativo al nivel de 10% contra una alternativa de dos colas.

¿Qué se debe concluir acerca de una raíz unitaria? De nuevo, no se puede rechazar una raíz unitaria, pero el estimador puntual de ρ no está especialmente cerca de uno. Cuando se tiene un tamaño pequeño de muestra $n = 35$ se considera como muy pequeño, es muy difícil rechazar la hipótesis nula de una raíz unitaria si el proceso tiene algo cercano a una raíz unitaria. Al usar más datos sobre periodos prologados de tiempo, muchos investigadores han concluido que hay poca evidencia contra la hipótesis de raíz unitaria para $\log(GDP)$. Esto ha llevado a la mayoría de ellos a suponer que el *crecimiento* en GDP es I(0), lo cual significa que $\log(GDP)$ es I(1). Por desgracia, dados los tamaños de muestra actualmente disponibles, no se puede confiar mucho en esta conclusión.

Si se omite la tendencia temporal, existe mucha menos evidencia contra H_0, pues $\hat{\theta} = -.023$ y $t_{\hat{\theta}} = -1.92$. Aquí, la estimación de ρ es mucho más cercana a uno, pero esto es engañoso debido a la tendencia temporal omitida.

Es tentador comparar el estadístico t en la tendencia temporal en (18.26), con el valor crítico de una distribución t o una normal estándar, para ver si la tendencia temporal es significativa. Por desgracia, el estadístico t en la tendencia no tiene una distribución normal estándar asintótica (a menos que $|\rho| < 1$). La distribución asintótica de este estadístico t es conocida, pero rara vez se utiliza. En general, depende de la intuición (o gráficas de las series de tiempo) para decidir si se incluye o no una tendencia en la prueba DF.

Existen muchas otras variantes en las pruebas de raíz unitaria. En una versión aplicable sólo a las series que claramente no tienen tendencia, el intercepto se omite de la regresión; es decir, α está en cero en (18.21). Esta variante de la prueba Dickey-Fuller rara vez se utiliza debido a los sesgos inducidos si $\alpha \neq 0$. También, se pueden permitir tendencias temporales más complicadas, como la cuadrática. Nuevamente, esto se usa con poca frecuencia.

Otra clase de pruebas intenta tomar en cuenta la correlación serial en Δy_t de una manera diferente al incluir rezagos en (18.21) o (18.25). El enfoque está relacionado con errores estándar seriales robustos a la correlación para los estimadores MCO que se analizaron en la sección 12.5. La idea es ser tan escéptico como sea posible en cuanto a la correlación serial de Δy_t. En la práctica, la prueba de Dickey-Fuller (aumentada) se ha sostenido muy bien. [Vea BDGH (1993, sección 4.3) para un análisis sobre otras pruebas.]

18.3 Regresión espuria

En un entorno de corte transversal, se utiliza la frase "correlación espuria" para describir una situación en la que dos variables están relacionadas a través de su correlación con una tercera variable. En particular si se hace la regresión de y sobre x, se encontrará una relación significativa. Pero cuando se controla otra variable, por ejemplo, z, el efecto parcial de x sobre y se convierte en cero. Naturalmente, esto también puede suceder en contextos de series de tiempo con variables I(0).

Como se analizó en la sección 10.5, es posible encontrar una relación espuria entre series de tiempo que tienen tendencias crecientes o decrecientes. Siempre que las series dependan débilmente de sus tendencias temporales, el problema se resuelve de forma efectiva al incluir una tendencia temporal en el modelo de regresión.

Cuando se trata con procesos que son integradas de orden uno, existe una complicación adicional. Inclusive si las dos series tienen medias sin tendencia, una simple regresión que implique a dos series *independientes* I(1) con frecuencia resultará en un estadístico t significativo.

Con más exactitud, sean $\{x_t\}$ y $\{y_t\}$ caminatas aleatorias generadas por

$$x_t = x_{t-1} + a_t, \quad t = 1, 2, \ldots, \tag{18.27}$$

y

$$y_t = y_{t-1} + e_t, \quad t = 1, 2, \ldots, \tag{18.28}$$

donde $\{a_t\}$ y $\{e_t\}$ son innovaciones independientes e idénticamente distribuidas, con media cero y varianzas σ_a^2 y σ_e^2, respectivamente. Para fines de concreción, se toman los valores iniciales de $x_0 = y_0 = 0$. Suponga además que $\{a_t\}$ y $\{e_t\}$ son procesos independientes. Esto implica que $\{x_t\}$ y $\{y_t\}$ también son independientes. Pero, ¿qué sucede si se ejecuta la regresión simple

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t \tag{18.29}$$

y se obtiene el estadístico usual t para $\hat{\beta}_1$ y la R-cuadrada habitual? Debido a que y_t y x_t son independientes, se esperaría que plim $\hat{\beta}_1 = 0$. Aun más importante, si se prueba H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$ al nivel de 5%, se espera que el estadístico t para $\hat{\beta}_1$ sea no significativo 95% de las veces. Por medio de una simulación, Granger y Newbold (1974) demostraron que este *no* es el caso: aunque y_t y x_t sean *independientes* la regresión de y_t sobre x_t proporciona un estadístico t estadísticamente significativo un gran porcentaje de las veces, mucho mayor que el nivel de significancia nominal. Granger y Newbold llamaron a esto el **problema de regresión espuria**: no hay forma en que y y x se relacionen, pero una regresión MCO que utilice el estadístico t habitual con frecuencia indicará una relación.

Pregunta 18.2

Con base en el escenario anterior, donde $\{x_t\}$ y $\{y_t\}$ se generan por (18.27) y (18.28) y $\{e_t\}$ y $\{a_t\}$ son secuencias i.i.d., ¿cuál es el plim del coeficiente de la variable, por decir, $\hat{\gamma}_1$, de la regresión de Δy_t sobre Δx_t? Describa el comportamiento del estadístico t de $\hat{\gamma}_1$.

Davidson y MacKinnon (1993, tabla 19.1) dan resultados recientes de simulaciones donde a_t y e_t se generan como variables aleatorias, normales, independientes e idénticamente distribuidas, generadas de 10,000 muestras diferentes. Para un tamaño de muestra de $n = 50$ al nivel de significancia de 5%, el estadístico t estándar para H_0: $\beta_1 = 0$ contra la alternativa de dos colas, rechaza H_0 aproximadamente 66.2% de las veces bajo H_0, más que 5% del tiempo. A medida que el tamaño de la muestra aumenta, las cosas *empeoran*: con $n = 250$, ¡la hipótesis nula se rechaza 84.7% de las veces!

He aquí una forma de ver qué sucede cuando se hace la regresión del nivel de y sobre el de x. Se escribe el modelo básico de (18.19) como

$$y_t = \beta_0 + \beta_1 x_t + u_t. \tag{18.30}$$

Para el estadístico t de $\hat{\beta}_1$ se tiene una distribución normal estándar aproximada en muestras grandes, como mínimo, $\{u_t\}$ debe ser un proceso no correlacionado serialmente con una media de cero. Pero bajo H_0: $\beta_1 = 0$, $y_t = \beta_0 + u_t$, y debido a que $\{y_t\}$ es una caminata aleatoria que comienza en $y_0 = 0$, la ecuación (18.30) es válida bajo H_0 sólo si $\beta_0 = 0$ y, más importante, si $u_t = y_t = \sum_{j=1}^{t} e_j$. En otras palabras, $\{u_t\}$ es una caminata aleatoria bajo H_0. Esto viola claramente inclusive la versión asintótica de los supuestos de Gauss-Markov del capítulo 11.

Incluir una tendencia temporal no altera en realidad la conclusión. Si y_t o x_t es una caminata aleatoria con tendencia (estocástica) y no se incluye una tendencia temporal, el problema de la regresión espuria empeora. Las mismas conclusiones cualitativas son válidas si $\{a_t\}$ y $\{e_t\}$ son procesos generales I(0), más que secuencias i.i.d.

Además del estadístico t usual que no tiene una distribución límite normal estándar, de hecho, tiende al infinito conforme $n \rightarrow \infty$ el comportamiento de R-cuadrada es no estándar. En contextos de corte transversal o en regresiones con variables de series de tiempo I(0), R-cuadrada converge en probabilidad con la población R-cuadrada: $1 - \sigma_u^2/\sigma_y^2$. Este no es el caso de las regresiones espurias con procesos I(1). En lugar de que R-cuadrada tenga un plim bien definido, en realidad converge a una variable aleatoria. Formalizar esta noción está fuera del alcance de este libro. [Un análisis de las propiedades asintóticas del estadístico t y de la R-cuadrada se puede encontrar en BDGH (sección 3.1).] La implicación es que, con una probabilidad muy alta, la R-cuadrada sea grande, aunque $\{y_t\}$ y $\{x_t\}$ sean procesos de series de tiempo independientes.

Las mismas consideraciones surgen con variables independientes múltiples, cada una de las cuales podrían ser I(1) o algunas podrían ser I(0). Si $\{y_t\}$ es I(1) y al menos algunas de las variables explicativas son I(1), la regresión resultante podría ser espuria.

La posibilidad de regresión espuria con variables I(1) es muy importante y ha llevado a los economistas a reexaminar muchas regresiones de series de tiempo acumuladas cuyos estadísticos t eran muy significativos y cuyas R-cuadradas eran extremadamente altas. En la siguiente sección se muestra que al efectuar una regresión de una variable dependiente I(1) sobre una variable independiente I(1) *puede* ser informativa, pero sólo si estas variables están relacionadas en un sentido preciso.

18.4 Modelos de cointegración y de corrección del error

El anterior análisis de la regresión espuria alerta sobre el uso de los niveles de las variables I(1) en el análisis de regresión. En los capítulos anteriores se sugirió que las variables I(1) se diferenciaran antes de utilizarse en los modelos de regresión lineal, ya sea que se estimen por MCO o por variables instrumentales. Esto es, sin lugar a dudas, un camino seguro, y es el método que se utiliza en muchas regresiones de series de tiempo después del trabajo original de Granger y Newbold sobre el problema de la regresión espuria. Por desgracia, diferenciar siempre las variables I(1) limita el alcance de las preguntas que se pueden responder.

Cointegración

La noción de **cointegración**, a la que Engle y Granger (1987) dieron un tratamiento formal, hace potencialmente significativas las regresiones que implican variables I(1). Un tratamiento completo de la cointegración exige un desarrollo matemático mas profundo, pero pueden describirse los temas y métodos básicos que se utilizan en muchas aplicaciones.

Si $\{y_t: t = 0, 1, \ldots\}$ y $\{x_t: t = 0, 1, \ldots\}$ son dos procesos I(1), entonces, en general, $y_t - \beta x_t$ es un proceso I(1) para cualquier número β. Sin embargo, es *posible* que para algunas $\beta \neq 0$, $y_t - \beta x_t$ sea un proceso I(0), lo cual significa que tiene media constante, varianza constante y las autocorrelaciones que dependen sólo del periodo transcurrido entre dos variables cualesquiera en la serie y no está correlacionada asintóticamente. Si tal β existe, se dice que y y x están *cointegradas* y β recibe el nombre de parámetro de cointegración. [De manera alterna, se podría ver a $x_t - \gamma y_t$ para $\gamma \neq 0$: si $y_t - \beta x_t$ es I(0), entonces $x_t - (1/\beta)y_t$ es I(0). Por tanto, la combinación lineal de y_t y x_t no es única, pero si se fija el coeficiente de y_t en una unidad, entonces β es única. Vea el problema 18.3. En concreto, se consideran combinaciones lineales de la forma $y_t - \beta x_t$.]

Pregunta 18.3

Sea $\{(y_t, x_t): t = 1, 2, \ldots\}$ una serie de tiempo bivariada donde cada serie es I(1) sin tendencia (estocástica). Explique por qué, si y_t y x_t están cointegradas, y_t y x_{t-1} también están cointegradas.

Con fines ilustrativos, sea $\beta = 1$, suponga que $y_0 = x_0 = 0$, y escriba $y_t = y_{t-1} + r_t$, $x_t = x_{t-1} + v_t$, donde $\{r_t\}$ y $\{v_t\}$ son dos procesos I(0) con medias cero. Entonces, y_t y x_t tienen una tendencia a oscilar y no regresar al valor inicial de cero con ninguna regularidad. En cambio, si $y_t - x_t$ es I(0), tiene media cero y regresa a cero con cierta regularidad.

Como ejemplo específico, sea $r6_t$ la tasa de interés anualizada para los certificados del Tesoro a seis meses (al final del trimestre t) y sea $r3_t$ la tasa de interés anualizada para los certificados del Tesoro a tres meses. (Éstos, por lo general, se llaman rendimientos de obligaciones y aparecen publicados en las páginas financieras.) En el ejemplo 18.2, mediante los datos en INTQRT.RAW, se encuentra poca evidencia contra la hipótesis de que $r3_t$ tiene una raíz unitaria; lo mismo es cierto para $r6_t$. Se define el diferencial entre las tasas de certificados del Tesoro a seis y tres meses como $spr_t = r6_t - r3_t$. Entonces, mediante la ecuación (18.21), el estadístico t de Dickey- Fuller para spr_t es -7.71 (con $\hat{\theta} = -.67$ o $\hat{\rho} = .33$). Así, se rechaza fuertemente una raíz unitaria para spr_t en favor de I(0). El resultado de esto es que, aun cuando $r6_t$ y $r3_t$ parecen ser procesos de raíces unitarias, la diferencia entre ellos es un proceso I(0). En otras palabras, $r6$ y $r3$ están cointegradas.

La cointegración en este ejemplo, como en muchos otros, tiene una interpretación económica. Si $r6$ y $r3$ no estuvieran cointegradas, la diferencia entre las tasas de interés podría volverse muy grande, sin tendencia a moverse conjuntamente. Con base en un simple argumento de arbitraje, esto parecería improbable. Suponga que el diferencial spr_t continúa aumentando por varios periodos de tiempo, lo que hace de los certificados del Tesoro a seis meses una inversión mucho más deseable. Entonces, los inversionistas se alejarían de dichos certificados a tres meses y se acercarían a los de seis meses, lo que aumentaría el precio de estos últimos y disminuiría el precio de los certificados del Tesoro a tres meses. Debido a que las tasas de interés están inversamente relacionadas con el precio, esto disminuiría $r6$ e incrementaría $r3$, hasta que el diferencial se redujera. Por tanto, no se espera que las grandes desviaciones entre $r6$ y $r3$ continúen: el diferencial tiene la tendencia a volver a su valor medio. (El diferencial, en realidad, tiene una media ligeramente positiva debido a que los inversionistas a largo plazo reciben mejores ganancias que los inversionistas a corto plazo.)

Existe otra forma de caracterizar el hecho de que spr_t no se desviará por largos periodos respecto a su valor promedio: $r6$ y $r3$ tienen una relación a *largo plazo*. Para explicar lo que esto quiere decir, sea $\mu = \mathrm{E}(spr_t)$ el valor esperado del diferencial. Entonces, es posible escribir

$$r6_t = r3_t + \mu + e_t,$$

donde $\{e_t\}$ tiene un proceso I(0) con media cero. El equilibrio o la relación a largo plazo ocurre cuando $e_t = 0$, o $r6^* = r3^* + \mu$. En cualquier periodo, pueden existir desviaciones respecto al equilibrio, pero serán temporales: hay fuerzas económicas que conducen a $r6$ y $r3$ nuevamente a la relación de equilibrio.

En el ejemplo de la tasa de interés, se utilizó el razonamiento económico para indicar el valor de β si y_t y x_t están cointegradas. Si se tiene un valor hipotético de β, entonces *probar* si dos series están cointegradas es fácil: simplemente se define una nueva variable, $s_t = y_t - \beta x_t$, y se aplica la prueba DF habitual o la aumentada a $\{s_t\}$. Si se *rechaza* una raíz unitaria en $\{s_t\}$ a favor de la alternativa I(0), entonces se encuentra que y_t y x_t *están cointegradas*. En otras palabras, la hipótesis nula es que y_t y x_t *no están* cointegradas.

Probar la cointegración es más difícil cuando el parámetro de cointegración (potencial) β es desconocido. Más que probar una raíz unitaria en $\{s_t\}$, primero debe estimarse β. *Si* y_t y x_t están cointegradas, resulta que el estimador MCO $\hat{\beta}$ de la regresión

$$y_t = \hat{\alpha} + \hat{\beta}x_t \qquad \boxed{\mathbf{18.31}}$$

es consistente para β. El problema es que la hipótesis nula establece que las dos series *no* están cointegradas, lo cual significa que, bajo H_0, se está realizando una regresión espuria. Por fortuna, es posible tabular los valores críticos aun cuando β es estimada, donde se aplica la prueba de Dickey-Fuller o la prueba aumentada de Dickey-Fuller a los residuales, por decir, $\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta}x_t$, de (18.31). La única diferencia es que los valores críticos dan cuenta de la estimación de β. La prueba resultante se conoce como **prueba de Engle-Granger**, y los valores asintóticos críticos se presentan en la tabla 18.4. Estos se tomaron de Davidson y MacKinnon (1993, tabla 20.2).

TABLA 18.4

Valores críticos asintóticos para la prueba de cointegración: sin tendencia temporal

Nivel de significancia	1%	2.5%	5%	10%
Valor crítico	−3.90	−3.59	−3.34	−3.04

En la prueba básica se realiza la regresión de $\Delta\hat{u}_t$ en $\hat{u}_{t-1}$ y se compara el estadístico t sobre $\hat{u}_{t-1}$ con el valor crítico deseado en la tabla 18.4. Si el estadístico t está por debajo del valor crítico, se tiene evidencia de que $y_t - \beta x_t$ es I(0) para cierta β; es decir y_t y x_t están cointegradas. Se pueden agregar rezagos de $\Delta\hat{u}_t$ para considerar correlación serial. Si se comparan los valores críticos en la tabla 18.4 con aquellos de la tabla 18.2, se debe obtener un estadístico t mucho mayor en magnitud para determinar la cointegración que si se utilizan los valores críticos DF. Esto sucede debido a que MCO, que minimiza la suma de los residuales cuadrados, tiende a producir residuales que se ven como una secuencia I(0) aun si y_t y x_t *no están* cointegradas.

Como con la prueba habitual de Dickey-Fuller, se puede aumentar la prueba Engle-Granger al incluir rezagos de $\Delta\hat{u}_t$ como regresores adicionales.

Si y_t y x_t no están cointegradas, una regresión de y_t en x_t es espuria y no muestra nada significativo: no hay relación a largo plazo entre y y x. Aún se puede hacer una regresión que implique las primeras diferencias, Δy_t y Δx_t, incluidos los rezagos. Pero es necesario interpretar estas regresiones como lo que son: ellas explican la diferencia en y en términos de la diferencia en x y no tiene que ver necesariamente nada con una relación en niveles.

Si y_t y x_t *están* cointegradas, se puede utilizar este hecho para especificar modelos dinámicos más generales, como se verá en la siguiente subsección.

El análisis anterior supone que ni y_t ni x_t tienen tendencia (estocástica). Esto es razonable para tasas de interés, pero no para otras series de tiempo. Si y_t y x_t contienen términos con tendencia (estocástica), $E(y_t)$ y $E(x_t)$ son funciones lineales (en general crecientes) de tiempo. La definición estricta de cointegración requiere que $y_t - \beta x_t$ sea I(0) *sin* una tendencia. Para ver qué implica esto, se escribe $y_t = \delta t + g_t$ y $x_t = \lambda t + h_t$, donde $\{g_t\}$ y $\{h_t\}$ son procesos I(1), δ es la tendencia (estocástica) en y_t [$\delta = E(\Delta y_t)$] y λ es la tendencia (estocástica) en x_t [$\lambda = E(\Delta x_t)$]. Ahora, si y_t y x_t están cointegradas, debe existir una β tal que $g_t - \beta h_t$ sea I(0). Pero entonces

$$y_t - \beta x_t = (\delta - \beta\lambda)t + (g_t - \beta h_t),$$

que por lo general es un proceso de *tendencia estacionaria*. La forma estricta de cointegración requiere que no exista una tendencia, lo cual significa que $\delta = \beta\lambda$. Para procesos I(1) con tendencia (estocástica) es posible que las partes estocásticas, es decir, g_t y h_t estén cointegrados, pero que el parámetro β que ocasiona que $g_t - \beta h_t$ sea I(0) no elimine la tendencia temporal lineal.

Se puede probar la cointegración entre g_t y h_t, sin detenerse en la parte de la tendencia, al realizar la regresión

$$\hat{y}_t = \hat{\alpha} + \hat{\eta} t + \hat{\beta} x_t \qquad \boxed{18.32}$$

y aplicar la prueba usual DF o la prueba DF aumentada a los residuales $\hat{u}_t$. Los valores críticos asintóticos se dan en la tabla 18.5 [de Davidson y MacKinnon (1993, tabla 20.2)].

TABLA 18.5

Valores críticos asintóticos para la prueba de cointegración: tendencia temporal lineal

Nivel de significancia	1%	2.5%	5%	10%
Valor crítico	−4.32	−4.03	−3.78	−3.50

Un hallazgo de cointegración en este caso deja abierta la posibilidad de que $y_t - \beta x_t$ tenga una tendencia lineal. Pero al menos no es I(1).

Ejemplo 18.5

[Cointegración entre fertilidad y exención personal]

En los capítulos 10 y 11 se estudiaron diferentes modelos para estimar la relación entre la tasa general de fertilidad (*gfr*) y el valor real de la exención personal de impuestos (*pe*) en Estados Unidos. Los resultados estadísticos de la regresión en niveles y primeras diferencias son notablemente distintos. La regresión en niveles, con una tendencia temporal incluida, da un coeficiente MCO sobre *pe* igual a .187 (ee = .035) y R^2 = .500. En primeras diferencias (sin una tendencia), el coeficiente sobre Δpe es −.043 (ee = .028), y R^2 = .032. Aunque existen otras razones para estas diferencias, como la especificación incorrecta en la dinámica de rezagos distribuidos, la discrepancia entre los niveles y cambios de regresiones sugiere que se debe probar la cointegración. Por supuesto, esto presupone que *gfr* y *pe* son procesos I(1). Éste parece ser el caso, las pruebas DF aumentadas, con un solo cambio rezagado y una tendencia temporal lineal, cada una produce estadísticos *t* de aproximadamente −1.47, y los coeficientes estimados AR(1) son cercanos a uno.

Cuando se obtienen los residuales de la regresión de *gfr* sobre *t* y *pe* y se aplica la prueba DF aumentada con un rezago, se obtiene un estadístico *t* en $\hat{u}_{t-1}$ de −2.43, el cual no está cerca de ninguna manera del valor crítico a 10%, −3.50. En consecuencia, se debe concluir que existe poca evidencia de cointegración entre *gfr* y *pe*, inclusive si se permiten tendencias separadas. Es muy probable que los primeros resultados de la regresión, que se obtuvieron en los niveles, tengan algún problema de regresión espuria.

La buena noticia es que, cuando se usaron las primeras diferencias y se permitieron dos rezagos, vea la ecuación (11.27), se encontró un efecto general positivo y significativo a largo plazo de Δpe sobre Δgfr.

Si se piensa que dos series están cointegradas, suele quererse probar las hipótesis acerca del parámetro cointegrador. Por ejemplo, una teoría puede afirmar que el parámetro cointegrador es uno. En teoría, se podría usar un estadístico *t* para probar esta hipótesis.

De forma explícita se cubre el caso sin tendencias temporales, aunque la extensión al caso de tendencia lineal es inmediato. Cuando y_t y x_t son I(1) y cointegradas, se puede escribir

$$y_t = \alpha + \beta x_t + u_t, \tag{18.33}$$

donde u_t es un proceso I(0) con media cero. En general, $\{u_t\}$ contiene una correlación serial, pero se sabe del capítulo 11 que esto no afecta la consistencia de MCO. Como se mencionó antes, MCO aplicado a (18.33) estima de forma consistente β (y α). Por desgracia, dado que x_t es I(1), los procedimientos acostumbrados de inferencia no aplican necesariamente: MCO no está distribuido de manera asintóticamente normal y el estadístico *t* para $\hat{\beta}$ no necesariamente tiene una distribución *t* aproximada. Se sabe del capítulo 10 que, si $\{x_t\}$ es estrictamente exógena (vea supuesto TS.3), y los errores son homocedásticos, serialmente no correlacionados y distribuidos normalmente, el estimador MCO también está distribuido normalmente (condicional sobre las variables explicativas) *además* que el estadístico *t* tiene una distribución *t* exacta. Por desgracia, estos supuestos son demasiado fuertes para aplicarse a la mayoría de las situaciones. La noción de cointegración no implica nada acerca de la relación entre $\{x_t\}$ y $\{u_t\}$, de hecho, pueden estar correlacionadas de forma arbitraria. Además, salvo el caso que se requiera que $\{u_t\}$ sea I(0), la cointegración entre y_t y x_t no restringe la dependencia serial en $\{u_t\}$.

Por fortuna, la característica de (18.33) que hace que la inferencia sea más difícil, la falta de exogeneidad estricta de $\{x_t\}$ puede ser fijada. Ya que x_t es I(1), la noción adecuada de la exogeneidad estricta es que u_t no está correlacionada con Δx_s, para toda *t* y *s*. Siempre se puede ordenar esto para un *nuevo* conjunto de errores, al menos de forma aproximada, al escribir u_t en función de Δx_s para toda s cercana a t. Por ejemplo,

$$u_t = \eta + \phi_0\Delta x_t + \phi_1\Delta x_{t-1} + \phi_2\Delta x_{t-2} + \gamma_1\Delta x_{t+1} + \gamma_2\Delta x_{t+2} + e_t, \tag{18.34}$$

donde, por construcción, e_t no está correlacionada con cada Δx_s que aparece en la ecuación. Es de esperar que e_t no esté correlacionada con más rezagos y adelantos de Δx_s. Se sabe que, a medida que $|s - t|$ aumenta, la correlación entre e_t y Δx_s se acerca a cero, debido a que éstos son procesos I(0). Ahora, si se inserta (18.34) en (18.33), se obtiene

$$y_t = \alpha_0 + \beta x_t + \phi_0\Delta x_t + \phi_1\Delta x_{t-1} + \phi_2\Delta x_{t-2} + \gamma_1\Delta x_{t+1} + \gamma_2\Delta x_{t+2} + e_t. \tag{18.35}$$

Esta ecuación parece un poco extraña debido a que las Δx_s futuras aparecen tanto con valores rezagados y actuales de Δx_t. La clave es que el coeficiente en x_t aún es β, y, por construcción, x_t ahora es estrictamente exógena en esta ecuación. El supuesto de exogeneidad estricta es la condición necesaria para obtener un estadístico t aproximadamente normal para $\hat{\beta}$. Si u_t no está correlacionada con ninguna Δx_s, $s \neq t$, entonces se pueden descartar las diferencias tanto hacia delante como hacia atrás y simplemente incluir la diferencia contemporánea, Δx_t. Entonces, la ecuación que se estima parece más estándar, pero sigue incluyendo la primera diferencia de x_t junto con su nivel: $y_t = \alpha_0 + \beta x_t + \phi_0\Delta x_t + e_t$. En efecto, sumar Δx_t resuelve toda endogeneidad contemporánea entre x_t y u_t. (Recuerde, cualquier endogeneidad no ocasiona inconsistencia. Pero se trata de obtener un estadístico t asintóticamente normal.) El que se necesiten incluir las diferencias tanto adelantadas como rezagadas y saber cuántas de ellas incorporar, en realidad es una cuestión empírica. Cada vez que se agrega una diferencia adelantada o rezagada adicional, se pierde una observación, y esto puede ser costoso a menos que se tenga un conjunto grande de datos.

El estimador MCO de la β a partir de (18.35) recibe el nombre de **estimador de adelantos y rezagos** de β debido a la forma en que emplea Δx. [Vea, por ejemplo, Stock y Watson (1993).] La única cuestión que debe preocupar en (18.35) es la posibilidad de correlación serial en $\{e_t\}$. Ésta puede manejarse calculando un error estándar robusto a la correlación serial para $\hat{\beta}$ (como se describió en la sección 12.5) o al utilizar una corrección estándar AR(1) (como la de Cochrane-Orcutt).

Ejemplo 18.6

[Parámetro de cointegración para las tasas de interés]

En secciones anteriores se probó la cointegración entre *r6* y *r3*, de las tasas de los certificados del Tesoro a seis y tres meses, al suponer que el parámetro de cointegración era igual a uno. Esto condujo a hallar la cointegración y, naturalmente, a concluir que el parámetro de cointegración es igual a la unidad. No obstante, se estimará directamente el parámetro de cointegración y se probará H_0: $\beta = 1$. Se aplica el estimador de adelantos y rezagos con dos adelantos y dos rezagos de $\Delta r3$, así como la diferencia contemporánea. La estimación de β es $\hat{\beta} = 1.038$, y el error estándar habitual MCO es .0081. Por tanto, el estadístico t para H_0: $\beta = 1$ es $(1.038 - 1)/.0081 \approx 4.69$, lo cual es un rechazo estadístico sólido de H_0. (Por supuesto, si 1.038 es diferente de 1 en términos económicos, es una consideración relevante.) Hay poca evidencia de correlación serial en los residuales, así que se pueda utilizar este estadístico t como si tuviera una distribución normal aproximada. [Por comparación, la estimación por MCO de β sin los adelantos, rezagos o términos $\Delta r3$ contemporáneos, y usar cinco observaciones más, es 1.026 (ee = .0077). Pero el estadístico t de (18.33) no necesariamente es válido.]

Existen muchos otros estimadores de parámetros de cointegración, lo que continúa siendo un área muy activa de la investigación. La noción de cointegración aplica a más de dos procesos, pero la interpretación, las pruebas y la estimación son mucho más complicadas. Un problema es que, incluso después de normalizar un coeficiente para que sea uno, puede haber diversas relaciones de cointegración. BDGH ofrece algunos análisis y varias referencias.

Modelos de corrección del error

Además de aprender acerca de una posible relación a largo plazo entre dos series, el concepto de cointegración enriquece el tipo de modelos dinámicos de los que se dispone. Si y_t y x_t son procesos I(1) y *no* están cointegrados, se puede estimar un modelo dinámico en primeras diferencias. Por ejemplo, considere la siguiente ecuación

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + u_t, \tag{18.36}$$

donde u_t tiene media cero dadas Δx_t, Δy_{t-1}, Δx_{t-1} y rezagos adicionales. Esta es en esencia la ecuación (18.16), pero en primeras diferencias y no en niveles. Si se ve esto como un modelo racional de rezagos distribuidos, se puede encontrar el multiplicador de impacto, el multiplicador de largo plazo y la distribución del rezago para Δy como un rezago distribuido en Δx.

Si y_t y x_t están cointegradas con el parámetro β, entonces se tienen variables I(0) adicionales que se pueden incluir en (18.36). Sea $s_t = y_t - \beta x_t$, de manera que s_t sea I(0) y suponga, por simplicidad, que s_t tiene media cero. Ahora, se pueden incluir rezagos de s_t en la ecuación. En el caso más simple, se incluye un rezago de s_t:

$$\begin{aligned} \Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta s_{t-1} + u_t \\ &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta (y_{t-1} - \beta x_{t-1}) + u_t, \end{aligned} \tag{18.37}$$

donde $\mathrm{E}(u_t|I_{t-1}) = 0$ y I_{t-1} contiene información sobre Δx_t y todos los valores pasados de x y y. El término $\delta(y_{t-1} - \beta x_{t-1})$ recibe el nombre de *término de corrección del error* y (18.37) es un ejemplo de un **modelo de corrección del error**. (En algunos modelos de corrección del error, el cambio contemporáneo en x, Δx_t, se omite. Que se incluya o no depende en parte del propósito de la ecuación. En el pronóstico, Δx_t rara vez se incluye, por razones que se verán en la sección 18.5.)

Un modelo de corrección del error, permite estudiar la dinámica a corto plazo en la relación entre y y x. Por simplicidad, considere el modelo sin los rezagos de Δy_t y Δx_t:

$$\Delta y_t = \alpha_0 + \gamma_0 \Delta x_t + \delta (y_{t-1} - \beta x_{t-1}) + u_t, \tag{18.38}$$

donde $\delta < 0$. Si $y_{t-1} > \beta x_{t-1}$, entonces y en el periodo anterior rebasa el equilibrio; debido a que $\delta < 0$, el término de corrección del error funciona para retornar y al equilibrio. Asimismo, si $y_{t-1} < \beta x_{t-1}$, el término de corrección de errores induce un cambio positivo en y que la devuelve al equilibrio.

¿Cómo se estiman los parámetros de un modelo de corrección de errores? Si se conoce β, es sencillo. Por ejemplo, en (18.38), simplemente se hace la regresión de Δy_t sobre Δx_t y s_{t-1}, donde $s_{t-1} = (y_{t-1} - \beta x_{t-1})$.

Ejemplo 18.7

[Modelo de corrección de errores para el rendimiento de bonos]

En el problema 11.6 se efectuó la regresión de $hy6_t$, el rendimiento de bonos a tres meses (porcentual) de comprar un certificado del Tesoro a seis meses en el tiempo $t - 1$ y venderlo en el tiempo t como un certificado del Tesoro a tres meses, en $hy3_{t-1}$, el rendimiento del bono a tres meses de comprar un certificado del Tesoro a tres meses en el tiempo $t - 1$. La hipótesis de expectativas implica que el coeficiente de la pendiente no debe ser estadísticamente diferente de uno. Resulta que existe evidencia de una raíz unitaria en $\{hy3_t\}$, que cuestiona el análisis de regresión estándar. Se supondrá que ambos rendimientos de los bonos son procesos I(1). La hipótesis de expectativas implica que, como mínimo, $hy6_t$ y $hy3_{t-1}$ están cointegradas con β igual a uno, lo cual parece ser el caso (vea el ejercicio para computadora C18.5). De acuerdo con este supuesto, un modelo de corrección de error es

Pregunta 18.4

¿Cómo probaría H_0: $\gamma_0 = 1$, $\delta = -1$ en el modelo de corrección del error del rendimiento?

$$\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2}) + u_t,$$

donde u_t tiene media cero, dado que $hy3$ y $hy6$ datan del tiempo $t - 1$ y anteriores. Los rezagos en las variables en el modelo de corrección del error están determinados por las hipótesis de expectativas.

Mediante los datos en INTQRT.RAW se obtiene

$$\begin{aligned} \widehat{\Delta hy6_t} &= \underset{(.043)}{.090} + \underset{(.264)}{1.218}\, \Delta hy3_{t-1} - \underset{(.244)}{.840}\, (hy6_{t-1} - hy3_{t-2}) \\ n &= 122,\ R^2 = .790. \end{aligned} \tag{18.39}$$

El coeficiente de la corrección del error es negativo y muy significativo. Por ejemplo, si el rendimiento de los certificados del Tesoro a seis meses está por encima de aquellos a tres meses por un punto, $hy6$ cae .84 puntos, en promedio, el siguiente trimestre. Es interesante que, $\hat{\delta} = -.84$ no sea estadísticamente diferente de -1, como se puede ver fácilmente al calcular el intervalo de confianza de 95%.

En muchos otros ejemplos se debe estimar el parámetro de cointegración. A continuación, se remplaza s_{t-1} con $\hat{s}_{t-1} = y_{t-1} - \hat{\beta}x_{t-1}$, donde $\hat{\beta}$ pueden ser varios estimadores de β. Se ha tratado el estimador estándar MCO así como el estimador de adelantos y rezagos. Esto plantea la cuestión de cómo afecta la variación muestral en $\hat{\beta}$ la inferencia sobre otros parámetros en el modelo de corrección de error. Por fortuna, como lo muestran Engle y Granger (1987), se puede ignorar la estimación preliminar de β (asintóticamente). Esta propiedad es muy conveniente e implica que la eficiencia asintótica de los estimadores de los parámetros en el modelo de corrección del error no se ve afectado por si se utiliza el estimador de MCO o el estimador de adelantos y rezagos de $\hat{\beta}$. Por supuesto, la elección de $\hat{\beta}$ por lo general tiene un efecto en los parámetros en el modelo de corrección de errores en cualquier muestra en particular, pero no hay forma sistemática de decidir qué estimador preliminar de β utilizar. El procedimiento de remplazar β con $\hat{\beta}$ recibe el nombre de **procedimiento de dos pasos de Engle-Granger**.

18.5 Elaboración de pronósticos

Pronosticar series de tiempo económicas es muy importante en algunas ramas de la economía y es un área que continúa siendo estudiada activamente. En esta sección se estudiará la regresión basada en métodos de pronóstico. Diebold (2001) ofrece una introducción exhaustiva de los pronósticos, incluidos sus desarrollos recientes.

Se asume, en esta sección, que el enfoque principal se centrará en el pronóstico de los valores futuros de un proceso de series de tiempo y no necesariamente en estimar modelos económicos causales o estructurales.

Es útil primero cubrir algunas cuestiones básicas de los pronósticos que no dependen de un modelo específico. Suponga que en el tiempo t se desea pronosticar el resultado de y en el tiempo $t + 1$, o y_{t+1}. El periodo podría corresponder a un año, un trimestre, un mes, una semana o incluso un día. Sea I_t la información que se puede observar en el tiempo t. Este **conjunto de información** incluye a y_t, los primeros valores de y seguidos con frecuencia de otras variables fechadas en el tiempo t o antes. Se puede combinar esta información de innumerables maneras para pronosticar y_{t+1}. ¿Hay una forma mejor?

La respuesta es sí, siempre y cuando se especifique la *pérdida* asociada con el error de pronóstico. Sea f_t el pronóstico de y_{t+1} hecho en el tiempo t. Se conoce a f_t como **pronóstico de un paso hacia delante**. El **error de pronóstico** es $e_{t+1} = y_{t+1} - f_t$, que se observa una vez que se obtiene el resultado en y_{t+1}. La medida más común de la pérdida es la misma que lleva a la estimación de los mínimos cuadrados ordinarios de un modelo de regresión lineal múltiple: el error cuadrático, e^2_{t+1}. El error cuadrático de pronóstico trata simétricamente errores de predicción positivos y negativos y los errores de pronósticos más amplios reciben relativamente más peso. Por ejemplo, los errores de $+2$ y -2 reportan la misma pérdida y esta es cuatro veces tan grande como los errores de pronósticos de $+1$ o -1. El error cuadrático de pronóstico es un ejemplo de **función de pérdida**. Otra función de pérdida popular es el valor absoluto del error de predicción, $|e_{t+1}|$. Por las razones que se verán en breve, ahora nos enfocaremos en la pérdida cuadrática del error.

Dada la función de pérdida cuadrática del error, se puede determinar cómo usar mejor la información en el tiempo t para pronosticar y_{t+1}. Pero se reconoce que en el tiempo t, no se conoce e_{t+1}: es una variable aleatoria, debido a que y_{t+1} es una variable aleatoria. Por tanto, cualquier criterio útil para elegir f_t debe basarse en lo que se conoce en el tiempo t. Es natural elegir el pronóstico para minimizar el error cuadrático *esperado* del pronóstico, dada I_t:

$$\mathrm{E}(e^2_{t+1}|I_t) = \mathrm{E}[(y_{t+1} - f_t)^2|I_t]. \tag{18.40}$$

Un hecho básico de la probabilidad (vea la propiedad CE.6 en el apéndice B) es que la esperanza condicional, $\mathrm{E}(y_{t+1}|I_t)$, minimiza (18.40). En otras palabras, si se desea minimizar el error cuadrático esperado del pronóstico dada la información en el tiempo t, el pronóstico debe ser el valor esperado de y_{t+1} dadas las variables que se conocen en el tiempo t.

Para muchos procesos conocidos de series de tiempo, la esperanza condicional es fácil de obtener. Suponga que $\{y_t: t = 0, 1, \ldots\}$ es una secuencia de una martingala en diferencia (SMD) y que I_t es $\{y_t, y_{t-1}, \ldots, y_0\}$, el pasado observado de y por definición, $\mathrm{E}(y_{t+1}|I_t) = 0$ para toda t; la mejor predicción de y_{t+1} en el tiempo t es siempre cero. De la sección 18.2 recuerde que una secuencia i.i.d. con una media de cero es una secuencia de una martingala en diferencia.

Una secuencia de una martingala en diferencia es aquella en la que el pasado no es útil para predecir el futuro. Se considera que los rendimientos de las acciones están bien aproximados a una SMD o, quizá, con una media positiva. La clave es que $\mathrm{E}(y_{t+1}|y_t, y_{t-1}, \ldots) = \mathrm{E}(y_{t+1})$: la media condicional es igual a la media incondicional, en cuyo caso, la y pasada no sirve para predecir la y futura.

Un proceso $\{y_t\}$ es una **martingala** si $E(y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = y_t$ para toda $t \geq 0$. [Si $\{y_t\}$ es una martingala, entonces $\{\Delta y_t\}$ es una secuencia de una martingala en diferencia, que es de donde procede su nombre.] El valor predicho de y para el siguiente periodo es siempre el valor de y para este periodo.

Un ejemplo más complicado es

$$E(y_{t+1}|I_t) = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \ldots + \alpha(1 - \alpha)^t y_0, \quad \textbf{18.41}$$

donde $0 < \alpha < 1$ es un parámetro que se debe elegir. Este método de pronóstico recibe el nombre de **suavizamiento exponencial**, porque los pesos de las y rezagadas declinan a cero exponencialmente.

La razón de escribir la esperanza como en (18.41) es que lleva a una relación de recurrencia muy simple. Sea $f_0 = y_0$. Entonces, para $t \geq 1$, los pronósticos pueden obtenerse como

$$f_t = \alpha y_t + (1 - \alpha)f_{t-1}.$$

En otras palabras, el pronóstico de y_{t+1} es un promedio ponderado de y_t y el pronóstico de y_t hecho en el tiempo $t - 1$. El suavizamiento exponencial es aconsejable sólo para series de tiempo muy específicas y requiere que se elija α. Los métodos de regresión, los cuales se analizarán en seguida, son más flexibles.

El análisis anterior se enfocó en el pronóstico de y sólo un periodo en el futuro. Las cuestiones generales que surgen de pronosticar y_{t+h} en el tiempo t, donde *h es* cualquier entero positivo, son similares. En particular, si se usa el error cuadrático esperado del pronóstico como una medida de la pérdida, el mejor pronóstico es $E(y_{t+h}|I_t)$. Cuando se trata con un **pronóstico de múltiples pasos hacia delante**, se emplea una notación $f_{t,h}$ para indicar el pronóstico de y_{t+h} hecho en el tiempo t.

Tipos de modelos de regresión empleados para pronósticos

Existen numerosos modelos de regresión diferentes que pueden utilizarse para pronosticar valores de una serie de tiempo. El primer modelo de regresión para datos de series de tiempo del capítulo 10 fue el modelo estático. Para ver cómo se puede pronosticar con este modelo, se supone que se tiene una sola variable explicativa:

$$y_t = \beta_0 + \beta_1 z_t + u_t. \quad \textbf{18.42}$$

Suponga, por el momento, que se conocen los parámetros β_0 y β_1. Escriba esta ecuación en el tiempo $t + 1$ a medida que $y_{t+1} = \beta_0 + \beta_1 z_{t+1} + u_{t+1}$. Ahora, *if* z_{t+1} se conoce en el tiempo t, así que es un elemento de I_t y $E(u_{t+1}|I_t) = 0$, entonces

$$E(y_{t+1}|I_t) = \beta_0 + \beta_1 z_{t+1},$$

donde I_t contiene $z_{t+1}, y_t, z_t, \ldots, y_1, z_1$. El lado derecho de esta ecuación es el pronóstico de y_{t+1} en el tiempo t. Este tipo de pronóstico suele conocerse como **pronóstico condicional** debido a que es condicional sobre el conocimiento del valor de z en el tiempo $t + 1$.

Por desgracia, en cualquier momento, rara vez se conoce el valor de las variables explicativas en periodos futuros. Las excepciones incluyen tendencias de tiempo y variables estacionales binarias, que se analizarán a detalle más adelante, pero de otra forma el conocimiento de z_{t+1} al tiempo t es raro. En ocasiones se desea generar pronósticos condicionales para varios valores de z_{t+1}.

Otro problema con (18.42) como modelo de pronóstico es que $E(u_{t+1}|I_t) = 0$ significa que $\{u_t\}$ no puede contener una correlación serial, algo que se ha visto como falso en la mayoría de los modelos de regresión estáticos. [El problema 18.8 le pide derivar el pronóstico en un modelo de rezagos distribuidos con errores AR(1).]

Si z_{t+1} no se conoce en el tiempo t, no se le puede incluir en I_t. Entonces, se tiene

$$\mathrm{E}(y_{t+1}|I_t) = \beta_0 + \beta_1 \mathrm{E}(z_{t+1}|I_t).$$

Esto significa que, con el fin de pronosticar y_{t+1}, primero se debe pronosticar z_{t+1}, con base en el mismo conjunto de información. Esto se conoce como **pronóstico no condicional**, pues no se supone el conocimiento de z_{t+1} al tiempo t. Por desgracia, este es un nombre inadecuado, ya que el pronóstico sigue siendo condicional de la información en I_t. Pero el nombre se ha arraigado en la literatura sobre pronósticos.

Para los pronósticos, a menos de que por otras razones se deba usar el modelo estático en (18.42), tiene más sentido especificar un modelo que sólo dependa de los valores rezagados de y y z. Esto ahorra el paso adicional de tener que pronosticar la variable del lado derecho antes de pronosticar y. El tipo de modelo que se tiene aquí en mente es

$$\begin{gathered} y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t \\ \mathrm{E}(u_t|I_{t-1}) = 0, \end{gathered} \qquad \boxed{\textbf{18.43}}$$

donde I_{t-1} contiene y y z fechadas al tiempo $t - 1$ y antes. Ahora, el pronóstico de y_{t+1} en el tiempo t es $\delta_0 + \alpha_1 y_t + \gamma_1 z_t$; si se conocen los parámetros, se pueden insertar los valores de y_t y z_t.

Si sólo se quiere utilizar las y pasadas para predecir las y futuras, entonces se puede descartar z_{t-1} de (18.43). Naturalmente, se pueden agregar más rezagos de y o z además de rezagos de otras variables. Tales modelos pueden ser muy útiles, en especial para pronósticos de un paso hacia delante.

Pronóstico de un paso hacia delante

Obtener el pronóstico de un periodo después de que la muestra ha finalizado es relativamente fácil mediante modelos como (18.43). Como es usual, sea n el tamaño de la muestra. El pronóstico de y_{n+1} es

$$\hat{f}_n = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n, \qquad \boxed{\textbf{18.44}}$$

donde se supone que los parámetros se estimaron mediante MCO. Se usa un gorro sobre f_n para enfatizar que se han estimado los parámetros en el modelo de regresión. (Si se conocieran los parámetros, no habría error de estimación en el pronóstico.) El error de pronóstico, que no se conocerá sino hasta el momento $n + 1$, es

$$\hat{e}_{n+1} = y_{n+1} - \hat{f}_n. \qquad \boxed{\textbf{18.45}}$$

Si se agregan más rezagos de y o z a la ecuación de pronóstico, simplemente se pierden más observaciones al principio de la muestra.

El pronóstico $\hat{f}_n$ de y_{n+1} por lo general se conoce como **pronóstico puntual**. También se puede obtener un **intervalo de pronóstico**. Un intervalo de pronóstico es, en esencia, lo mismo que el intervalo de predicción, que se estudió en la sección 6.4. Ahí se mostró cómo obtener un intervalo de predicción exacto de 95%, bajo los supuestos del modelo lineal clásico. Un intervalo de pronóstico se obtiene *exactamente* de la misma forma. Si el modelo no satisface los supuestos del modelo lineal clásico, por ejemplo, si contiene variables dependientes rezagadas, como en (18.44), el intervalo de pronóstico todavía es aproximadamente válido, siempre y cuando u_t dada I_{t-1}, tenga una distribución normal con media cero y varianza constante. (Esto asegura que los estimadores MCO estén distribuidos de manera aproximadamente normal con las varianzas MCO usuales y que u_{n+1} sea independiente de los estimadores MCO con media cero y varianza σ^2.) Sea ee($\hat{f}_n$) el error estándar del pronóstico y sea $\hat{\sigma}$ el error estándar de la regresión. [De la sección 6.4, se puede obtener $\hat{f}_n$ y ee($\hat{f}_n$) como el intercepto y su error estándar de la regresión de y_t en $(y_{t-1} - y_n)$ y $(z_{t-1} - z_n)$, $t = 1, 2, \ldots, n$; es decir, se resta el valor de y al tiempo n de cada y rezagada y se hace lo mismo con z antes de efectuar la regresión.] Así,

$$\mathrm{ee}(\hat{e}_{n+1}) = \{[\mathrm{ee}(\hat{f}_n)]^2 + \hat{\sigma}^2\}^{1/2}, \tag{18.46}$$

y el intervalo de pronóstico (aproximado) a 95% es

$$\hat{f}_n \pm 1.96 \cdot \mathrm{ee}(\hat{e}_{n+1}). \tag{18.47}$$

Debido a que ee($\hat{f}_n$) es aproximadamente proporcional a $1/\sqrt{n}$, ee($\hat{f}_n$) por lo general, es pequeño con relación a la incertidumbre en el error u_{n+1}, medido por $\hat{\sigma}$. [Algunos paquetes de econometría calculan de manera rutinaria intervalos de pronóstico, pero otros requieren algunas sencillas manipulaciones para obtener (18.47).]

Ejemplo 18.8

[Pronóstico de la tasa de desempleo en Estados Unidos]

Se utilizan los datos en PHILLIPS.RAW, pero sólo para el periodo de 1948 a 1996, para pronosticar la tasa de desempleo civil en Estados Unidos para 1997. Se utilizan dos modelos. El primero es un modelo simple AR(1) para *unem*:

$$\begin{aligned} \widehat{unem}_t &= \underset{(.577)}{1.572} + \underset{(.097)}{.732}\, unem_{t-1} \\ n &= 48, \bar{R}^2 = .544, \hat{\sigma} = 1.049. \end{aligned} \tag{18.48}$$

En un segundo modelo se suma la inflación con rezago de un año:

$$\begin{aligned} \widehat{unem}_t &= \underset{(.490)}{1.304} + \underset{(.084)}{.647}\, unem_{t-1} + \underset{(.041)}{.184}\, inf_{t-1} \\ n &= 48, \bar{R}^2 = .677, \hat{\sigma} = .883. \end{aligned} \tag{18.49}$$

La tasa de inflación rezagada es muy significativa en (18.49) ($t \approx 4.5$) y la R-cuadrada ajustada de la segunda ecuación es mucho más alta que aquella de la primera. Sin embargo, esto *no* necesariamente significa que la segunda ecuación producirá un mejor pronóstico para 1997. Todo lo que se puede decir hasta ahora es que, usando los datos hasta 1996, un rezago en la inflación ayuda a explicar la variación en la tasa de desempleo.

Para obtener los pronósticos para 1997, se necesita conocer *unem* e *inf* en 1996. Éstos son de 5.4 y 3.0, respectivamente. Por tanto, el pronóstico de $unem_{1997}$ de la ecuación (18.48) es 1.572 + .732(5.4), o alrededor de 5.52. El pronóstico de la ecuación (18.49) es 1.304 + .647(5.4) + .184(3.0), o aproximadamente 5.35. La tasa de desempleo civil real para 1997 fue de 4.9, así que ambas ecuaciones predicen un monto excesivo de la tasa real. La segunda ecuación proporciona un pronóstico un poco mejor.

Se puede obtener fácilmente un intervalo de pronóstico a 95%. Cuando se hace una regresión de $unem_t$ sobre ($unem_{t-1}$ − 5.4) y (inf_{t-1} − 3.0), se obtiene 5.35 como intercepto, que ya se calculó como el pronóstico, y ee($\hat{f}_n$) = .137. Por consiguiente, dado que $\hat{\sigma}$ = .883, se observa que ee($\hat{e}_{n+1}$) = $[(.137)^2 + (.883)^2]^{1/2} \approx .894$. El intervalo de pronóstico a 95% de (18.47) es 5.35 ± 1.96(.894), o aproximadamente [3.6, 7.1]. Este es un amplio intervalo y el valor conocido de 1997, 4.9, está dentro del intervalo. Como se esperaba, el error estándar de u_{n+1}, que es .883 es una fracción muy grande de ee($\hat{e}_{n+1}$).

Un pronosticador profesional debe elaborar un pronóstico para cada periodo. Por ejemplo, en el tiempo n, él o ella produce un pronóstico de y_{n+1}. Después, cuando y_{n+1} y z_{n+1} se conocen él o ella deben pronosticar y_{n+2}. Incluso si el pronosticador ha elegido el modelo (18.43), existen dos opciones para pronosticar y_{n+2}. La primera es usar $\hat{\delta}_0 + \hat{\alpha}_1 y_{n+1} + \hat{\gamma}_1 z_{n+1}$, donde los parámetros se estiman utilizando las primeras n observaciones. La segunda posibilidad es *volver a estimar* los parámetros utilizando todas las observaciones $n + 1$ y después utilizar la misma fórmula para pronosticar y_{n+2}. Para pronosticar en periodos consecutivos, por lo general, se usan las estimaciones de parámetros obtenidas de las n observaciones iniciales, o se pueden actualizar los parámetros de regresión cada vez para obtener un nuevo punto de datos. Aunque el último método requiere más cálculos, la carga adicional es relativamente menor y puede funcionar mejor debido a que los coeficientes de regresión se ajustan, al menos en parte, a los nuevos puntos de datos.

Como ejemplo específico, suponga que desea pronosticar la tasa de desempleo para 1998, mediante el modelo con un solo rezago de *unem* e *inf*. La primera posibilidad es sólo insertar los valores de 1997 de desempleo e inflación del lado derecho de (18.49). Con *unem* = 4.9 e *inf* = 2.3 en 1997, se tiene un pronóstico para $unem_{1998}$ de alrededor de 4.9. (Sólo es coincidencia que éste sea igual a la tasa de desempleo de 1997.) La segunda posibilidad es volver a estimar la ecuación al agregar la observación de 1997 y después utilizar esta nueva ecuación (vea el ejercicio para computadora C18.6).

El modelo en la ecuación (18.43) es una ecuación en lo que se conoce como **modelo de vectores autorregresivos (VAR)**. Se sabe que un modelo autorregresivo, del capítulo 11: modela una única serie, $\{y_t\}$, en términos de su propio pasado. En los modelos de vectores autorregresivos, se modelan varias series, de las cuales, si usted está familiarizado con el álgebra lineal, es de donde proviene la palabra "vector", en términos de su propio pasado. Si se tienen dos series, y_t y z_t, un vector autorregresivo consiste en ecuaciones parecidas a

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + \alpha_2 y_{t-2} + \gamma_2 z_{t-2} + \ldots \qquad \boxed{18.50}$$

y

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + \beta_2 y_{t-2} + \rho_2 z_{t-2} + \ldots,$$

donde cada ecuación contiene un error que tiene un valor esperado de cero dada la información pasada en y y z. En la ecuación (18.43), y en el ejemplo estimado en (18.49), se supone que un rezago en cada variable capturó toda la dinámica. (Una prueba F para la significancia conjunta de $unem_{t-2}$ y inf_{t-2} confirma que sólo se necesita un rezago de cada una.)

Como lo ilustra el ejemplo 18.8, los modelos VAR pueden ser útiles para pronosticar. En varios casos se tiene interés en pronosticar sólo una variable, y, en cuyo caso sólo es necesario estimar y analizar la ecuación para y. Nada impide que se agreguen otras variables rezagadas, por ejemplo, w_{t-1}, w_{t-2}, …, a la ecuación (18.50). Tales ecuaciones se estiman eficientemente por MCO, siempre y cuando se hayan incluido suficientes rezagos de todas las variables y la ecuación satisfaga el supuesto de homocedasticidad para regresiones de series de tiempo.

Las ecuaciones como (18.50) permiten probar si, *después de controlar las y pasadas*, las z pasadas ayudan a pronosticar y_t. En general, se dice que *z causa en el sentido de Granger* a y si

$$E(y_t|I_{t-1}) \neq E(y_t|J_{t-1}), \tag{18.51}$$

donde I_{t-1} contiene información pasada sobre y *además de z* y J_{t-1} contiene sólo información sobre el pasado de y. Cuando (18.51) es válida, el pasado de z es útil, *además del pasado de y*, para predecir y_t. El término "causa" en "causa en el sentido de Granger" debe interpretarse con cuidado. El único sentido en que z "causa" y se da en (18.51). En particular, no muestra nada acerca de la causalidad *contemporánea* entre y y z, así que no permite determinar si z_t es una variable exógena o endógena en una ecuación que relaciona y_t con z_t. (Ésta también es la razón de que la noción de **causalidad de Granger** no aplique en contextos de cortes transversales puros.)

Una vez que se hace el supuesto de un modelo lineal y se decide cuántos rezagos de y se deben incluir en $E(y_t|y_{t-1}, y_{t-2}, \ldots)$, es fácil probar la hipótesis nula de que z *no* causa en el sentido de Granger a y. Para ser más específicos, suponga que $E(y_t|y_{t-1}, y_{t-2}, \ldots)$ depende de sólo tres rezagos:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + u_t$$
$$E(u_t|y_{t-1}, y_{t-2}, \ldots) = 0.$$

Ahora, bajo la hipótesis nula de que z no causa en el sentido de Granger a y, cualquier rezago de z que se agrega a la ecuación debe tener cero coeficientes poblacionales. Si se agrega z_{t-1}, entonces simplemente se hace una prueba t sobre z_{t-1}. Si se agregan dos rezagos de z, entonces se puede realizar una prueba F para la significancia conjunta de z_{t-1} y z_{t-2} en la ecuación

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + u_t.$$

(Si existe heterocedasticidad, se puede utilizar una forma robusta de la prueba. No puede ser una correlación serial bajo H_0 debido a que el modelo está dinámicamente completo.)

Como cuestión práctica, ¿cómo se decide qué rezagos de y y z incluir? Primero, se empieza por estimar un modelo autorregresivo para y, se realizan pruebas t y F para determinar cuántos rezagos de y deben aparecer. Con datos anuales, el número de rezagos suele ser pequeño, por ejemplo, de uno o dos. Con datos trimestrales o mensuales, suele haber muchos más rezagos. Una vez que se ha elegido un modelo autorregresivo para y, se pueden probar los rezagos de z. La elección de rezagos de z tiene menor importancia debido a que cuando z no causa en el sentido de Granger a y, ningún conjunto de z rezagadas debe ser significativo. Con datos anuales, 1 o 2 rezagos se suelen utilizar; con datos trimestrales, por lo general 4 u 8; y con datos mensuales, 6, 12 o quizás hasta 24, dados suficientes datos.

Ya se ha mostrado un ejemplo de prueba de la causalidad de Granger en la ecuación (18.49). El modelo autorregresivo que mejor se adapta al desempleo es un AR(1). En la ecuación (18.49), se agregó un solo rezago de inflación y fue muy significativo. Por tanto, la inflación es causa Granger del desempleo.

Existe una definición extendida de la causalidad de Granger que suele ser útil. Sea $\{w_t\}$ una tercera serie (o podría representar varias series adicionales). Entonces, *z es causa Granger de y condicionada sobre w* si (18.51) es válida, pero ahora I_{t-1} contiene información pasada sobre y, z además de w, mientras que J_{t-1} contiene información pasada sobre y y w. Con seguridad es posible que y cause en el sentido de Granger a z, pero que z no cause en el sentido de Granger a y condicionada sobre w. Una prueba de la hipótesis nula de que y no cause en el sentido de Granger a z condicionada sobre w, se obtiene al probar la significancia de z rezagada en un modelo para y que también depende de las y y w rezagadas. Por ejemplo, para probar si el crecimiento en la oferta monetaria causa en el sentido de Granger al crecimiento en el GDP real, condicionado sobre el cambio en las tasas de interés, se realizaría una regresión $gGDP_t$ en los rezagos de $gGDP$, Δint y gM y se hacen pruebas de significancia en los rezagos de gM. [Vea, por ejemplo, Stock y Watson (1989).]

Comparación de pronósticos de un paso hacia delante

En casi cualquier problema de pronóstico, existen varios modelos que compiten entre sí para tal propósito. Inclusive cuando se restringe la atención a los modelos de regresión, existen muchas posibilidades. ¿Qué variables se deben incluir y con cuántos rezagos? ¿Se deben usar logs, niveles de variables o primeras diferencias?

Con el fin de decidir sobre un método de pronóstico, es necesaria una forma de elegir cuál es el idóneo. En términos generales, se puede distinguir entre **criterios dentro de la muestra** y **criterios fuera de la muestra**. En un contexto de regresión, los criterios dentro de la muestra incluyen a la R-cuadrada y en especial a la R-cuadrada ajustada. Existen muchos otros *estadísticos de selección de modelos*, pero sólo se cubrirán éstos [vea, por ejemplo, Ramanathan (1995, capítulo 4)].

Para realizar pronósticos es mejor utilizar los criterios fuera de la muestra, puesto que los pronósticos son, en esencia, un problema fuera de la muestra. Un modelo podría ofrecer un buen ajuste a y en la muestra utilizada para estimar los parámetros. Pero esta necesidad no se traduce en un buen desempeño de pronósticos. Una comparación fuera de la muestra implica utilizar la primera parte de una muestra para estimar los parámetros del modelo y ahorrar la última parte de la muestra para calibrar sus capacidades de pronóstico. Esto imita lo que se habría de hacer en la práctica si no se conocieran aún los valores futuros de las variables.

Suponga que se tienen $n + m$ observaciones, donde se utilizan las primeras n observaciones para estimar los parámetros en el modelo y ahorrar las últimas m observaciones para el pronóstico. Sea $\hat{f}_{n+h}$ un pronóstico de un paso hacia delante de y_{n+h+1} para $h = 0, 1, \ldots, m - 1$. Los errores de pronóstico m son $\hat{e}_{n+h+1} = y_{n+h+1} - \hat{f}_{n+h}$. ¿Cómo se deben medir qué tan bien pronostica y el modelo cuando está fuera de la muestra? Dos mediciones son las más comunes. La primera es la **raíz del error cuadrático medio (RECM)**:

$$RECM = \left(m^{-1}\sum_{h=0}^{m-1}\hat{e}_{n+h+1}^2\right)^{1/2}. \tag{18.52}$$

Esta es, en esencia, la desviación estándar de la muestra de los errores pronosticados (sin ningún grado de ajuste de libertad). Si se calcula la RECM para dos o más métodos de pronóstico, entonces se preferirá el método con la RECM fuera de la muestra menor.

Una segunda medición común es el **error absoluto medio (EAM)**, que es el promedio de los errores pronosticados absolutos:

$$EAM = m^{-1}\sum_{h=0}^{m-1}|\hat{e}_{n+h+1}|. \tag{18.53}$$

De nuevo se prefiere un EAM menor. Otros posibles criterios incluyen minimizar el mayor de los valores absolutos de los errores de pronóstico.

Ejemplo 18.9

[Comparaciones fuera de la muestra de los pronósticos de desempleo]

En el ejemplo 18.8 se encontró que la ecuación (18.49) se adapta notablemente mejor durante los años 1948 a 1996 que la ecuación (18.48) y, al menos para pronosticar el desempleo en 1997, el modelo que incluía la inflación rezagada funcionó mejor. Ahora se utilizan los dos modelos, y se continúa estimando mediante los datos sólo hasta 1996, para comparar pronósticos de un paso hacia delante de 1997 a 2003. Esto deja siete observaciones fuera de la muestra ($n = 48$ y $m = 7$) para utilizar en las ecuaciones (18.52) y (18.53). Para el modelo AR(1), RECM = .962 y EAM = .778. Para el modelo que agrega la inflación rezagada (un modelo VAR de orden uno), RECM = .673 y EAM = .628. Así, por cualquier medición, el modelo que incluye inf_{t-1} produce mejores pronósticos fuera de la muestra para 1997 a 2003. En este caso los criterios fuera y dentro de la muestra eligen el mismo modelo.

En lugar de utilizar sólo las primeras n observaciones para estimar los parámetros del modelo, se pueden volver a estimar los modelos cada vez que se agrega una nueva observación y se utiliza el nuevo modelo para pronosticar el siguiente periodo de tiempo.

Pronósticos de múltiples pasos hacia delante

Pronosticar más de un periodo por adelantado, por lo general es más difícil que pronosticar uno solo. Se puede formalizar esto de la siguiente manera. Suponga que se considera pronosticar y_{t+1} al tiempo t y en un periodo de tiempo anterior s (así que $s < t$). Entonces $\mathrm{Var}[y_{t+1} - \mathrm{E}(y_{t+1}|I_t)] \leq \mathrm{Var}[y_{t+1} - \mathrm{E}(y_{t+1}|I_s)]$, donde la desigualdad suele ser estricta. No se demostrará que este resultado tiene sentido de manera general, sino intuitiva: la varianza en el error de pronóstico en la predicción y_{t+1} es mayor cuando se efectúa ese pronóstico con base en menos información.

Si $\{y_t\}$ se desprende de un modelo AR(1) (que incluye una caminata aleatoria, posiblemente con tendencia (estocástica)), se puede mostrar con facilidad que la varianza del error aumenta con el horizonte de pronóstico. El modelo es

$$y_t = \alpha + \rho y_{t-1} + u_t$$
$$\mathrm{E}(u_t|I_{t-1}) = 0,\ I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\},$$

y $\{u_t\}$ tiene una varianza constante σ^2 condicional sobre I_{t-1}. Al tiempo $t + h - 1$, el pronóstico de y_{t+h} es $\alpha + \rho y_{t+h-1}$ y el error de pronóstico simplemente es u_{t+h}. Por tanto, la varianza del pronóstico de un paso hacia delante es σ^2. Para elaborar pronósticos de múltiples pasos hacia delante, se tiene por sustitución repetida,

$$\begin{aligned} y_{t+h} = {} & (1 + \rho + \ldots + \rho^{h-1})\alpha + \rho^h y_t \\ & + \rho^{h-1}u_{t+1} + \rho^{h-2}u_{t+2} + \ldots + u_{t+h}. \end{aligned}$$

Al tiempo t, el valor esperado de u_{t+j}, para toda $j \geq 1$, es cero. Así que

$$\mathrm{E}(y_{t+h}|I_t) = (1 + \rho + \ldots + \rho^{h-1})\alpha + \rho^h y_t, \tag{18.54}$$

y el error de pronóstico es $e_{t,h} = \rho^{h-1}u_{t+1} + \rho^{h-2}u_{t+2} + \ldots + u_{t+h}$. Esta es una suma de las variables aleatorias no correlacionadas y por tanto, la varianza de la suma es la suma de las varianzas: $\mathrm{Var}(e_{t,h}) = \sigma^2[\rho^{2(h-1)} + \rho^{2(h-2)} + \ldots + \rho^2 + 1]$. Debido a que $\rho^2 > 0$, cada término que multiplica a σ^2 es positivo, así que la varianza del error del pronóstico aumenta con h. Cuando $\rho^2 < 1$, a medida que h aumenta la varianza de pronóstico converge a $\sigma^2/(1 - \rho^2)$, que es sólo la varianza incondicional de y_t. En el caso de una caminata aleatoria ($\rho = 1$), $f_{t,h} = \alpha h + y_t$ y $\mathrm{Var}(e_{t,h}) = \sigma^2 h$: la varianza del pronóstico aumenta sin límite a medida que el horizonte h aumenta. Esto demuestra que es muy difícil pronosticar una caminata aleatoria, con o sin tendencia (estocástica), en el futuro lejano. Por ejemplo, los pronósticos de tasas de interés en el futuro lejano pierden radicalmente su precisión.

La ecuación (18.54) muestra que al utilizar el modelo AR(1) para el pronóstico de múltiples pasos es fácil, una vez que se ha estimado ρ mediante MCO. El pronóstico de y_{n+h} al tiempo n es

$$\hat{f}_{n,h} = (1 + \hat{\rho} + \ldots + \hat{\rho}^{h-1})\hat{\alpha} + \hat{\rho}^h y_n. \quad \mathbf{18.55}$$

Obtener los intervalos de pronóstico es más difícil, a menos que $h = 1$, debido a que obtener el error estándar de $\hat{f}_{n,h}$ es difícil. Sin embargo, el error estándar de $\hat{f}_{n,h}$ por lo general, es pequeño en comparación con la desviación estándar del término de error y éste se puede estimar como $\hat{\sigma}[\hat{\rho}^{2(h-1)} + \hat{\rho}^{2(h-2)} + \ldots + \hat{\rho}^2 + 1]^{1/2}$, donde $\hat{\sigma}$ es el error estándar de la regresión de la estimación AR(1). Es posible utilizar este hecho para obtener un intervalo de confianza aproximado. Por ejemplo, cuando $h = 2$, un intervalo de confianza aproximado a 95% (para n grande) es

$$\hat{f}_{n,2} \pm 1.96\hat{\sigma}(1 + \hat{\rho}^2)^{1/2}. \quad \mathbf{18.56}$$

Debido a que se está subestimando la desviación estándar de y_{n+h}, este intervalo es demasiado estrecho, pero quizás no tanto, en especial si n es grande.

Un enfoque menos tradicional, pero útil es estimar un modelo diferente para cada horizonte de pronóstico. Por ejemplo, suponga que se desea pronosticar y dos periodos adelante. Si I_t depende sólo de y hasta el tiempo t, se supondría que $E(y_{t+2}|I_t) = \alpha_0 + \gamma_1 y_t$ [que, como se vio antes, es válido si $\{y_t\}$ sigue a un modelo AR(1)]. Se puede estimar α_0 y γ_1 al realizar una regresión de y_t sobre un intercepto y sobre y_{t-2}. Aunque los errores en esta ecuación contienen una correlación serial, los errores en periodos adyacentes están correlacionados, se pueden obtener estimadores consistentes y aproximadamente normales de α_0 y γ_1. El pronóstico de y_{n+2} al tiempo n es simplemente $\hat{f}_{n,2} = \hat{\alpha}_0 + \hat{\gamma}_1 y_n$. Además, y muy importante, el error estándar de la regresión, es justo lo que se necesita para calcular un intervalo de confianza para el pronóstico. Por desgracia, para obtener el error estándar de $\hat{f}_{n,2}$, mediante el truco para un pronóstico de un paso hacia delante se requiere que se obtenga un error estándar robusto a la correlación serial del tipo que se describió en la sección 12.5. Este error estándar tiende a cero a medida que n aumenta, mientras que la varianza del error es constante. Por tanto, se puede obtener un intervalo aproximado al usar (18.56) y al poner EER de la regresión de y_t sobre y_{t-2} en lugar de $\hat{\sigma}(1 + \hat{\rho}^2)^{1/2}$. Pero se debe recordar que esto ignora el error de estimación en $\hat{\alpha}_0$ y $\hat{\gamma}_1$.

También se pueden calcular los pronósticos de múltiples pasos hacia delante con modelos autorregresivos más complicados. Por ejemplo, suponga que $\{y_t\}$ sigue un modelo AR(2) y que en el tiempo n, se desea pronosticar y_{n+2}. Ahora, $y_{n+2} = \alpha + \rho_1 y_{n+1} + \rho_2 y_n + u_{n+2}$, así que

$$E(y_{n+2}|I_n) = \alpha + \rho_1 E(y_{n+1}|I_n) + \rho_2 y_n.$$

Se puede escribir esto como

$$f_{n,2} = \alpha + \rho_1 f_{n,1} + \rho_2 y_n,$$

De manera que se pueden obtener pronósticos con dos pasos hacia delante en el tiempo n una vez que se obtiene el pronóstico con un paso hacia delante. Si los parámetros del modelo AR(2) se estimaron mediante MCO, entonces se opera como

$$\hat{f}_{n,2} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,1} + \hat{\rho}_2 y_n. \quad \mathbf{18.57}$$

Ahora, $\hat{f}_{n,1} = \hat{\alpha} + \hat{\rho}_1 y_n + \hat{\rho}_2 y_{n-1}$, el cual es posible calcularlo en el tiempo n. Entonces, se inserta esto en (18.57), junto con y_n, para obtener $\hat{f}_{n,2}$. Para toda $h > 2$, es fácil obtener de manera recursiva cualquier pronóstico h pasos hacia delante para un modelo AR(2): $\hat{f}_{n,h} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,h-1} + \hat{\rho}_2 \hat{f}_{n,h-2}$.

Un razonamiento similar se puede utilizar para obtener pronósticos de pasos múltiples hacia delante para modelos VAR. Para ilustrar esto suponga que se tiene

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t \tag{18.58}$$

y

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + v_t.$$

Ahora, si se desea pronosticar y_{n+1} al tiempo n, simplemente se utiliza $\hat{f}_{n,1} = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n$. Asimismo, el pronóstico de z_{n+1} al tiempo n es (por ejemplo) $\hat{g}_{n,1} = \hat{\eta}_0 + \hat{\beta}_1 y_n + \hat{\rho}_1 z_n$. Ahora, suponga que se desea obtener el pronóstico de dos pasos hacia delante de y al tiempo n. De (18.58), se tiene

$$\mathrm{E}(y_{n+2}|I_n) = \delta_0 + \alpha_1 \mathrm{E}(y_{n+1}|I_n) + \gamma_1 \mathrm{E}(z_{n+1}|I_n)$$

[debido a que $\mathrm{E}(u_{n+2}|I_n) = 0$], así que se puede escribir el pronóstico como

$$\hat{f}_{n,2} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,1} + \hat{\gamma}_1 \hat{g}_{n,1}. \tag{18.59}$$

Esta ecuación muestra que el pronóstico de dos pasos hacia delante para y depende de los pronósticos de un paso hacia delante para y y z. En general, se pueden construir pronósticos de múltiples pasos hacia delante para y mediante la fórmula recursiva

$$\hat{f}_{n,h} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,h-1} + \hat{\gamma}_1 \hat{g}_{n,h-1}, \quad h \geq 2.$$

Ejemplo 18.10

[Pronóstico a dos años de la tasa de desempleo]

Para utilizar la ecuación (18.49) para pronosticar el desempleo a dos años, por ejemplo, la tasa de 1998 utilizando los datos hasta 1996, es necesario un modelo para la inflación. El mejor modelo para *inf* en términos *unem* e *inf* rezagadas parece ser un simple modelo AR(1) ($unem_{-1}$ no es significativo cuando se agrega a la regresión):

$$\begin{aligned} \widehat{inf}_t &= \underset{(.558)}{1.277} + \underset{(.107)}{.665}\, inf_{t-1} \\ n &= 48,\ R^2 = .457,\ \bar{R}^2 = .445. \end{aligned}$$

Si se inserta el valor de 1996 de *inf* en esta ecuación, se obtiene el pronóstico de *inf* para 1997: $\widehat{inf}_{1997} = 3.27$. Ahora, se puede insertar esto, junto con $\widehat{unem}_{1997} = 5.35$ (que se obtuvo antes), en (18.59) para pronosticar $unem_{1998}$:

$$\widehat{unem}_{1998} = 1.304 + .647(5.35) + .184(3.27) \approx 5.37.$$

Recuerde, este pronóstico utiliza información sólo hasta 1996. El pronóstico de un paso hacia delante de $unem_{1998}$, obtenido al insertar los valores de 1997 de *unem* e *inf* en (18.48), fue de aproximadamente 4.90. La tasa de desempleo real en 1998 fue 4.5%, lo cual significa que, en este caso, el pronóstico de un paso hacia delante es un poco mejor que el pronóstico de dos pasos hacia delante.

Tal como con el pronóstico de un paso hacia delante, es posible emplear la raíz del error cuadrático medio o error absoluto medio fuera de la muestra para elegir entre métodos de pronóstico de múltiples pasos.

Pronóstico de tendencia, estacionalidad y procesos integrados

Ahora, se analizará la forma de pronosticar series que exhiben tendencias, contienen estacionalidad o tienen raíces unitarias. Recuerde de los capítulos 10 y 11 que un método para manejar variables dependientes o independientes en los modelos de regresión es incluir tendencias de tiempo, la más popular de ellas es la tendencia lineal. Las tendencias se pueden incluir también en las ecuaciones de pronósticos, aunque deben utilizarse con cuidado.

En el caso más simple, suponga que $\{y_t\}$ tiene una tendencia lineal pero es impredecible alrededor de esa tendencia. Entonces, se puede escribir

$$y_t = \alpha + \beta t + u_t,\ \mathrm{E}(u_t|I_{t-1}) = 0,\ t = 1, 2, \ldots, \qquad \boxed{\textbf{18.60}}$$

donde, como es costumbre, I_{t-1} contiene la información observada a través del tiempo $t - 1$ (que incluye al menos la y pasada). ¿Cómo se pronostica y_{n+h} al tiempo n para toda $h \geq 1$? Esto es simple debido a que $\mathrm{E}(y_{n+h}|I_n) = \alpha + \beta(n + h)$. La varianza del error del pronóstico es simplemente $\sigma^2 = \mathrm{Var}(u_t)$ (si se supone una varianza constante en el tiempo). Si se estima α y β por MCO mediante las primeras n observaciones, entonces el pronóstico para y_{n+h} al tiempo n es $\hat{f}_{n,h} = \hat{\alpha} + \hat{\beta}(n + h)$. En otras palabras, simplemente se inserta el periodo de tiempo correspondiente a y en la función de tendencia estimada. Por ejemplo, si se utilizan $n = 131$ observaciones en BARIUM.RAW para pronosticar las importaciones chinas mensuales a Estados Unidos de cloruro de bario, se obtiene $\hat{\alpha} = 249.56$ y $\hat{\beta} = 5.15$. El periodo de la muestra termina en diciembre de 1988, así que el pronóstico de importaciones chinas seis meses después es $249.56 + 5.15(137) = 955.11$, medido como toneladas cortas. En términos comparativos el valor de diciembre de 1988 es 1,087.81, así que es mayor que el valor pronosticado seis meses después. La serie y su línea de tendencia estimada se muestran en la figura 18.2.

Como se analizó en el capítulo 10, la mayoría de las series de tiempo económicas se caracterizan mejor cuando tienen, al menos aproximadamente, una tasa constante de crecimiento, lo cual sugiere que $\log(y_t)$ sigue una tendencia temporal lineal. Suponga que se utilizan n observaciones para obtener la ecuación

$$\widehat{\log(y_t)} = \hat{\alpha} + \hat{\beta}t,\ t = 1, 2, \ldots, n. \qquad \boxed{\textbf{18.61}}$$

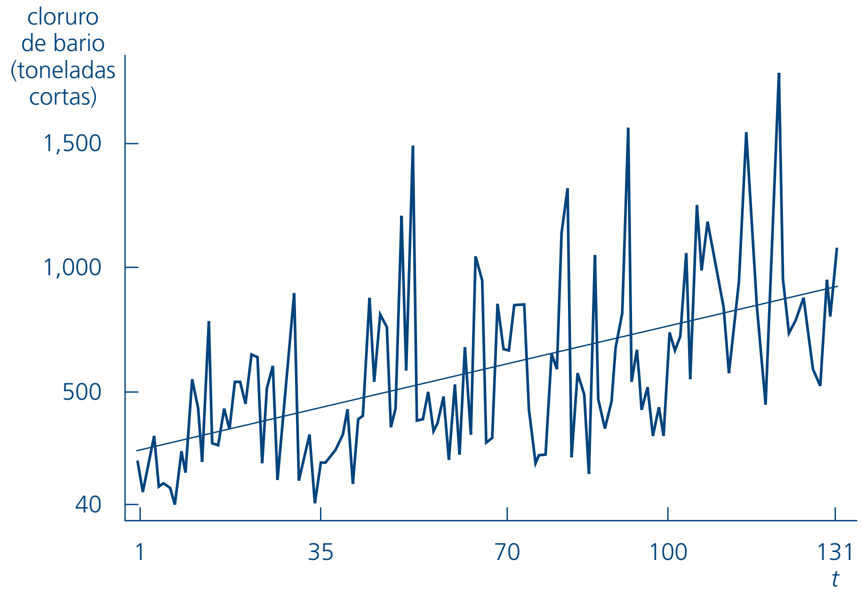
Entonces, para pronosticar $\log(y)$ en cualquier periodo futuro $n + h$, se debe insertar $n + h$ en la ecuación de tendencia, como antes. Pero esto no permite pronosticar y, que suele ser lo deseado. Es tentador simplemente exponenciar $\hat{\alpha} + \hat{\beta}(n + h)$ para obtener el pronóstico de y_{n+h}, pero esto no es correcto, por las mismas razones que se dieron en la sección 6.4. Se debe tomar en cuenta el error implícito en (18.61). La forma más sencilla de hacer esto es utilizar n observaciones para efectuar la regresión de y_t sobre $\exp(\widehat{logy_t})$ *sin* un intercepto.

Pregunta 18.5

Suponga que se tiene un modelo $\{y_t: t = 1, 2, \ldots, 46\}$ como tendencia temporal lineal, donde los datos comienzan en 1950 y terminan en 1995. Defina la variable $year_t$ como $t = 1$ a partir de 1950, de tal modo que en 1995 $t = 46$. Si se estima la ecuación $\hat{y}_t = \hat{\gamma} + \hat{\delta}year_t$, ¿cómo se comparan $\hat{\gamma}$ y $\hat{\delta}$ con $\hat{\alpha}$ y $\hat{\beta}$ en $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$? ¿Cómo se compararán los pronósticos de las dos ecuaciones?

FIGURA 18.2

Importaciones chinas a Estados Unidos de cloruro de bario (en toneladas cortas) y su línea de tendencia lineal estimada, 249.56 + 5.15*t*.



Sea $\hat{\gamma}$ el coeficiente de la pendiente en $\exp(\widehat{logy_t})$. Entonces, el pronóstico de y en el periodo $n + h$ es simplemente

$$\hat{f}_{n,h} = \hat{\gamma}\exp[\hat{\alpha} + \hat{\beta}(n + h)]. \tag{18.62}$$

Como un ejemplo, con los datos de las primeras 687 semanas del índice de Precios y Cotizaciones de la Bolsa de Nueva York en NYSE.RAW se obtiene $\hat{\alpha} = 3.782$ y $\hat{\beta} = .0019$ [al efectuar la regresión de $\log(price_t)$ sobre una tendencia temporal lineal]; esto muestra que el índice crece alrededor de .2% a la semana, en promedio. Cuando se hace la regresión de *price* sobre los valores ajustados exponenciados, se obtiene $\hat{\gamma} = 1.018$. Ahora se pronostica *price* que habrá cuatro semanas después, lo cual es la última semana en la muestra, utilizando (18.62): $1.018 \cdot \exp[3.782 + .0019(691)] \approx 166.12$. El valor real resultó ser 164.25, así que se observa que se sobrestimó, pero este resultado es mucho mejor que si se hubiera estimado la tendencia temporal lineal para las primeras 687 semanas: el valor pronosticado para la semana 691 es 152.23, lo cual es una predicción sustancialmente inferior.

Aunque los modelos de tendencia son útiles para hacer predicciones, deben utilizarse con cuidado, en especial cuando se pronostican, en el futuro lejano, series integradas que tienen tendencias (estocásticas). El problema puede observarse cuando se considera una caminata aleatoria con tendencia (estocástica). En el tiempo $t + h$, se escribe y_{t+h} como

$$y_{t+h} = \beta h + y_t + u_{t+1} + \ldots + u_{t+h},$$

donde β es el término de la tendencia (estocástica) (por lo general, $\beta > 0$), y cada u_{t+j} tiene media cero dada I_t y una varianza constante σ^2. Como se vio antes, el pronóstico de y_{t+h} al tiempo

t es $E(y_{t+h}|I_t) = \beta h + y_t$ y la varianza del error de pronóstico es de $\sigma^2 h$. ¿Qué sucede si se utiliza el modelo de tendencia lineal? Sea y_0 el valor inicial del proceso al tiempo cero, que se considera como no aleatorio. Así se puede escribir también

$$\begin{aligned} y_{t+h} &= y_0 + \beta(t + h) + u_1 + u_2 + \ldots + u_{t+h} \\ &= y_0 + \beta(t + h) + v_{t+h}. \end{aligned}$$

Esto parece un modelo de tendencia lineal con el intercepto $\alpha = y_0$. Pero el error, v_{t+h}, si bien tiene media cero, tiene varianza $\sigma^2(t + h)$. Por tanto, si se utiliza la tendencia lineal $y_0 + \beta(t + h)$ para pronosticar y_{t+h} al tiempo t, la varianza del error de pronóstico es $\sigma^2(t + h)$, comparada con $\sigma^2 h$ cuando se utiliza $\beta h + y_t$. La razón de las varianzas pronosticadas es $(t + h)/h$, la cual puede ser importante para una t grande. El resultado final es que no se debe utilizar una tendencia lineal para pronosticar una caminata aleatoria con tendencia (estocástica). (El ejercicio para computadora C18.8 le pide comparar pronósticos de una línea de tendencia cúbica y aquellos del modelo simple de caminata aleatoria para la tasa de fertilidad general en Estados Unidos).

Las tendencias deterministas también pueden producir pronósticos pobres si se estiman los parámetros de tendencias mediante datos antiguos y si el proceso tiene un cambio posterior en la línea de la tendencia. Algunas veces, influencias exógenas, como las crisis petroleras de la década de los setenta, pueden cambiar la trayectoria de las variables con tendencia. Si se recurre a una tendencia antigua para pronosticar el futuro lejano, es posible que el pronóstico esté equivocado. Este problema se puede mitigar utilizando los datos disponibles más recientes para obtener los parámetros de la línea de tendencia.

Nada impide combinar tendencias con otros modelos de pronóstico. Por ejemplo, se puede agregar una tendencia lineal a un modelo AR(1), lo cual puede funcionar bien para pronosticar series con tendencias lineales, pero que también sean procesos estables AR en torno a la tendencia.

También es fácil pronosticar procesos con estacionalidad determinista (series mensuales o trimestrales). Por ejemplo, el archivo BARIUM.RAW contiene la producción mensual de gasolina en Estados Unidos de 1978 a 1988. Esta serie no tiene una tendencia evidente, pero sí tiene un fuerte patrón estacional. (La producción de gasolina es más alta en los meses de verano y en diciembre.) En el modelo más simple, se realizaría una regresión de *gas* (medida en galones) sobre 11 binarias mensuales, por ejemplo, de febrero a diciembre. Después, el pronóstico para cualquier mes futuro es simplemente el intercepto más el coeficiente en la binaria del mes apropiado. (Para enero, el pronóstico es sólo el intercepto en la regresión.) También se pueden agregar rezagos de variables y tendencias temporales para considerar series generales con estacionalidad.

Pronosticar procesos con raíces unitarias también merece atención especial. Ya se obtuvo antes el valor esperado de una caminata aleatoria condicionada a la información en el tiempo n. Para pronosticar una caminata aleatoria, con una posible tendencia (estocástica) α, h periodos en el futuro al tiempo n, se utiliza $\hat{f}_{n,h} = \hat{\alpha}h + y_n$, donde $\hat{\alpha}$ es el promedio muestral de Δy_t hasta $t = n$. (Si no existe tendencia (estocástica), se considera $\hat{\alpha} = 0$.) Este método impone la raíz unitaria. Una alternativa sería estimar un modelo AR(1) para $\{y_t\}$ y utilizar la fórmula de pronóstico (18.55). Este enfoque no impone una raíz unitaria, pero si una está presente, $\hat{\rho}$ converge en probabilidad a uno a medida que n crece. Sin embargo, $\hat{\rho}$ puede ser sustancialmente diferente de uno, en especial si el tamaño muestral no es muy grande. La cuestión de qué método produce mejores pronósticos fuera de la muestra es empírica. Si en el modelo AR(1), ρ es menor que uno, inclusive ligeramente, el modelo AR(1) tenderá a producir mejores pronósticos a largo plazo.

En general, existen dos enfoques para producir pronósticos para los procesos I(1). El primero es imponer una raíz unitaria. Para un pronóstico de un paso hacia delante, se obtiene un modelo para pronosticar el cambio en y, Δy_{t+1}, dada la información a través del tiempo t. Entonces, debido a que $y_{t+1} = \Delta y_{t+1} + y_t$, $\mathrm{E}(y_{t+1}|I_t) = \mathrm{E}(\Delta y_{t+1}|I_t) + y_t$. Por tanto, el pronóstico de y_{n+1} al tiempo n es

$$\hat{f}_n = \hat{g}_n + y_n,$$

donde $\hat{g}_n$ es el pronóstico de Δy_{n+1} al tiempo n. Suele emplearse un modelo AR (que necesariamente es estable) para Δy_t, o un vector autorregresivo.

Esto se puede extender a los pronósticos de múltiples pasos hacia delante al escribir y_{n+h} como

$$y_{n+h} = (y_{n+h} - y_{n+h-1}) + (y_{n+h-1} - y_{n+h-2}) + \ldots + (y_{n+1} - y_n) + y_n,$$

o

$$y_{n+h} = \Delta y_{n+h} + \Delta y_{n+h-1} + \ldots + \Delta y_{n+1} + y_n.$$

Por consiguiente, el pronóstico de y_{n+h} al tiempo n es

$$\hat{f}_{n,h} = \hat{g}_{n,h} + \hat{g}_{n,h-1} + \ldots + \hat{g}_{n,1} + y_n, \tag{18.63}$$

donde $\hat{g}_{n,j}$ es el pronóstico de Δy_{n+j} en el tiempo n. Por ejemplo, se puede hacer un modelo de Δy_t como AR(1) estable, obtener los pronósticos de pasos múltiples hacia delante de (18.55) (pero con $\hat{\alpha}$ y $\hat{\rho}$ obtenidas de Δy_t en Δy_{t-1} y y_n remplazada con Δy_n), y después insertar éstos en (18.63).

El segundo enfoque para pronosticar variables I(1) es usar un modelo AR o VAR para $\{y_t\}$. Esto no impone la raíz unitaria. Por ejemplo, si se emplea un modelo AR(2),

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + u_t, \tag{18.64}$$

entonces $\rho_1 + \rho_2 = 1$. Si se inserta en $\rho_1 = 1 - \rho_2$ y se reordena, se obtiene $\Delta y_t = \alpha - \rho_2 \Delta y_{t-1} + u_t$, que es un modelo estable AR(1) en la diferencia que lleva al primer método que se describió antes. Nada impide estimar (18.64) directamente mediante MCO. Un asunto curioso acerca de esta regresión es que *se puede* emplear el estadístico habitual t sobre $\hat{\rho}_2$ para determinar si y_{t-2} es significativa. (Esto es que el supuesto de homocedasticidad es válido; si no, se puede utilizar la forma robusta de homocedasticidad.) No se demostrará esto formalmente, pero de manera intuitiva, esto se deduce al reescribir la ecuación como $y_t = \alpha + \gamma y_{t-1} - \rho_2 \Delta y_{t-1} + u_t$, donde $\gamma = \rho_1 + \rho_2$. Aun si $\gamma = 1$, ρ_2 es menor que el coeficiente en un proceso estacionario y débilmente dependiente $\{\Delta y_{t-1}\}$. Debido a que los resultados de la regresión serán idénticos a (18.64), se puede emplear directamente (18.64).

Por ejemplo, se estimará un modelo AR(2) para la tasa de fertilidad general en FERTIL3.RAW, mediante las observaciones hasta 1979. (En el ejercicio para computadora C18.8, se le pide utilizar este modelo para pronosticar, lo cual es la razón de que se ahorren algunas observaciones al final de la muestra).

$$\begin{aligned} \widehat{gfr}_t &= \underset{(2.92)}{3.22} + \underset{(.120)}{1.272}\, gfr_{t-1} - \underset{(.121)}{.311}\, gfr_{t-2} \\ n &= 65,\ R^2 = .949,\ \bar{R}^2 = .947. \end{aligned} \tag{18.65}$$

El estadístico t en el segundo rezago es de alrededor de -2.57, que es estadísticamente diferente de cero al nivel de 1%. (El primer rezago también tiene un estadístico t muy significativo que tiene una distribución t aproximada, por el mismo razonamiento empleado para $\hat{\rho}_2$.) La R-cuadrada, ajustada o no, es especialmente informativa como medida de bondad de ajuste debido a que *gfr* contiene en apariencia una raíz unitaria y tiene poco sentido preguntar qué tanto de la varianza en *gfr* se está explicando.

Los coeficientes en los dos rezagos en (18.65) suman .961, el cual es cercano a, y no es estadísticamente diferente de uno (como puede verificarse al aplicar la prueba de Dickey-Fuller aumentada a la ecuación $\Delta gfr_t = \alpha + \theta gfr_{t-1} + \delta_1 \Delta gfr_{t-1} + u_t$). Aunque no se impuso la restricción de la raíz unitaria, aún se puede utilizar (18.65) para el pronóstico, como se analizó antes.

Antes de terminar esta sección, se debe señalar una posible mejora al pronóstico en el contexto de los modelos de vectores autorregresivos con variables I(1). Suponga que $\{y_t\}$ y $\{z_t\}$ son, cada una, procesos I(1). Un método para obtener pronósticos de y es estimar una autorregresión bivariada en las variables Δy_t y Δz_t y después emplear (18.63) para generar pronósticos de uno o de múltiples pasos hacia delante; esto es en esencia, el primer método que antes se describió. No obstante, si y_t y z_t están *cointegradas*, se tienen más variables estables y estacionarias en el conjunto de información que se puede utilizar para pronosticar Δy: a saber, rezagos de $y_t - \beta z_t$, donde β es el parámetro de cointegración. Un modelo simple de corrección de errores es

$$\begin{gathered}\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_1 \Delta z_{t-1} + \delta_1 (y_{t-1} - \beta z_{t-1}) + e_t, \\ \mathrm{E}(e_t | I_{t-1}) = 0. \end{gathered} \tag{18.66}$$

Para pronosticar y_{n+1}, se emplean las observaciones hasta n para estimar el parámetro de cointegración, β, y después estimar los parámetros del modelo de corrección de error por MCO, como se describió en la sección 18.4. Pronosticar Δy_{n+1} es fácil: sólo se deben insertar Δy_n, Δz_n y $y_n - \hat{\beta} z_n$ en la ecuación estimada. Una vez obtenido el pronóstico de Δy_{n+1}, se agrega a y_n.

Al reordenar el modelo de corrección de error, se obtiene

$$y_t = \alpha_0 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \tag{18.67}$$

donde $\rho_1 = 1 + \alpha_1 + \delta$, $\rho_2 = -\alpha_1$, y así sucesivamente, lo cual es la primera ecuación en un modelo VAR para y_t y z_t. Observe que esto depende de cinco parámetros, tantos como en el modelo de corrección de error. El punto es que, para fines de pronóstico, el modelo VAR en los niveles y el modelo de corrección de error son esencialmente lo mismo. Este no es el caso en los modelos de corrección del error más generales. Por ejemplo, suponga que $\alpha_1 = \gamma_1 = 0$ en (18.66), pero se tiene un segundo término de corrección del error, $\delta_2(y_{t-2} - \beta z_{t-2})$. Entonces, el modelo de corrección del error implica sólo cuatro parámetros, mientras que (18.67), que tiene el mismo orden de rezagos para y y z contiene cinco parámetros. Por tanto, los modelos de corrección del error pueden economizar en parámetros; es decir, por lo general son más *parsimoniosos* que los modelos VAR en niveles.

Si y_t y z_t son I(1) pero no están cointegrados, el modelo apropiado es (18.66) sin el término de corrección de errores. Esto se puede emplear para pronosticar Δy_{n+1}, se puede agregar esto a y_n para pronosticar y_{n+1}.

RESUMEN

Los temas de series de tiempo que se analizaron en este capítulo suelen utilizarse en macroeconomía empírica, finanzas empíricas y varios otros campos aplicados. Se comenzó mostrando cómo se pueden interpretar y estimar los modelos de rezagos distribuidos. Éstos pueden proporcionar distribuciones flexibles de rezagos con menos parámetros que un modelo de rezagos distribuidos finito similar. El rezago distribuido geométrico y, en general, los modelos de rezagos distribuidos racionales son los más conocidos. Se pueden estimar mediante procedimientos econométricos estándares sobre ecuaciones dinámicas simples.

Probar una raíz unitaria se ha vuelto muy común en series de tiempo de econometría. Si una serie tiene una raíz unitaria, entonces, en numerosos casos, las aproximaciones normales de muestras grandes dejarán de ser válidas. Además, un proceso de raíz unitaria tiene la propiedad de que una innovación tiene un efecto de larga duración, lo cual es un tema de interés en sí mismo. Si bien hay muchas pruebas para raíces unitarias, la prueba *t* de Dickey-Fuller, y su extensión, la prueba de Dickey-Fuller aumentada, quizá sea la más conocida y más fácil de implementar. Se puede permitir una tendencia lineal cuando se prueban raíces unitarias al agregar una tendencia a la regresión de Dickey-Fuller.

Cuando se efectúa una regresión de una serie I(1), y_t, sobre otra serie I(1), x_t, existe una seria preocupación acerca de la regresión espuria, inclusive si las series no contienen tendencias obvias. Esto se ha estudiado a detalle en el caso de una caminata aleatoria: aun si dos caminatas aleatorias son independientes, la prueba *t* habitual para la significancia del coeficiente de la pendiente, con base en los valores críticos usuales, rechazará con mayor evidencia que el tamaño nominal de la prueba. Además, la R^2 tiende a una variable aleatoria, más que a cero (que sería el caso si se realizara la regresión de la diferencia en y_t sobre la diferencia en x_t).

En un caso importante, una regresión que implica variables I(1) no es espuria, y lo es cuando las series están cointegradas. Esto significa que una función lineal de las dos variables I(1) es I(0). Si y_t y x_t son I(1) pero $y_t - x_t$ es I(0), y_t y x_t no pueden comportarse arbitrariamente para alejarse. Existen formas simples de probar la hipótesis nula de no cointegración frente a la alternativa de cointegración, una de las cuales está basada en la prueba de raíz unitaria de Dickey-Fuller a los residuales de una regresión estática. Existen también estimadores simples del parámetro de cointegración que produce estadísticos *t* con distribuciones normales estándar aproximadas (e intervalos de confianza asintóticamente válidos). Se analizó el estimador de adelantos y rezagos en la sección 18.4.

La cointegración entre y_t y x_t implica que los términos de corrección de errores pueden aparecer en el modelo que relaciona a Δy_t con Δx_t; los términos de corrección de errores son rezagos en $y_t - \beta x_t$, donde β es el parámetro de cointegración. Un procedimiento de estimación simple de dos pasos está disponible para estimar modelos de corrección de errores. Primero, β se estima mediante una regresión estática (o la regresión de adelantos y rezagos). A continuación se emplea MCO para estimar un modelo dinámico simple en primeras diferencias que incluye los términos de corrección de errores.

La sección 18.5 contenía una introducción al pronóstico, con énfasis en los métodos de pronóstico basados en la regresión. Los modelos estáticos o, en términos más generales, los modelos que contienen variables explicativas contemporáneas a la variable dependiente, están limitados, pues se deben pronosticar las variables explicativas. Si se insertan los valores hipotéticos de variables explicativas futuras, se obtiene un pronóstico condicional. Los pronósticos incondicionales son similares a simplemente obtener un modelo y_t en función de la información *pasada* que se ha observado en el tiempo en que se necesitó el pronóstico. Los modelos de regresión dinámica, incluso las autorregresiones y vectores autorregresivos, se emplean de manera habitual. Además de obtener pronósticos puntuales de un paso hacia delante, se analizó también la construcción de intervalos de pronósticos, que son muy similares a los intervalos de predicción.

Se utilizaron varios criterios para elegir entre los diferentes métodos de pronóstico. Las medidas de desempeño más comunes son la raíz del error cuadrático medio y el error absoluto medio. Ambas estiman el tamaño del error de pronóstico promedio. Estas medidas contienen mayor información cuando se emplean pronósticos fuera de la muestra.

Los pronósticos de múltiples pasos hacia delante presentan nuevos desafíos y están sujetos a grandes varianzas en los errores de pronóstico. Sin embargo, para modelos como autorregresiones y vectores autorregresivos, se pueden calcular pronósticos de múltiples pasos hacia delante, y se pueden obtener intervalos de pronóstico aproximados.

Pronosticar la tendencia y series I(1) requiere cuidados especiales. Los procesos con tendencias deterministas se pueden pronosticar incluyendo tendencias temporales en los modelos de regresión, tal vez con rezagos de variables. Una posible desventaja es que las tendencias deterministas pueden ofrecer pronósticos pobres para pronósticos de horizontes largos: una vez que se estima, una tendencia lineal continúa aumentando o disminuyendo. El enfoque típico para pronosticar un proceso I(1) es pronosticar la diferencia en el proceso y agregar el nivel de la variable al de la diferencia pronosticada. Por otra parte, los modelos de vector autorregresivo se pueden utilizar en los niveles de las series. Si las series están cointegradas, es posible emplear en su lugar los modelos de corrección de errores.

TÉRMINOS CLAVE

Causalidad de Granger
Cointegración
Conjunto de información
Criterios dentro de la muestra
Criterios fuera de la muestra
Distribución de Dickey-Fuller
Error absoluto medio (EAM)
Error de pronóstico
Estimador de adelantos y rezagos
Función de pérdida
Intervalo de pronóstico
Martingala
Modelo de corrección de errores
Modelo de rezagos distribuidos infinitos (RDI)
Modelo de rezagos distribuidos racionales (RDR)
Modelo de vector autorregresivo (VAR)
Problema de regresión espuria
Procedimiento de dos pasos de Engle-Granger
Pronóstico condicional
Pronóstico de múltiples pasos hacia delante
Pronóstico de un paso hacia delante
Pronóstico incondicional
Pronóstico puntual
Prueba de Dickey-Fuller (DF)
Prueba de Dickey-Fuller aumentada
Prueba de Engle-Granger
Raíces unitarias
Raíz del error cuadrático medio (RECM)
Rezago distribuido geométrico (o Koyck)
Secuencia de una martingala en diferencia (SMD)
Suavizamiento exponencial

PROBLEMAS

18.1 Considere la ecuación (18.15) con $k = 2$. Mediante el enfoque VI para estimar γ_h y ρ, ¿qué usaría como instrumento para y_{t-1}?

18.2 Un modelo económico interesante que lleva a un modelo econométrico con una variable dependiente rezagada relaciona y_t con el *valor esperado* de x_t, sea, x_t^*, donde la expectativa está basada en toda la información observada en el tiempo $t - 1$:

$$y_t = \alpha_0 + \alpha_1 x_t^* + u_t. \qquad \boxed{18.68}$$

Un supuesto natural sobre $\{u_t\}$ es que $E(u_t|I_{t-1}) = 0$, donde I_{t-1} denota toda la información sobre y y x observadas en el tiempo $t - 1$; esto significa que $E(y_t|I_{t-1}) = \alpha_0 + \alpha_1 x_t^*$. Para completar este modelo, es necesario un supuesto acerca de cómo se formó la expectativa x_t^*.

Se vio un ejemplo simple de expectativas adaptativas en la sección 11.2, donde $x_t^* = x_{t-1}$. Un esquema más complicado de expectativas adaptativas es

$$x_t^* - x_{t-1}^* = \lambda(x_{t-1} - x_{t-1}^*), \qquad \boxed{18.69}$$

donde $0 < \lambda < 1$. Esta ecuación implica que el cambio en las expectativas reacciona si el valor realizado del último periodo estuvo por encima o por debajo de su expectativa. El supuesto $0 < \lambda < 1$ implica que el cambio en las expectativas es una fracción del error del último periodo.

i) Muestre que las dos ecuaciones implican que

$$y_t = \lambda\alpha_0 + (1 - \lambda)y_{t-1} + \lambda\alpha_1 x_{t-1} + u_t - (1 - \lambda)u_{t-1}.$$

[*Sugerencia:* rezague un periodo la ecuación (18.68), multiplique por $(1 - \lambda)$ y reste esto de (18.68). A continuación, utilice (18.69).]

ii) De acuerdo con $\mathrm{E}(u_t|I_{t-1}) = 0$, $\{u_t\}$ no está correlacionada serialmente. ¿Qué implica esto acerca de los nuevos errores, $v_t = u_t - (1 - \lambda)u_{t-1}$?

iii) Si se escribe la ecuación del inciso i) como

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t,$$

¿Cómo estimaría usted consistentemente β_j?

iv) Dados los estimadores consistentes de β_j, ¿cómo estimaría usted consistentemente λ y α_1?

18.3 Suponga que $\{y_t\}$ y $\{z_t\}$ son series I(1), pero $y_t - \beta z_t$ es I(0) para alguna $\beta \neq 0$. Muestre que para toda $\delta \neq \beta$, $y_t - \delta z_t$ debe ser I(1).

18.4 Considere el modelo de corrección de errores en la ecuación (18.37). Muestre que si usted añade otro rezago del término de corrección de errores, $y_{t-2} - \beta x_{t-2}$, la ecuación sufre de colinealidad perfecta. (*Sugerencia:* muestre que $y_{t-2} - \beta x_{t-2}$ es una función de colinealidad perfecta de $y_{t-1} - \beta x_{t-1}$, Δx_{t-1}, y Δy_{t-1}.)

18.5 Suponga que el proceso $\{(x_t, y_t): t = 0, 1, 2, \dots\}$ satisface las ecuaciones

$$y_t = \beta x_t + u_t$$

y

$$\Delta x_t = \gamma \Delta x_{t-1} + v_t,$$

donde $\mathrm{E}(u_t|I_{t-1}) = \mathrm{E}(v_t|I_{t-1}) = 0$, I_{t-1} contiene información sobre x y y en el tiempo $t - 1$ y antes, $\beta \neq 0$ y $|\gamma| < 1$ [de manera que x_t, y por tanto y_t, es I(1)]. Muestre que estas dos ecuaciones implican un modelo de corrección de errores de la forma

$$\Delta y_t = \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + e_t,$$

donde $\gamma_1 = \beta\gamma$, $\delta = -1$ y $e_t = u_t + \beta v_t$. (*Sugerencia:* primero, reste y_{t-1} de ambos lados de la primera ecuación. Después, sume y reste βx_{t-1} del lado derecho y reordene. Por último, emplee la segunda ecuación para obtener el modelo de corrección de errores que contiene Δx_{t-1}.)

18.6 El siguiente modelo se estimó con los datos mensuales de VOLAT.RAW:

$$\begin{aligned}\widehat{pcip} = {} & \underset{(.56)}{1.54} + \underset{(.042)}{.344}\, pcip_{-1} + \underset{(.045)}{.074}\, pcip_{-2} + \underset{(.042)}{.073}\, pcip_{-3} + \underset{(.013)}{.031}\, pcsp_{-1}\\ n = {} & 554,\ R^2 = .174,\ \bar{R}^2 = .168,\end{aligned}$$

donde *pcip* es el cambio porcentual en la producción mensual industrial, a una tasa anualizada y *pcsp* es el cambio porcentual en el índice 500 de Standard & Poor, también a una tasa anualizada.

i) Si los pasados tres meses de *pcip* son cero y $pcsp_{-1} = 0$, ¿cuál es la predicción de crecimiento en la producción industrial para este mes? ¿Es estadísticamente diferente de cero?

ii) Si los pasados tres meses de *pcip* son cero pero $pcsp_{-1} = 10$, ¿cuál es la predicción de crecimiento en la producción industrial?

iii) ¿Qué concluye usted acerca de los efectos del mercado accionario sobre la actividad económica real?

18.7 Sea gM_t el crecimiento anual en la oferta de dinero y $unem_t$ la tasa de desempleo. En el caso de que $unem_t$ siga un proceso AR(1) estable, explique a detalle cómo probaría que gM sea causa Granger de *unem*.

18.8 Suponga que y_t sigue el modelo

$$y_t = \alpha + \delta_1 z_{t-1} + u_t$$
$$u_t = \rho u_{t-1} + e_t$$
$$\mathrm{E}(e_t|I_{t-1}) = 0,$$

donde I_{t-1} contiene y y z en el tiempo $t - 1$ y antes.

i) Muestre que $\mathrm{E}(y_{t+1}|I_t) = (1 - \rho)\alpha + \rho y_t + \delta_1 z_t - \rho\delta_1 z_{t-1}$. (*Sugerencia*: escriba $u_{t-1} = y_{t-1} - \alpha - \delta_1 z_{t-2}$ e inserte esto en la segunda ecuación; después, inserte el resultado en la primera ecuación y tome la esperanza condicional.)

ii) Suponga que utiliza n observaciones para estimar α, δ_1 y ρ. Escriba la ecuación para pronosticar y_{n+1}.

iii) Explique por qué el modelo con un rezago de z y una correlación serial AR(1) es un caso especial del modelo

$$y_t = \alpha_0 + \rho y_{t-1} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + e_t.$$

iv) ¿Qué sugiere el inciso iii) acerca de utilizar modelos con correlación serial AR(1) para el pronóstico?

18.9 Sea $\{y_t\}$ una secuencia I(1). Suponga que $\hat{g}_n$ es un pronóstico de un paso hacia delante de Δy_{n+1} y sea $\hat{f}_n = \hat{g}_n + y_n$ un pronóstico de un paso hacia delante de y_{n+1}. Explique por qué los errores de pronóstico para pronosticar Δy_{n+1} y y_{n+1} son idénticos.

EJERCICIOS EN COMPUTADORA

C18.1 Utilice los datos en WAGEPRC.RAW para este ejercicio. El problema 11.5 da estimaciones de un modelo de rezagos distribuidos distintos de *gprice* sobre *gwage*, donde se emplearon 12 rezagos de *gwage*.

i) Estime un modelo RD geométrico simple de *gprice* sobre *gwage*. En particular, estime la ecuación (18.11) por MCO. ¿Cuáles son el multiplicador de impacto estimado y PLP? Grafique la distribución estimada de los rezagos.

ii) Compare los estimadores de PI y PLP con los obtenidos en el problema 11.5. ¿Cómo se comparan las distribuciones estimadas rezagadas?

iii) Ahora estime el modelo de rezagos distribuidos racionales de (18.16). Diagrame la distribución de rezagos y compare el estimador de PI y PLP con los obtenidos en el inciso ii).

C18.2 Para este ejercicio utilice los datos en HSEINV.RAW.

i) Pruebe una raíz unitaria en log(*invpc*), incluida la tendencia temporal lineal y dos rezagos de $\Delta\log(invpc_t)$. Utilice un nivel de significancia de 5%.

ii) Emplee el método del inciso i) para probar una raíz unitaria en log(*price*).

iii) Dados los resultados en los incisos i) y ii), ¿tiene sentido probar la cointegración entre log(*invpc*) y log(*price*)?

C18.3 Utilice los datos en VOLAT.RAW para este ejercicio.

i) Estime un modelo AR(3) para *pcip*. Ahora, agregue un cuarto rezago y verifique que sea muy insignificante.

ii) Al modelo AR(3) del inciso i), agréguele tres rezagos de *pcsp* para probar si *pcsp* es causa Granger de *pcip*. Exprese con cuidado su conclusión.

iii) Al modelo del inciso ii), agréguele tres rezagos del cambio en *i3*, la tasa de certificados del Tesoro a tres meses. ¿Es *pcsp* causa Granger de *pcip* condicionada sobre $\Delta i3$ pasada?

C18.4 Al probar la cointegración entre *gfr* y *pe* en el ejemplo 18.5, agregue t^2 a la ecuación (18.32) para obtener los residuales MCO. Incluya un rezago en la prueba DF aumentada. El valor crítico de 5% para la prueba es -4.15.

C18.5 Utilice INTQRT.RAW para este ejercicio.

i) En el ejemplo 18.7, se estimó un modelo de corrección de errores para los rendimientos de certificados del Tesoro a seis meses, donde un rezago en el rendimiento sobre certificados del Tesoro a tres meses es la variable explicativa. Se supone que el parámetro de cointegración fue uno en la ecuación $hy6_t = \alpha + \beta hy3_{t-1} + u_t$. Ahora, agregue la diferencia adelantada, $\Delta hy3_t$, la variación contemporánea, $\Delta hy3_{t-1}$ y la diferencia rezagada, $\Delta hy3_{t-2}$, de $hy3_{t-1}$. Es decir, estime la ecuación

$$hy6_t = \alpha + \beta hy3_{t-1} + \phi_0\Delta hy3_t + \phi_1\Delta hy3_{t-1} + \rho_1\Delta hy3_{t-2} + e_t$$

y reporte los resultados en forma de ecuación. Pruebe H_0: $\beta = 1$ frente a la alternativa de dos colas. Suponga que el adelanto y el rezago son suficientes de manera que $\{hy3_{t-1}\}$ sea estrictamente exógeno en esta ecuación y haga caso omiso de la correlación serial.

ii) Al modelo de corrección de errores en (18.39), agregue $\Delta hy3_{t-2}$ y $(hy6_{t-2} - hy3_{t-3})$. ¿Estos términos en conjunto son significativos? ¿Qué concluye acerca del modelo de corrección de errores apropiado?

C18.6 Utilice los datos en PHILLIPS.RAW para contestar estas preguntas.

i) Estime los modelos en (18.48) y (18.49) mediante los datos hasta 1997. ¿Los estimadores de parámetros cambian mucho en comparación con (18.48) y (18.49)?

ii) Emplee las nuevas ecuaciones para pronosticar $unem_{1998}$; redondee a dos decimales. ¿Qué ecuación produce un mejor pronóstico?

iii) Como se analizó en el texto, el pronóstico para $unem_{1998}$ utilizando (18.49) es 4.90. Compare esto con el pronóstico obtenido empleando los datos hasta 1997. ¿Al utilizar el año adicional de datos para obtener el estimador del parámetro produce un mejor pronóstico?

iv) Emplee el modelo estimado en (18.48) para obtener un pronóstico de dos pasos hacia delante. Es decir, pronostique $unem_{1998}$ mediante la ecuación (18.55) con $\hat{\alpha} = 1.572$,

$\hat{\rho} = .732$ y $h = 2$. ¿Esto es mejor o peor que un pronóstico de un paso hacia delante obtenido al insertar $unem_{1997} = 4.9$ en (18.48)?

C18.7 Utilice los datos en BARIUM.RAW para este ejercicio.

i) Estime el modelo de tendencia lineal $chnimp_t = \alpha + \beta t + u_t$, mediante las primeras 119 observaciones (esto excluye los últimos 12 meses de observaciones para 1988). ¿Cuál es el error estándar de la regresión?

ii) Ahora estime un modelo AR(1) para *chnimp*, nuevamente emplee todos los datos menos los últimos 12 meses. Compare el error estándar de la regresión con el del inciso i). ¿Qué modelo ofrece un mejor ajuste dentro de la muestra?

iii) Utilice los modelos de los incisos i) y ii) para calcular los errores de pronóstico de un paso hacia delante para los 12 meses en 1988. (Debe obtener 12 errores de pronóstico para cada método.) Calcule y compare los RECM y los EAM para los dos métodos. ¿Qué método de pronóstico funciona mejor fuera de la muestra para pronósticos de un paso hacia adelante?

iv) Agregue variables binarias mensuales a la regresión del inciso i). ¿Son conjuntamente significativas? (No se preocupe por la ligera correlación serial en los errores de esta regresión al hacer la prueba conjunta.)

C18.8 Utilice los datos en FERTIL3.RAW para este ejercicio.

i) Grafique *gfr* en función del tiempo. ¿Contienen una tendencia clara ascendente o descendente durante todo el periodo de la muestra?

ii) Mediante los datos hasta 1979, estime el modelo de tendencia temporal cúbica para *gfr* (es decir, ejecute una regresión de *gfr* sobre t, t^2 y t^3, junto con un intercepto). Comente sobre la R-cuadrada de la regresión.

iii) Mediante el modelo de la parte ii), calcule el error absoluto medio de los errores de pronóstico de un paso hacia delante para los años de 1980 a 1984.

iv) Mediante los datos hasta 1979, haga una regresión de Δgfr_t sólo sobre una constante. ¿La constante es estadísticamente diferente de cero? ¿Tiene sentido suponer que cualquier término de tendencia (estocástica) es cero, si se ha supuesto que gfr_t sigue una caminata aleatoria?

v) Ahora, pronostique *gfr* para 1980 a 1984, empleando un modelo de caminata aleatoria: el pronóstico de gfr_{n+1} es simplemente gfr_n. Calcule el EAM. ¿Cómo se compara éste con el del inciso iii)? ¿Qué método de pronóstico prefiere?

vi) Ahora, estime un modelo AR(2) para *gfr*, de nuevo, utilice los datos sólo hasta 1979. ¿El segundo rezago es significativo?

vii) Obtenga el EAM de 1980 a 1984, mediante el modelo AR(2). ¿Este modelo más general funciona mejor fuera de la muestra que el modelo de caminata aleatoria?

C18.9 Utilice CONSUMP.RAW para este ejercicio.

i) Sea y_t el ingreso disponible real *per cápita*. Estime el modelo con los datos hasta 1989

$$y_t = \alpha + \beta t + \rho y_{t-1} + u_t$$

y reporte los resultados de la forma habitual.

ii) Emplee la ecuación estimada del inciso i) para pronosticar y en 1990. ¿Cuál es el error de pronóstico?

iii) Calcule el error absoluto medio del pronóstico de un paso hacia delante para la década de los noventa, mediante los parámetros estimados en el inciso i).

iv) Ahora, calcule el EAM para el mismo periodo, pero descarte y_{t-1} de la ecuación. ¿Es mejor o no incluir y_{t-1} en el modelo?

C18.10 Emplee los datos en INTQRT.RAW para este ejercicio.

i) Mediante los datos de al menos los últimos cuatro años (16 trimestres), estime un modelo AR(1) para $\Delta r6_t$. (Se utiliza la diferencia debido a que parece que $r6_t$ tiene una raíz unitaria.) Determine la RECM de los pronósticos de un paso hacia delante para $\Delta r6$, mediante los últimos 16 trimestres.

ii) Ahora, sume el término de corrección de error $spr_{t-1} = r6_{t-1} - r3_{t-1}$ a la ecuación del inciso i). (Esto supone que el parámetro de cointegración es uno.) Calcule la RECM para los últimos 16 trimestres. ¿El término de corrección de error ayuda con el pronóstico fuera de la muestra en este caso?

iii) Ahora, estime el parámetro de cointegración, en lugar de fijarlo en uno. Emplee los últimos 16 trimestres de nuevo, para producir la RECM fuera de la muestra. ¿Cómo se compara esto con los pronósticos de los incisos i) y ii)?

iv) ¿Sus conclusiones cambiarían si quisiera predecir $r6$ en lugar de $\Delta r6$? Explique.

C18.11 Utilice los datos en VOLAT.RAW para este ejercicio.

i) Confirme que $lsp500 = \log(sp500)$ y $lip = \log(ip)$ parecen contener raíces unitarias. Utilice las pruebas de Dickey-Fuller con cuatro cambios rezagados y haga las pruebas con y sin la tendencia temporal lineal.

ii) Realice una regresión simple de $lsp500$ sobre lip. Comente sobre las dimensiones del estadístico t y R-cuadrada.

iii) Utilice los residuales del inciso ii) para probar si $lsp500$ y lip están cointegradas. Emplee la prueba estándar de Dickey-Fuller y la prueba de DF aumentada con dos rezagos. ¿Qué concluye?

iv) Agregue una tendencia temporal lineal a la regresión del inciso ii) y ahora pruebe la cointegración utilizando las mismas pruebas del inciso iii).

v) ¿Parece que los precios de las acciones y la actividad económica real tienen una relación de equilibrio a largo plazo?

C18.12 Este ejercicio también utiliza datos de VOLAT.RAW. El ejercicio para computadora C18.11 estudia la relación a largo plazo entre los precios de las acciones y la producción industrial. Aquí, usted estudiará la pregunta de la causalidad de Granger mediante los cambios porcentuales.

i) Estime un modelo AR(3) para $pcip_t$, el cambio porcentual en la producción industrial (reportado a una tasa anualizada). Muestre que el segundo y tercer rezagos son conjuntamente significantes al nivel de 2.5%.

ii) Agregue un rezago de $pcsp_t$ a la ecuación estimada en el inciso i). ¿El rezago es estadísticamente significante? ¿Qué le dice esto acerca de la causalidad de Granger entre el crecimiento en la producción industrial y el crecimiento en los precios de las acciones?

iii) Reelabore el inciso ii) pero obtenga el estadístico t robusto a heterocedasticidad. ¿La prueba robusta cambió sus conclusiones del inciso ii)?

C18.13 Utilice los datos en TRAFFIC2.RAW para este ejercicio. Estos datos mensuales, sobre los accidentes de tránsito en California entre los años 1981 y 1989, se utilizaron en el ejercicio para computadora C10.11.

i) Mediante la regresión estándar de Dickey-Fuller, pruebe si $ltotacc_t$ tiene una raíz unitaria. ¿Puede rechazar una raíz unitaria al nivel de 2.5%?

ii) Ahora agregue dos cambios rezagados a la prueba del inciso i) y calcule la prueba de Dickey-Fuller aumentada. ¿Qué concluye?

iii) Agregue una tendencia temporal lineal a la regresión DF aumentada del inciso ii). Ahora ¿qué sucede?

iv) Dados los resultados del inciso i) a iii), ¿qué diría que es la mejor caracterización de $ltotacc_t$: un proceso I(1) o un proceso I(0) acerca de una tendencia temporal lineal?

v) Pruebe el porcentaje de muertes, $prcfat_t$, para una raíz unitaria, mediante dos rezagos en una regresión ADF. En este caso, ¿importa si usted incluye una tendencia temporal lineal?

C18.14 Utilice los datos en MINWAGE.DTA para el sector 232 a fin de responder las siguientes preguntas.

i) Confirme que $lwage232_t$ y $lemp232_t$ se caracterizan mejor como procesos I(1). Utilice la prueba DF aumentada con un rezago de $gwage232$ y $gemp232$, respectivamente, y una tendencia temporal lineal. ¿Tiene duda de que se deba suponer que estas series tienen raíces unitarias?

ii) Realice una regresión de $lemp232_t$ sobre $lwage232_t$ y pruebe la cointegración, con y sin tendencia temporal, y permita dos rezagos en la prueba de Engle-Granger aumentada. ¿Qué concluye?

iii) Ahora realice una regresión de $lemp232_t$ sobre el logaritmo de la tasa de salario real $lrwage232_t = lwage232_t - lcpi_t$, y una tendencia temporal. ¿Encuentra cointegración? ¿Estarán "más cerca" de cointegrarse cuando utilice salarios reales en lugar de salarios nominales?

iv) ¿Cuáles son algunos factores que pueden faltar en la regresión de cointegración del inciso iii)?

CAPÍTULO 19

Realización de un proyecto empírico

En este capítulo se analizan los ingredientes de un análisis empírico exitoso, con énfasis en la realización de un proyecto de equipo. Además de repasar las cuestiones importantes que han surgido en el libro, se hace énfasis en los temas recurrentes que son fundamentales para la investigación aplicada. También se ofrecen sugerencias de temas como una forma de estimular la imaginación. Se proporcionan varias fuentes de investigación y datos económicos.

19.1 Plantear una pregunta

Es crucial plantear una pregunta muy específica. Sin que la meta del análisis que se está realizando esté claramente definida, no se puede saber por dónde comenzar. La difusión de conjuntos de datos ricos puede inducir al investigador a lanzarse en un conjunto de datos con base en ideas mal concebidas, lo cual resulta contraproducente. Es probable que, sin formular con cuidado las hipótesis y el tipo de modelo que se necesitará estimar, se olvide recabar información sobre variables importantes, obtener una muestra de la población equivocada o recabar datos del periodo equivocado.

Esto no significa que se deba plantear la pregunta en un vacío. En especial para un proyecto de un solo plazo, no se puede ser demasiado ambicioso. Por tanto, cuando se elija un tema, se debe estar razonablemente seguro de que existen fuentes de datos que permitirán responder la pregunta en el tiempo asignado.

Al elegir un tema, se necesita decidir qué áreas de la economía o de otras ciencias sociales son de interés. Por ejemplo, si se ha tomado un curso en economía laboral, quizá se hayan visto teorías que pueden probarse empíricamente o relaciones que tengan alguna relevancia política. Los economistas laborales constantemente están ideando nuevas variables que puedan explicar los diferenciales salariales. Algunos ejemplos incluyen la calidad del bachillerato [Card y Krueger (1992) y Betts (1995)], la cantidad de matemáticas y ciencias que se cursaron en el bachillerato [Levine y Zimmerman (1995)] y la apariencia física [Hamermesh y Biddle (1994), Averett y Korenman (1996) y Biddle y Hamermesh (998)]. Los investigadores de finanzas públicas estatales y locales estudian cómo depende la actividad económica local de las variables de las políticas económicas, como impuestos sobre la propiedad, impuestos sobre las ventas, nivel y calidad de servicios (como escuelas, bomberos y policía): etc. [Vea, por ejemplo, White (1986), Papke (1987), Bartik (1991) y Netzer (1992).]

Los economistas que estudian temas de educación están interesados en determinar cómo el gasto afecta el desempeño [Hanushek (1986)], si asistir a cierto tipo de escuelas mejora el desempeño [por ejemplo, Evans y Schwab (1995)], y qué factores afectan dónde eligen las escuelas primarias ubicarse [Downes y Greenstein (1996)].

Los macroeconomistas están interesados en las relaciones entre varios agregados de series de tiempo, como el vínculo entre el crecimiento en el producto interno bruto y el crecimiento en la inversión fija o maquinaria [vea De Long y Summers (1991)] o el efecto de los impuestos sobre las tasas de interés [por ejemplo, Peek (1982)].

Sin lugar a dudas, existen razones para estimar modelos que en su mayoría son descriptivos. Por ejemplo, los asesores de impuestos sobre la propiedad utilizan modelos (llamados *modelos de precios hedónicos*) para estimar el valor del alojamiento de las casas que no se han vendido recientemente. Esto implica un modelo de regresión que relacione el precio de una casa con sus características (tamaño, número de recámaras, número de baños, etc.). Como tema para una investigación no es muy emocionante: puesto que tal análisis no tiene implicaciones evidentes sobre las políticas. Agregar la tasa de delitos en el vecindario como variable explicativa podría permitir determinar qué importante es el factor de la criminalidad en los precios de las viviendas, algo que sería útil para estimar los costos de los delitos.

Se han estimado varias relaciones mediante los datos macroeconómicos que son más descriptivos. Por ejemplo, se puede utilizar una función del ahorro agregado para estimar la propensión agregada marginal para ahorrar, así como la respuesta de ahorrar para obtener rendimientos sobre los activos (como tasas de interés). Tal análisis podría ser más interesante utilizando datos de series de tiempo en un país que tiene una historia de perturbaciones políticas y determinar si las tasas de ahorro disminuyen durante épocas de incertidumbre política.

Una vez que se decida un área de investigación, existen muchas maneras de encontrar datos específicos sobre el tema. La revista *Journal of Economic Literature* (JEL) tiene un sistema de clasificación detallado en el cual se le da a cada ensayo un conjunto de códigos de identificación que lo coloca dentro de ciertos subtemas de economía. La JEL también contiene una lista de artículos publicados en una amplia variedad de revistas, organizados por tema e incluso contiene algunos resúmenes breves de algunos artículos.

Los servicios de **Internet** son especialmente convenientes para hallar ensayos publicados sobre varios temas, como *EconLit*, al que muchas universidades están suscritas. *EconLit* permite a los usuarios hacer una búsqueda integral de casi todos los temas económicos por autor, tema, palabras en el título, etc. El *Social Sciences Citation Index* es útil para encontrar ensayos sobre una amplia variedad de temas en ciencias sociales, incluidos ensayos populares que se citan a menudo en otros trabajos publicados.

Google Scholar es un motor de búsqueda en Internet que puede ser muy útil para rastrear investigaciones sobre varios temas o investigar mediante un autor particular. Sobre todo para temas que no se han publicado en una revista académica o que están por publicarse.

Cuando piense en un tema, debe tener en mente algunas cosas. Primero, para que una pregunta sea interesante, no es necesario tener implicaciones políticas muy generales; sino que debe ser de interés local. Por ejemplo, quizá se esté interesado en saber si vivir en una fraternidad de la universidad ocasiona que los estudiantes tengan calificaciones superiores o menores al promedio. Esto puede, o no, interesar a las personas fuera de su universidad, pero probablemente interese a algunas personas dentro de ella. Por otra parte, puede estudiar un problema que comience siendo un problema local, pero que después se convierta en un tema de interés general, como determinar qué factores afectan y qué políticas universitarias pueden originar el abuso del alcohol en los campus universitarios.

Segundo, es muy difícil, en especial para un proyecto trimestral o semestral, hacer una investigación verdaderamente original usando los agregados macroeconómicos estándares de la economía estadounidense. Por ejemplo, la pregunta de si el crecimiento monetario, el crecimiento en el gasto gubernamental, etc., afectan el crecimiento económico ha sido y continúa siendo estudiado por los profesionales de la macroeconomía. La pregunta de si los rendimientos sobre las acciones u otros activos pueden predecirse sistemáticamente utilizando información conocida, por obvias razones, se ha estudiado con mucho cuidado. Esto no significa que se deba evitar estimar los modelos financieros macroeconómicos o empíricos, con la creencia de que con sólo utilizar datos más recientes se podrá complementar una argumentación. Además, en ocasiones se puede encontrar una nueva variable que tenga un efecto importante sobre los agregados económicos o los rendimientos financieros; tal descubrimiento puede ser muy emocionante.

El punto es que, ejercicios como utilizar algunos años adicionales para estimar una curva de Phillips estándar o una función de consumo agregado para la economía estadounidense, o alguna otra economía grande, tiene pocas probabilidades de producir una comprensión más profunda, aunque pueden ser instructivas para el estudiante. De hecho, se pueden emplear datos sobre un país pequeño para estimar la curva estática o dinámica de Phillips, o para probar la hipótesis de los mercados eficientes, etcétera.

En el nivel no macroeconómico, existe también una infinidad de cuestiones que se han estudiado ampliamente. Por ejemplo, los economistas laborales han publicado muchos trabajos relacionados con la estimación del rendimiento sobre la educación. Esta pregunta sigue estudiándose por su importancia, y nuevos conjuntos de datos, y nuevos enfoques econométricos continúan desarrollándose. Por ejemplo, como se vio en el capítulo 9, ciertos conjuntos de datos tienen mejores variables proxy para la capacidad inobservable que otros. (Compare WAGE1.RAW y WAGE2.RAW.) En otros casos, se pueden obtener paneles de datos o datos de un experimento natural (vea el capítulo 13) lo cual permite una aproximación a un tema antiguo desde una perspectiva distinta.

Por ejemplo, los penalistas están interesados en estudiar los efectos de diversas legislaciones sobre el crimen. La pregunta sobre si la pena capital tiene un efecto disuasivo ha estado en debate durante mucho tiempo. Asimismo, los economistas han estado interesados en si los impuestos a los cigarros y el alcohol reducen su consumo (como siempre, en un sentido *ceteris paribus*). A medida que se tiene acceso a más datos a nivel estatal, se puede crear un panel de datos más sustancial y esto puede ayudar a mejorar las respuestas a las principales interrogantes políticas. Además, la efectividad de las recientes innovaciones en contra de la delincuencia, como vigilancia vecinal, se puede evaluar de manera empírica.

Cuando se formule la pregunta, será útil analizar las ideas con compañeros de clase, profesor y amigos. Se debe ser capaz de convencer a las personas de que dar respuesta a la pregunta entraña algún interés. (Por supuesto, el que se pueda contestar persuasivamente la pregunta es otra cuestión, lo cierto aquí es que se necesita comenzar con una pregunta interesante.) Si alguien pregunta acerca de la investigación y se responde con un "Es una investigación sobre la delincuencia" o "El trabajo es sobre las tasas de interés", las probabilidades de que su trabajo sea muy general, carente de un verdadero planteamiento, son altas. Debe poder decirse algo como "Es un estudio sobre los efectos de la vigilancia vecinal sobre las tasas de delitos en Estados Unidos" o "Se estudia cómo afecta la volatilidad inflacionaria las tasas de interés de corto plazo en Brasil."

19.2 Revisión bibliográfica

Toda investigación, incluso las que son relativamente cortas, deben contar con una revisión de la bibliografía relevante. Es poco frecuente que uno intente realizar un proyecto empírico para el cual no existen precedentes publicados. Si la investigación se basa en revistas o en **servicios de búsqueda en línea** como *EconLit* para elegir un tema, se está por buen camino para una revisión bibliográfica. Si se elije un tema por propia cuenta, como el estudio de los efectos del consumo de drogas en el rendimiento académico en la universidad, entonces probablemente se requiera un mayor esfuerzo. Pero los servicios de investigación en línea hacen que ese trabajo sea mucho más fácil, puesto que se puede buscar por palabras clave, por palabras en los títulos, por autor, etc. Después, es posible leer los resúmenes de los trabajos para saber qué tan relevantes son para la investigación.

Cuando se elabore investigación bibliográfica, se deben tener en mente temas relacionados que quizá no aparezcan en una búsqueda mediante unas cuantas palabras clave. Por ejemplo, si se están estudiando los efectos del consumo de drogas en los salarios o en el rendimiento académico, probablemente se deba buscar bibliografía acerca de cómo el consumo del alcohol afecta tales factores. Saber cómo realizar una búsqueda bibliográfica detallada es una habilidad adquirida, pero pueden ahorrarse muchos problemas si se reflexiona antes de buscar.

Los investigadores difieren en cuanto a cómo se debe incorporar la revisión bibliográfica en un trabajo. Algunos prefieren tener una sección separada llamada "revisión bibliográfica", mientras a otros les gusta incluir la revisión bibliográfica como parte de la introducción. Esto es en gran medida cuestión de gustos, sin embargo, una revisión bibliográfica extensa probablemente merezca su propia sección. Si el trabajo final constituye la parte central del curso, por ejemplo, de un seminario de último año de la carrera o de un curso de econometría avanzada, la revisión bibliográfica quizá sea larga. Los trabajos de final de curso de los primeros años de la carrera, por lo general son más cortos y las revisiones bibliográficas más breves.

19.3 Recolección de datos

Decidir el conjunto apropiado de datos

Recabar datos para un trabajo de final de curso puede ser educativo, emocionante y, en ocasiones, hasta frustrante. Primero debe decidir el tipo de datos necesario para responder a su pregunta. Como se analizó en la introducción y se cubre a lo largo de este libro, los conjuntos de datos tienen una variedad de formas. Los tipos más comunes son los conjuntos de datos de corte transversal, series de tiempo, cortes transversales combinados y datos de panel.

Se puede dar respuesta a muchas preguntas mediante cualquiera de las estructuras de datos que se han descrito. Por ejemplo, para estudiar si una imposición de leyes más severa reduce la delincuencia, se podría utilizar un corte transversal de ciudades, una serie de tiempo para una ciudad determinada o un panel de datos de las ciudades, que consiste en las mismas ciudades durante dos o más años.

Decidir qué tipo de datos recabar suele depender de la naturaleza del análisis. Para responder preguntas a nivel individual o familiar, por lo general, sólo se tiene acceso a un único corte transversal; a menudo, éstos se obtienen a través de encuestas. Después, se debe preguntar si es posible obtener un conjunto de datos lo suficientemente sustancioso para realizar un análisis *ceteris paribus* convincente. Por ejemplo, suponga que se quiere saber si las familias que ahorran a través de cuentas individuales de retiro (IRA, *individual retirement accounts*), que tienen ciertas ventajas fiscales, tienen menores ahorros, diferentes de los IRA. En otras palabras, ¿el ahorro IRA simplemente reduce otras formas de ahorro? Existen conjuntos de datos, como los que publica el estudio *Survey of Consumer Finances*, que contiene información sobre varios tipos de instrumentos de ahorro para una muestra diferente de familias cada año. Varias cuestiones surgen cuando se utiliza tal conjunto de datos. Quizá la más importante sea si existen los suficientes controles (incluido el ingreso, la demografía y proxy para las preferencias de ahorro) para hacer un análisis razonable *ceteris paribus*. Si éstos son los únicos tipos de datos disponibles, se debe hacer lo que se pueda con ellos.

Las mismas cuestiones surgen con datos de corte transversal sobre empresas, ciudades, estados, etc. En la mayoría de los casos, no es evidente que sea posible hacer un análisis *ceteris paribus* con un único corte transversal. Por ejemplo, cualquier estudio de los efectos de una imposición más severa de la ley sobre la delincuencia debe reconocer la endogeneidad de los gastos de la imposición de las leyes. Cuando se utilizan métodos estándar de regresión, puede ser muy difícil completar un análisis *ceteris paribus* convincente, sin importar cuántos controles se tengan. (Vea la sección 19.4 para más detalles.)

Si usted ha leído los capítulos avanzados sobre métodos de datos de panel, sabrá que tener las mismas unidades de corte transversal en dos o más puntos diferentes en el tiempo puede permitir el control de los efectos inobservables constantes en el tiempo que normalmente podrían confundir la regresión sobre una sola sección de corte transversal. Los conjuntos de datos de panel son relativamente difíciles de obtener para individuos o familias, aunque existen algunos importantes, como el *Panel Study of Income Dynamics*, pero se pueden utilizar en formas muy convincentes. También existen conjuntos de datos de panel sobre empresas. Por ejemplo, *Compustat* y el *Center for Research in Security Prices (CRSP)* administran conjuntos de datos de panel muy grandes de información financiera sobre las empresas. Es más fácil obtener los conjuntos de datos de panel sobre unidades más grandes, como escuelas, ciudades, condados o municipios y estados, y éstos no tienden a desaparecer con el tiempo, y las agencias gubernamentales son responsables de recabar la información sobre las mismas variables cada año. Por ejemplo, el *Federal Bureau of Investigation* recaba y reporta información detallada sobre las tasas de delitos a nivel de ciudad. Varias fuentes de datos se listan al final de este capítulo.

Los datos aparecen en diversas formas. Algunos conjuntos de datos, en especial los históricos, están disponibles sólo en forma impresa. Para conjuntos pequeños de datos, introducir uno mismo los datos, a partir de la fuente impresa, es más fácil y cómodo. En ocasiones, los artículos se publican junto con pequeños conjuntos de datos, en especial aplicaciones de series de tiempo. Éstos se pueden utilizar en un estudio empírico, quizá para complementar los datos con información de años más recientes.

Existen muchos conjuntos de datos disponibles en forma electrónica. Varias agencias gubernamentales ofrecen datos en sus sitios web. Las empresas privadas, en ocasiones, compilan datos para hacerlos fáciles de usar, y que después venden por una cuota. Los autores de trabajos de investigación suelen estar dispuestos a proporcionar sus conjuntos de datos en forma electrónica. Cada vez más y más conjuntos de datos están disponibles en Internet. La web es un recurso vasto de **bases de datos en línea**. Se han creado numerosos sitios web que contienen conjuntos de datos económicos y relacionados. Varios otros sitios web contienen vínculos a conjuntos de datos que son de interés para los economistas; algunos de estos se listan al final del capítulo. Por lo general, buscar fuentes de datos en Internet es fácil y lo será más en el futuro.

Ingresar y almacenar los datos

Una vez que se haya decidido un tipo de datos y haya localizado una fuente de datos, debe colocarlos en un formato que sea útil. Si vienen en forma electrónica, ya están en un formato, y con suerte quizás estén en uno de uso bien conocido. La forma más flexible de obtener datos en forma electrónica es un **archivo de texto (ASCII)** estándar. Todos los paquetes de software de estadística y econometría permiten que los datos brutos se almacenen de esta forma. En general, es fácil leer directamente un archivo de texto en un paquete de econometría, siempre que el archivo esté estructurado de la manera adecuada. Los archivos de datos que se han utilizado en todo el libro ofrecen varios ejemplos de cómo los datos de corte transversal, series de tiempo, cortes transversales combinados y datos de panel suelen almacenarse. Como regla, los datos deben tener una forma tabular, en la que cada observación representa una fila diferente; las columnas en el conjunto de datos representan diferentes variables. En ocasiones, quizás encuentre conjuntos de datos almacenados en cada columna que representen una observación y cada fila una variable diferente. Esto no es lo ideal, pero la mayoría de los paquetes de software permiten que los datos se lean de esta forma y se reconfiguren. Como es natural, es crucial saber cómo están organizados los datos antes de leerlos en su paquete de econometría.

Para los conjuntos de datos de series de tiempo, sólo hay una forma sensible de ingresar y almacenar los datos: por nombre, de manera cronológica, con el periodo de tiempo más antiguo listado como la primera observación y el periodo más reciente como la última. Suele ser útil incluir variables que indiquen el año y, si son relevantes, el trimestre o el mes. Esto, más adelante, facilita la estimación de una variedad de modelos como permitir la estacionalidad e interrupciones en periodos de tiempo diferentes. Para los cortes transversales combinados con el paso del tiempo, suele ser mejor llenar el primer bloque de observaciones con el corte transversal del primer año, y así sucesivamente. (Vea FERTIL1.RAW, como ejemplo). Esta distribución no es común, pero es muy importante tener una variable que exprese el año junto a cada observación.

Para los datos de panel, como se analizó en la sección 13.5, es mejor si los años de la observación de corte transversal son adyacentes y están en orden cronológico. Con este ordenamiento se pueden utilizar todos los métodos de datos de panel de los capítulos 13 y 14. Con los datos de panel es importante incluir un identificador único para cada unidad de corte transversal, junto con una variable de año.

Si se obtienen datos en forma impresa, se tendrán varias opciones para ingresarlos en una computadora. Primero, puede crear un archivo de texto mediante un **editor de texto** estándar. (Esta es la forma en que varios de los conjuntos de datos brutos incluidos en este libro se crearon en un principio.) Por lo general, se requiere que cada fila comience con una nueva observación, que contenga el mismo orden de las variables, en particular, que cada fila deba tener el mismo número de entradas, y que los valores estén separados por al menos un espacio. Algunas veces, un separador diferente, como una coma, es mejor pero depende del software que se esté utilizando. Si no se tienen observaciones sobre algunas variables, debe decidirse cómo señalar esto; generalmente no funciona dejar sólo un espacio en blanco. Varios paquetes de regresión aceptan un periodo como el símbolo del valor faltante. Algunas personas prefieren utilizar un número, supuestamente un valor imposible para la variable, para denotar valores faltantes. Si no se tiene cuidado, esto puede ser peligroso, lo cual se analizará con mayor detalle más adelante.

Si se tienen datos no numéricos, por ejemplo, se quiere incluir los nombres en una muestra de colegios o los nombres de las ciudades, entonces debe revisarse el paquete de econometría que utilizará para ver la mejor forma de ingresar tales variables (que suelen llamarse *cadenas*). Por lo general, las cadenas se colocan entre comillas sencillas y dobles. O el archivo de texto puede seguir un formato rígido, lo cual requiere, por lo general, un programa pequeño para leer el archivo de texto. Pero es necesario revisar los detalles del paquete de econometría.

Otra opción generalmente disponible es utilizar una **hoja de cálculo**, como Excel, para ingresar los datos. Esto tiene algunas ventajas sobre un archivo de texto. Primero, dado que cada observación en cada variable está en una celda, es menos probable que los números corran juntos (como sucedería si olvida ingresar un espacio en un archivo de texto). Segundo, las hojas de cálculo permiten la manipulación de datos, como la clasificación o el cálculo de promedios. Este beneficio es menos importante si se utiliza un software que permita un manejo sofisticado de datos; muchos paquetes, como EViews y Stata, se encuentran en esta categoría. Si se utiliza una hoja de cálculo para el ingreso inicial de datos, entonces se deben exportar los datos en una forma que el paquete de econometría pueda leer. Esto suele ser sencillo, pues las hojas de cálculo se exportan a archivos de texto mediante una variedad de formatos.

Una tercera alternativa es ingresar directamente los datos a su paquete de econometría. Aunque esto elimina la necesidad de un editor de texto o una hoja de cálculo, puede ser engorroso si no se puede mover con libertad a través de las diferentes observaciones para realizar correcciones o sumas.

Los datos descargados de Internet pueden tener una variedad de formatos. Los datos suelen venir como archivos de texto, pero se utilizan diferentes convenciones para separar las variables; para los conjuntos de datos de panel, las convenciones sobre cómo ordenar los datos pueden diferir. Algunos conjuntos de datos de Internet aparecen como archivos de hojas de cálculo, en cuyo caso se debe utilizar una hoja apropiada de cálculo para leerlos.

Inspección, depuración y resumen de los datos

Es crucial que se familiarice con el conjunto de datos que se usará en un análisis empírico. Si se ingresan los datos, se estará obligado a saber todo acerca de ellos. Pero si se obtienen datos de una fuente externa, se tendrá que invertir algún tiempo para comprender su estructura y convenciones. Incluso los conjuntos de datos amplia y altamente documentados pueden contener defectos. Si se está utilizando un conjunto de datos, obtenido del autor de un trabajo, debe estar consciente de que las reglas para la construcción de conjuntos de datos pueden haberse olvidado.

En los párrafos anteriores se revisaron las formas usuales en las que se almacenan varios conjuntos de datos. También se necesita saber cómo están codificados los valores faltantes. Si se utiliza un número como un código de valor faltante, como "999" o "-1", debe tener cuidado cuando se utilicen estas observaciones en el cómputo de cualquier estadística. El paquete de econometría quizás ignore que un cierto número representa en realidad un valor faltante: es probable que tales observaciones se utilicen como si fueran válidas y esto puede producir resultados muy equivocados. El mejor método es establecer todos los códigos numéricos para valores faltantes con algún otro símbolo (como un punto) que no pueda confundirse con los datos reales.

También se debe conocer la naturaleza de las variables en el conjunto de datos. ¿Cuáles son las variables binarias? ¿Cuáles son las variables ordinarias (como la calificación crediticia)? ¿Cuáles son las unidades de medida de las variables? Por ejemplo, ¿los valores monetarios están expresados en dólares, miles de dólares, millones de dólares, etc.? ¿Las variables que representan una tasa, como las tasas de deserción escolar, las tasas de inflación, las tasas de sindicalización o las tasas de interés, están medidas como porcentaje o como una proporción?

En particular, para los datos de series de tiempo, es crucial saber si los valores monetarios están expresados en dólares nominales (actuales) o reales (constantes). Si los valores están expresados en términos reales, ¿cuál es el año o periodo base?

Si usted recibe un conjunto de datos de un autor, algunas variables pueden transformarse de ciertas maneras. Por ejemplo, algunas veces sólo el log de una variable (como el sueldo o el salario) se reporta en el conjunto de datos.

Detectar errores en un conjunto de datos es necesario para preservar la integridad de cualquier análisis de datos. Siempre es útil encontrar los mínimos, máximos, medias y desviaciones estándar de todas las variables en el análisis, al menos de las más importantes. Por ejemplo, si se encuentra que el valor mínimo de la educación en su muestra es -99, usted sabe que al menos una entrada sobre la educación se debe establecer como un valor faltante. Si, después de una mayor inspección, encuentra que varias observaciones tienen -99 como el nivel de educación, puede tener la confianza de que se ha topado con el código de valor faltante para la educación. Por ejemplo, si encuentra que la tasa promedio de condenas por asesinato a través de una muestra de ciudades es .632, sabe que la tasa de condenas se mide como una proporción y no como porcentaje. Entonces, si el valor máximo es superior a uno, es probable que exista un error tipográfico. (No es poco común encontrar conjuntos de datos que se ingresaron como una proporción, y viceversa. Tales errores de codificación de datos pueden ser difíciles de detectar, pero es importante intentar.)

También se debe tener cuidado cuando se utilicen datos de series de tiempo. Si se usan datos mensuales o trimestrales, se debe saber qué variables, si las hay, se han ajustado estacionalmente. Transformar los datos también requiere un gran cuidado. Suponga que se tiene un conjunto de datos mensuales y se quiere crear el cambio en una variable de un mes al siguiente. Para hacerlo, es necesario asegurarse que los datos están ordenados cronológicamente, del periodo más antiguo al más reciente. Si por alguna razón éste no es el caso, esta diferenciación generará resultados inservibles. Para asegurarse de que los datos están ordenados adecuadamente, es útil tener un indicador de periodos de tiempo. Con datos anuales, es suficiente conocer el año, pero entonces se debe saber si el año se ingresó con cuatro o con dos dígitos (por ejemplo 1998 o 98). Con datos mensuales o trimestrales, también es útil tener una variable o más variables que indiquen el mes o el trimestre. Con los datos mensuales se puede tener un conjunto de variables binarias (11 o 12) o una variable que indique el mes (1 a 12 o una variable en cadena, como *ene*, *feb*, etcétera).

Con o sin indicadores anuales, mensuales o trimestrales, es posible construir tendencias temporales en todos los paquetes de software de econometría. Crear variables binarias estacionales es fácil si se indican el mes o el trimestre, al menos se necesita saber el mes o el trimestre de la primera observación.

Manipular los datos de panel puede ser aún más difícil. En el capítulo 13, se analizaron los datos combinados MCO sobre los datos diferenciados como un enfoque general para controlar los efectos inobservables. Cuando se construyen los datos diferenciados, se debe tener cuidado de no crear observaciones fantasma. Suponga que se tiene un panel balanceado de ciudades de 1992 a 1997. Incluso si los datos estuvieran ordenados de manera cronológica dentro de cada unidad de corte transversal, algo que se debe hacer antes de comenzar, una diferenciación descuidada creará una observación de 1992 para todas las ciudades, salvo la primera en la muestra. Esta observación será el valor de 1992 para la ciudad i, menos el valor de 1997 para la ciudad $i - 1$; esto claramente es un sin sentido. Por tanto, debe estar seguro de que 1992 falta en todas las variables diferenciadas.

19.4 Análisis econométrico

Este libro se ha enfocado en el análisis econométrico, y no se dará una revisión de los métodos econométricos en esta sección. No obstante, se pueden dar algunos lineamientos generales acerca del tipo de cuestiones que se deben considerar en un análisis empírico.

Como se analizó antes, después de decidir un tema, se debe recabar un conjunto adecuado de datos. En el supuesto de que esto también se haya hecho, se deben decidir, a continuación, los métodos econométricos adecuados.

Si su curso se enfocó en la estimación por mínimos cuadrados ordinarios de un modelo de regresión lineal múltiple, usando datos de corte transversal o de series de tiempo, el enfoque econométrico es el idóneo. Esto no necesariamente es una debilidad, puesto que MCO sigue siendo el método econométrico que más ampliamente se utiliza. Por supuesto, aún se debe decidir si alguna de las variantes de MCO, como los mínimos cuadrados ponderados o corregir una correlación serial en una regresión de series de tiempo, están garantizadas.

Con el fin de justificar la estimación por MCO, se debe hacer una argumentación convincente de que los supuestos clave de MCO se satisfacen en el modelo. Como se ha analizado con cierta amplitud, la primera cuestión es si el término de error no está correlacionado con las variables explicativas. Idealmente, usted ha podido controlar suficientes factores para suponer que los que se dejaron en el error no estaban relacionados con los regresores. En especial, cuando se trata con datos de corte familiar individuales, familiares o a nivel de una empresa, el problema de la autoselección, del que se habló en los capítulos 7 y 15, suele ser relevante. Por ejemplo, en el caso de IRA de la sección 19.3, puede ser que las familias con un gusto inobservable para el ahorro también sean los que abren IRA. Se debe poder argumentar que otras posibles fuentes de endogeneidad, por ejemplo, el error de medición o la simultaneidad, no son un problema serio.

Cuando se especifique el modelo también se deben tomar decisiones funcionales. ¿Algunas variables aparecen en forma logarítmica? (En las aplicaciones econométricas la respuesta suele ser afirmativa.) ¿Algunas variables deben incluirse en los niveles y cuadrados para capturar posiblemente un efecto decreciente? ¿Cómo deben aparecer los factores cualitativos? ¿Basta con sólo incluir variables binarias para diferentes atributos o grupos? O, ¿necesitan interactuar con las variables cuantitativas? (Vea los detalles en el capítulo 7.)

Un error común, en especial entre principiantes, es incluir de forma incorrecta las variables explicativas en un modelo de regresión que estén listadas como valores numéricos, pero que no tengan significado. Por ejemplo, en un conjunto de datos a nivel individual que contiene información sobre salarios, educación, experiencia y otras variables, se debe incluir una variable de "ocupación". Por lo general, éstos son sólo códigos arbitrarios que se han asignado a diferentes ocupaciones; el hecho de que se le dé a un profesor de enseñanza elemental un valor de, por ejemplo, 453 mientras que a un técnico de cómputo se le dé, por decir, 751, es relevante sólo en cuanto a que permite distinguir entre las dos ocupaciones. No tiene sentido incluir la variable ocupacional bruta en un modelo de regresión. (¿Qué sentido tendría medir el efecto de una ocupación creciente en uno, cuando el incremento unitario no tiene significado cuantitativo?) En lugar de esto, se deben definir diferentes variables binarias para distintas ocupaciones (o grupos de ocupaciones, si hay muchas de ellas). Entonces, las variables binarias se pueden incluir en el modelo de regresión. Un problema menos egregio ocurre cuando una variable cualitativa ordenada se incluye como una variable explicativa. Suponga que en conjunto de datos salariales se incluye una variable que mide la "satisfacción laboral", definida en una escala de 1 a 7, con el 7 como la más satisfactoria. Siempre que se tengan datos suficientes, se querría definir un conjunto de seis variables binarias para, por ejemplo, niveles de satisfacción laboral de 2 a 7, donde el nivel de satisfacción laboral de 1 es el grupo base. Al incluir las seis variables binarias de satisfacción laboral en la regresión, se permite una relación completamente flexible entre la variable de respuesta y la satisfacción laboral. Colocar la variable de satisfacción laboral en forma bruta supone de forma implícita que un incremento unitario en la variable ordinal tiene un significado cuantitativo. Si bien la dirección del efecto se estimará con frecuencia, de manera adecuada, interpretar el coeficiente sobre una variable ordinal es difícil. Si una variable ordinal asume varios valores, se puede definir un conjunto de variables binarias para rangos de valores. Vea la sección 7.3 para un ejemplo.

En ocasiones se quiere explicar una variable que sea una respuesta ordinal. Por ejemplo, podría pensarse en utilizar la variable de satisfacción laboral, del tipo descrito antes, como la variable dependiente en un modelo de regresión, con las características tanto del trabajador como del empleador entre las variables independientes. Por desgracia, con la variable de satisfacción laboral en su forma original, los coeficientes del modelo son difíciles de interpretar: cada uno mide el cambio en la satisfacción laboral dado un incremento unitario en la variable independiente. Ciertos modelos, como *probit ordenados* y *logit ordenados* son los más comunes e idóneos para respuestas ordenadas. Estos modelos en esencia extienden los modelos binarios probit y logit que se analizaron en el capítulo 17. [Vea Wooldridge (2002, capítulo 15) para un tratamiento de los modelos de respuesta ordenada.] Una solución simple es convertir cualquier respuesta ordenada en una respuesta binaria. Por ejemplo, se podría definir una variable igual a uno si la satisfacción laboral es de al menos 4, y cero en caso contrario. Por desgracia, crear una variable binaria elimina información y requiere que se utilice algún corte arbitrario.

Para un análisis de corte transversal, una cuestión secundaria, pero no menos importante, es si existe heterocedasticidad. En el capítulo 8 se explicó cómo manejarla. La forma más simple es calcular los estadísticos de heterocedasticidad robusta.

Como se enfatizó en los capítulos 10, 11 y 12, las aplicaciones de series de tiempo requieren cuidados adicionales. ¿Una ecuación debe estar estimada en niveles? Si se usan niveles, ¿se necesitan tendencias temporales? ¿Diferenciar los datos es más apropiado? ¿Si los datos son mensuales o trimestrales, la estacionalidad se debe tomar en cuenta? Si se permiten dinámicas, por ejemplo, dinámica de rezagos distribuidos, ¿cuántos rezagos se deben incluir? Debe comenzar con algún rezago basado en la intuición o el sentido común, pero eventualmente es una cuestión empírica.

Si el modelo tiene algún error potencial, como variables omitidas y utiliza MCO, debe intentar alguna clase de **análisis de error de especificación** de los tipos que se analizaron en los capítulos 3 y 5. ¿Es posible determinar, con base en supuestos razonables, la dirección de cualquier sesgo en los estimadores?

Del estudio del método de las variables instrumentales, se sabe que es posible utilizarlo para resolver varias formas de endogeneidad, incluidas las variables omitidas (capítulo 15) y simultaneidad (capítulo 16). Como es natural, es necesario analizar con profundidad si es probable que las variables instrumentales en consideración sean válidas.

Los buenos trabajos en ciencias sociales empíricas contienen **análisis de sensibilidad**. En términos generales, esto significa que se estima el modelo original y es modificado de formas que parezcan razonables. Con algo de suerte las conclusiones importantes no cambian. Por ejemplo, si se usa como variable explicativa una medida del consumo de alcohol (por ejemplo, en una ecuación de promedio de calificaciones), ¿obtiene resultados cualitativamente similares si remplaza la medida cuantitativa con una variable binaria que refleje el consumo de alcohol? Si la variable binaria sobre el consumo es significativa, pero la variable de la cantidad de alcohol no lo es, podría tal consumo reflejar algún atributo inobservable que influya el GPA y también esté correlacionado con el consumo del alcohol. Pero esto necesita considerarse de manera casuística.

Si algunas observaciones son muy diferentes del grueso de la muestra, por ejemplo, si se tienen algunas empresas en una muestra que sean mucho más grandes que las otras empresas, ¿los resultados cambiarán mucho cuando tales observaciones se excluyan de la estimación? Si es así, quizá se deban alterar las formas funcionales para tomar en consideración estas observaciones o argumentar que siguen un modelo completamente diferente. La cuestión de las observaciones aberrantes se estudió en el capítulo 9.

Utilizar los datos de panel plantea algunos problemas econométricos adicionales. Suponga que ha recabado dos periodos. Existen al menos cuatro formas de usar dos periodos de datos de panel sin recurrir a variables instrumentales. Se pueden combinar los dos años en un análisis estándar MCO, como se estudió en el capítulo 13. Aunque esto podría incrementar el tamaño muestral con relación a un solo corte transversal, no controla los inobservables constantes en el tiempo. Además, los errores en una ecuación de este tipo casi siempre están serialmente correlacionados debido a un efecto inobservable. La estimación de los efectos aleatorios corrige el problema de correlación serial y produce estimadores asintóticamente eficientes, siempre que el efecto inobservable tenga una media cero, dados los valores de las variables explicativas en todos los periodos de tiempo.

Otra posibilidad es incluir una variable dependiente rezagada en la ecuación para el segundo año. En el capítulo 9 se presentó esta posibilidad como una forma de mitigar al menos el problema de las variables omitidas, ya que en todo caso se mantiene fijo el resultado inicial de la variable dependiente. Esto suele generar resultados similares a hacer una diferenciación de los datos, como se analizó en el capítulo 13.

Con más años de datos de panel, se tienen las mismas opciones, más una alternativa adicional. Se puede utilizar la transformación de efectos fijos para eliminar el efecto inobservable. (Con datos de dos años, esto es lo mismo que hacer la diferenciación.) En el capítulo 15 se demostró cómo se pueden combinar las técnicas de variables instrumentales con las transformaciones de datos de panel para relajar los supuestos de exogeneidad. Como regla general, es buena idea aplicar varios métodos econométricos y comparar los resultados. Esto permite determinar cuál de los supuestos planteados probablemente será falso.

Aun si se tiene mucho cuidado al diseñar un tema, postular el modelo, recabar los datos y llevar a cabo la econometría, es muy posible que se obtengan resultados desconcertantes, al menos en algún momento. Cuando eso suceda, la tendencia natural es intentar diferentes modelos, distintas técnicas de estimación o quizá diferentes subconjuntos de datos hasta que los resultados correspondan más a lo que se esperaba. Prácticamente todas las personas que realizan una investigación aplicada investigan varios modelos antes de hallar el "mejor" de ellos. Por desgracia, esta práctica de **minería de datos** viola los supuestos que se han planteado en el análisis econométrico. Los resultados sobre el insesgamiento de MCO y de otros estimadores, así como las distribuciones t y F que se derivaron de las pruebas de hipótesis, suponen que se observa una muestra que sigue el modelo poblacional y que ya se estimó ese modelo alguna vez. Estimar modelos que son variantes del modelo original viola aquel supuesto pues, se está utilizando el mismo conjunto de datos en una *búsqueda de especificación*. En efecto, se utiliza el resultado de las pruebas con ayuda de los datos para volver a especificar este modelo. Las estimaciones y pruebas de diferentes especificaciones de modelos no son independientes entre sí.

Algunas búsquedas de especificación se han programado en paquetes estándar de software. Uno muy conocido es la *regresión por pasos*, donde se utilizan diferentes combinaciones de variables explicativas en el análisis de regresión múltiple en un intento por obtener el mejor modelo. Existen varias formas en que es posible la regresión por pasos, y en este libro no se tiene la intención de hacer un repaso de ellas. La idea general es, o comenzar con un modelo general y mantener variables cuyos valores-p estén por debajo de un cierto nivel de significancia, o comenzar con un modelo simple y agregar variables que tengan valores-p significativos. En ocasiones, los grupos de variables se evalúan mediante una prueba F. Por desgracia, el modelo final suele depender del orden en que las variables se eliminaron o agregaron. [Para más información sobre la regresión por pasos, vea Draper y Smith (1981).] Además, esta es una forma rigurosa de minería de datos y resulta difícil interpretar los estadísticos t y F en el modelo final. Se podría argumentar que la regresión por pasos simplemente automatiza lo que los investigadores hacen de cualquier modo al buscar entre diversos modelos. No obstante, en la mayoría de las aplicaciones, una o dos variables explicativas son de interés fundamental y, entonces, la meta es ver qué tan robustos son los coeficientes de esas variables si se agregan o eliminan otras o se modifica la forma funcional.

En principio, es posible incorporar los efectos de la minería de datos a la inferencia estadística; sin embargo, en la práctica, es muy difícil y pocas veces se hace, en especial en el trabajo empírico sofisticado. [Vea Leamer (1983) para un estudio fascinante sobre este problema.] Pero se puede intentar minimizar la minería de datos si se deja de buscar en numerosos modelos o métodos de estimación hasta hallar un resultado significativo y después reportar sólo ese resultado. Si una variable es estadísticamente significativa en sólo una pequeña fracción de los modelos estimados, es muy probable que la variable no tenga efecto en la población.

19.5 La redacción de un trabajo empírico

Redactar un ensayo que utilice un análisis econométrico es todo un desafío, pero también puede ser gratificante. Un trabajo exitoso combina un análisis de datos cuidadoso y convincente, con una buena explicación y exposición. Por tanto, se debe tener un buen dominio del tema, una buena comprensión de los métodos econométricos, y sólidas habilidades de redacción. No se debe desanimar si se le dificulta escribir un trabajo empírico; la mayoría de los investigadores profesionales han pasado varios años aprendiendo el oficio de crear con destreza un análisis empírico y escribir los resultados de una forma convincente.

Si bien los estilos ensayísticos varían, muchos trabajos siguen el mismo patrón general. Los siguientes párrafos incluyen algunas ideas para los títulos de las secciones y explicaciones acerca de qué debe contener cada una de ellas. Éstas son sólo sugerencias y no necesitan seguirse al pie de la letra. En el trabajo final, a cada sección puede asignársele un número, que suele comenzar con uno para la introducción.

Introducción

La introducción plantea los objetivos básicos del estudio y explica por qué es importante. Por lo general, incluye una revisión bibliográfica, la cual indica qué se ha hecho antes y cómo pueden mejorarse los trabajos previos. (Como se analizó en la sección 19.2, cuando se trata de una revisión extensa ésta se puede escribir en una sección aparte.) Mostrar estadísticas o gráficas simples que revelen una relación aparentemente paradójica es una forma útil de presentar el tema de trabajo. Por ejemplo, suponga que se está escribiendo un trabajo acerca de los factores que afectan la fertilidad en un país en vías de desarrollo, con un enfoque en los niveles educativos de las mujeres. Una forma atractiva de presentar el tema sería presentar una tabla o gráfica, la cual muestre que la fertilidad ha estado (por decir) disminuyendo con el paso del tiempo, y una breve explicación de cómo se espera examinar los factores que han contribuido a esta disminución. En este punto quizá ya sepa que, *ceteris paribus*, más mujeres con altos niveles educativos tienen menos hijos y que los niveles educativos promedio han estado aumentando con el tiempo.

A la mayoría de los investigadores les gusta resumir los hallazgos de su trabajo en la introducción. Esto puede ser una estrategia útil para captar la atención del lector. Por ejemplo, quizá se afirme que la mejor estimación del efecto de faltar a 10 horas de clase durante un curso de 30 horas es cerca de medio punto de la calificación. Pero el resumen no debe ser demasiado detallado, porque aún no se han presentado los métodos ni los datos usados para obtener las estimaciones.

Marco conceptual (o teórico)

En esta sección se describe el método general que se empleó para responder la pregunta planteada. Puede consistir en teoría económica formal, pero en muchos casos, es un análisis intuitivo acerca de qué problemas conceptuales surgen al tratar de responder a la pregunta.

Por ejemplo, suponga que se están estudiando los efectos de las oportunidades económicas y la severidad del castigo para la conducta delictiva. Un enfoque para explicar la participación en los delitos es especificar un problema de maximización de utilidad donde el individuo elige la cantidad de tiempo que va invertir en actividades legales e ilegales, dadas las tasas salariales en ambos tipos de actividades, así como una variable que mide la probabilidad y la severidad del castigo para la actividad delictiva. La utilidad de tal ejercicio sugiere qué variables se deben incluir en el análisis empírico; esto da una guía (pero rara vez detalles) en cuanto a cómo deben aparecer las variables en el modelo econométrico.

Con frecuencia no hay necesidad de escribir una teoría económica. Para el análisis de políticas econométricas, el sentido común suele bastar para especificar un modelo. Por ejemplo, suponga que se está interesado en estimar los efectos de la participación en Ayuda a Familias con Menores de Edad (AFDC, por sus siglas en inglés) sobre el desempeño de los niños en la escuela. AFDC ofrece ingresos complementarios, pero la participación también facilita que reciban ayuda del Medicaid y otras prestaciones. La parte difícil de tal análisis es decidir el conjunto de variables que se debe controlar. En este ejemplo es posible controlar el ingreso familiar (incluido el AFDC y cualquier otro ingreso por concepto de asistencia social), la educación de la madre, si la familia vive en un área urbana y otras variables. Por tanto, la inclusión de un indicador de participación AFDC, con algo de suerte, medirá las prestaciones de la participación AFDC no provenientes del ingreso. Un análisis de qué factores deben controlarse y los mecanismos a través de los que la participación AFDC podría mejorar el rendimiento escolar se sustituye con la teoría económica formal.

Métodos econométricos y métodos de estimación

Es muy útil contar con una sección que contenga algunas ecuaciones del tipo que se estima y presentarlas en la sección de resultados del trabajo. Esto le permitirá fijar sus ideas acerca de cuál es la variable explicativa clave y qué otros factores controlará. Escribir ecuaciones que contengan términos de error le permite analizar si MCO es un método de estimación adecuado.

Esta es la sección en la que debe hacerse la distinción entre un *modelo* y un método de estimación. Un modelo representa una relación *poblacional* (definida en términos generales para dar cuenta de las ecuaciones de series de tiempo). Por ejemplo, se debe escribir

$$colGPA = \beta_0 + \beta_1 alcohol + \beta_2 hsGPA + \beta_3 SAT + \beta_4 female + u \quad \textbf{19.1}$$

para describir la relación entre GPA escolar y el consumo de alcohol, con algunos otros controles en la ecuación. Supuestamente, esta ecuación representa una población, como todos los estudiantes universitarios en una universidad determinada. No hay "gorros" (ˆ) en β_j o en *colGPA* debido a que este es un modelo, no una ecuación estimada. Aunque no se colocan números para β_j porque esos números se ignoran (y siempre se ignorarán), más adelante, se *estimarán*. En esta sección no se debe anticipar la presentación de los resultados empíricos. En otras palabras, no se debe comenzar con un modelo general y después mencionar que se omitieron ciertas variables porque resultaron ser insignificantes. Tales argumentos deben reservarse para la sección de resultados.

Un modelo de series de tiempo para relacionar los robos de automóviles en una ciudad con la tasa de desempleo y las tasas de condena podría ser así

$$\begin{aligned} thefts_t = \beta_0 + \beta_1 unem_t + \beta_2 unem_{t-1} + \beta_3 cars_t \\ + \beta_4 convrate_t + \beta_5 convrate_{t-1} + u_t, \end{aligned} \quad \textbf{19.2}$$

donde el subíndice t es útil para enfatizar cualquier dinámica en la ecuación (en este caso, dar cuenta de que las tasas de desempleo y de condena por robos de automóviles tienen efectos rezagados).

Después de especificar un modelo o modelos, es adecuado analizar los métodos de estimación. En la mayoría de los casos, el análisis será de los MCO pero, por ejemplo, en una ecuación de series de tiempo, quizá se utilicen MCG factibles para hacer una correlación serial (como en el capítulo 12). No obstante, el método para estimar un modelo es muy distinto del modelo mismo. No tiene importancia, por ejemplo, hablar de un "modelo MCO". El método de los mínimos cuadrados ordinarios es un método de estimación, como lo son también los mínimos cuadrados ponderados, Cochrane-Orcutt, etc. Por lo general, hay varias formas de estimar cualquier modelo. Debe explicarse por qué el método elegido es el idóneo.

Cualquier supuesto que se utilice para obtener un modelo econométrico estimado del modelo económico básico debe analizarse con claridad. Vea el caso en la calidad del bachillerato del ejemplo mencionado en la sección 19.1, la cuestión de cómo medir la calidad de la escuela es fundamental para el análisis. ¿Deberá basarse en el promedio de puntuaciones SAT, en el porcentaje de graduados que asisten a la universidad, en la proporción entre profesores y estudiantes, en el nivel educativo promedio de los profesores, alguna combinación entre éstos o quizás en otras medidas?

Sin importar si se presentó o no un modelo teórico, siempre se tienen que hacer suposiciones acerca de la forma funcional. Como ya se sabe, los modelos de elasticidad y semielasticidad constantes son atractivos, porque los coeficientes son fáciles de interpretar (como efectos porcentuales). No hay reglas estrictas sobre cómo elegir una forma funcional, pero en la práctica, los lineamientos que se analizaron en la sección 6.2 parecen funcionar bien. No es necesario un análisis extenso de la forma funcional, pero es útil mencionar si se estimarán las elasticidades o una semielasticidad. Por ejemplo, si se estima el efecto de alguna variable en el sueldo o salario, la variable dependiente, casi con seguridad, estará en forma logarítmica, y quizá también se incluya en cualquier ecuación desde el principio. No es necesario que se presenten todas, ni siquiera la mayoría de las variaciones de la forma funcional que se reportarán más adelante en la sección de resultados.

Con frecuencia, los datos que se utilizan en economía empírica están al nivel de ciudad o país. Por ejemplo, suponga que para la población de las ciudades pequeñas o medianas se desea probar la hipótesis de que tener una liga menor de béisbol ocasiona que la ciudad posea una tasa de divorcios menor. En este caso debe tomar en cuenta el hecho de que en las ciudades más grandes habrá más divorcios. Una forma de dar cuenta del tamaño de la ciudad es graduar los divorcios según la población de la ciudad o la población adulta. Por tanto, un modelo razonable sería

$$\log(div/pop) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc/pop) + otros\ factores, \qquad \boxed{19.3}$$

donde *mlb* es una variable binaria igual a uno si la ciudad tiene un equipo de béisbol de liga menor y *perCath* es el porcentaje de la población que es católica (así que un número como 34.6 significa 34.6%). Observe que *div*/*pop* es una tasa de divorcio, la cual es más fácil de interpretar que el número absoluto de divorcios.

Otra forma de controlar la población es estimar el modelo

$$\log(div) = \gamma_0 + \gamma_1 mlb + \gamma_2 perCath + \gamma_3 \log(inc) + \gamma_4 \log(pop) + otros\ factores. \qquad \boxed{19.4}$$

El parámetro de interés, γ_1, cuando se multiplica por 100, produce la diferencia porcentual entre tasas de divorcio, manteniendo constantes la población, el porcentaje de católicos, el ingreso y cualquier otra cosa que esté en "otros factores" constante. En la ecuación (19.3), β_1 mide el efecto porcentual de una liga menor de béisbol sobre *div*/*pop*, que puede cambiar también debido al número de divorcios o cambios en la población. Mediante el hecho de que $\log(div/pop) = \log(div) - \log(pop)$ y $\log(inc/pop) = \log(inc) - \log(pop)$, se puede reescribir (19.3) como

$$\log(div) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc) + (1 - \beta_3)\log(pop) + otros\ factores,$$

lo cual muestra que (19.3) es un caso especial de (19.4) con $\gamma_4 = (1 - \beta_3)$ y $\gamma_j = \beta_j$, $j = 0,1,2,3$. Por otra parte, (19.4) es equivalente a agregar $\log(pop)$ como una variable explicativa adicional a (19.3). Esto facilita probar un efecto poblacional separado en la tasa de divorcios.

Si se está utilizando un método de estimación más avanzado, como el de los mínimos cuadrados en dos etapas (MC2E), es necesario dar algunas razones por las que se hizo. Si se usa MC2E, debe ofrecerse un análisis detallado sobre por qué las opciones de VI para la variable (o variables) explicativa endógena son válidas. Como se mencionó en el capítulo 15, existen dos requisitos para que una variable se considere una buena VI. Primero, dentro de la ecuación de interés (ecuación estructural), debe omitirse o ser exógena. Esto es algo que se debe suponer. Segundo, se debe tener alguna correlación parcial con la variable explicativa endógena. Esto se puede probar. Por ejemplo, en la ecuación (19.1), se debe utilizar una variable binaria para el caso de que un estudiante viva en un dormitorio (*dorm*) como una VI para el consumo de alcohol. Esto requiere que la situación real no tenga un impacto directo sobre *colGPA*, así que se omitió en (19.1), y no está correlacionada con los factores observables en *u* que tienen un efecto sobre *colGPA*. También se tendría que verificar que *dorm* esté parcialmente correlacionado con *alcohol* al hacer una regresión de *alcohol* sobre *dorm*, *hsGPA*, *SAT* y *female*. (Vea detalles en el capítulo 15.)

Quizá se dé cuenta del problema de la variable omitida (o heterogeneidad omitida) mediante datos de panel. De nuevo, esto se describe con facilidad escribiendo una ecuación o dos. De hecho, es útil mostrar cómo diferenciar las ecuaciones con el tiempo para eliminar las inobservables de tiempo constante; esto produce una ecuación que se puede estimar mediante MCO. O, si está utilizando una estimación de efectos fijos, simplemente lo debe indicar.

Por ejemplo, suponga que se está probando si las tasas fiscales superiores de un condado o municipio reducen la actividad económica, medido como producción manufacturera *per cápita*. Suponga que para los años 1982, 1987 y 1992, el modelo es

$$\log(manuf_{it}) = \beta_0 + \delta_1 d87_t + \delta_2 d92_t + \beta_1 tax_{it} + \ldots + a_i + u_{it},$$

donde $d87_t$ y $d92_t$ son las variables binarias y tax_{it} es la tasa fiscal para el condado o municipio i en el tiempo t (en forma porcentual). Se tendrían otras variables que cambian con el tiempo en la ecuación, incluidas las medidas de los costos de hacer negocios (como los salarios promedio), medias de la productividad laboral (como nivel educativo promedio), etc. El término a_i es el efecto fijo, que contiene todos los valores que no varían con el tiempo, y u_{it} es el término de error idiosincrático. Para eliminar a_i, se puede diferenciar a través de los años, o usar la transformación de efectos fijos.

Los datos

Siempre se debe tener una sección que describa con cuidado los datos utilizados en el análisis empírico. Esto es particularmente importante si sus datos son no estándar u otros investigadores no los han usado ampliamente. Se debe presentar suficiente información a fin de que el lector, en principio, pueda obtener los datos y rehacer su análisis. En particular, todas las fuentes de datos públicos aplicables se deben incluir en las referencias, y los conjuntos de datos escasos se pueden listar en el apéndice. Si se usó una encuesta propia para recabar los datos, se debe presentar una copia del cuestionario en el apéndice.

Junto con un análisis de las fuentes de datos, se deben analizar las unidades de cada una de las variables (por ejemplo, ¿el ingreso está medido en cientos o en miles de dólares?). Incluir una tabla de las definiciones de variables es muy útil para el lector. Los nombres en la tabla deben corresponder a los nombres que se utilizan para describir los resultados econométricos en la siguiente sección.

También es muy ilustrativo presentar una tabla resumen de estadística con valores mínimos y máximos, medias, y desviaciones estándar para cada variable. Tener esta tabla hace más fácil interpretar los coeficientes estimados en la siguiente sección y esto enfatiza las unidades de medición de las variables. Para variables binarias, el único resumen de estadística necesario es la fracción de los uno en la muestra (que es la misma para la media muestral). Para variables con tendencia, las cosas relacionadas con las medias son menos interesantes. A menudo es útil calcular la tasa de crecimiento promedio a lo largo de los años en su muestra.

También se debe indicar con claridad cuántas observaciones posee. Para los conjuntos de datos de series de tiempo, se deben identificar los años que se están usando en el análisis, incluida una descripción de cualquier periodo especial en la historia (como la Segunda Guerra Mundial). Si se utiliza una combinación de cortes transversales o de datos de panel, se debe estar seguro del reporte de cuántas unidades de corte transversal (personas, ciudades, etc.) tiene para cada año.

Resultados

La sección de resultados debe incluir las estimaciones de cualquier modelo formulado en la sección de modelos. Se puede empezar con un análisis muy sencillo. Por ejemplo, suponga que el porcentaje de estudiantes que asiste a la universidad después de graduarse (*percoll*) se usa como una medida de la calidad del bachillerato al que asistió una persona. Entonces, la ecuación a estimar es

$$\log(wage) = \beta_0 + \beta_1 percoll + u.$$

Por supuesto, esto no controla varios otros factores que pueden determinar los salarios y que pueden estar correlacionados con *percoll*. Pero un análisis simple puede llevar al lector a un análisis más sofisticado y revelar la importancia de controlar otros factores.

Si sólo se estiman algunas ecuaciones, se pueden presentar los resultados en forma de ecuación con los errores estándar entre paréntesis debajo de los coeficientes estimados. Si el modelo posee muchas variables explicativas y se están presentando muchas variaciones en el modelo general, lo mejor es reportar los resultados en forma tabular y no en forma de ecuación. La mayoría de los trabajos realizados deben tener al menos una tabla, la cual debe incluir siempre al menos la *R*-cuadrada y el número de observaciones de cada ecuación. También se puede listar otros estadísticos, como la *R*-cuadrada ajustada.

Lo más importante es analizar la interpretación y fortalecer sus resultados empíricos. ¿Los coeficientes tienen los signos esperados? ¿Son estadísticamente significativos? Si un coeficiente es estadísticamente significativo, pero tiene un signo contrario a la intuición, ¿por qué sería verdadero? Puede estar revelando un problema con los datos o el método econométrico (por ejemplo, los MCO pueden ser inapropiados debido a problemas de variables omitidas).

Se debe asegurar que las *magnitudes* de los coeficientes, en las principales variables explicativas, se describan. Con frecuencia, existen una o dos variables de política fundamentales para el estudio. Sus signos, magnitudes y significancia estadística se deben tratar con detalle. No se debe olvidar distinguir entre significancia económica y estadística. Si un estadístico *t* es pequeño, ¿esto se debe a que el coeficiente es prácticamente pequeño o a que el error estándar es grande?

Además de analizar las estimaciones del modelo más general, es posible proporcionar casos especiales interesantes, en especial, aquellos necesarios para probar ciertas hipótesis múltiples. Por ejemplo, en un estudio para determinar los diferenciales de salarios entre las industrias, se puede presentar la ecuación sin variables binarias industriales; esto permite al lector probar con facilidad si los diferenciales industriales son estadísticamente significativos (mediante la *R*-cuadrada de la prueba *F*). No es necesario preocuparse demasiado por eliminar muchas variables para encontrar la "mejor" combinación de variables explicativas. Como se mencionó antes, esta es una tarea difícil y no siempre bien definida. Esto será importante si al eliminar un conjunto de variables se alteran sustancialmente las magnitudes y/o la significancia de los coeficientes de interés. Eliminar un grupo de variables para simplificar el modelo, como cuadráticas o interacciones, puede justificarse mediante la prueba *F*.

Si se han utilizado al menos dos métodos diferentes, como MCO y MC2E, o niveles y diferenciación de una serie de tiempo, o MCO combinados en contraste con la diferenciación mediante un conjunto de datos de panel, entonces se debe comentar cualquier diferencia crítica. Si MCO produce resultados contrarios a la intuición, ¿usar métodos MC2E o de datos de panel mejorará las estimaciones? o, ¿sucederá lo contrario?

Conclusiones

Ésta puede ser una breve sección que resuma lo que se ha aprendido. Por ejemplo, quizá se desee presentar la magnitud de un coeficiente que fue de interés particular. La conclusión también debe analizar las advertencias a las conclusiones a las que llegó y quizá, hasta sugerir las direcciones para una investigación ulterior. Es útil imaginar que los lectores leen primero las conclusiones para decidir si leerán el resto del trabajo.

Sugerencias de estilo

Se debe dar al trabajo un título que refleje el tema. Los trabajos deben estar mecanografiados y escritos a doble espacio. Todas las ecuaciones deben comenzar en una línea nueva y estar centradas y numeradas consecutivamente, es decir, (1), (2), (3), etc. Los párrafos y tablas grandes pueden incluirse después del cuerpo principal. En el texto, refiérase a los trabajos por autor y fecha, por ejemplo, White (1980). La sección de referencias al final del trabajo debe tener un formato estándar. Se deben dar varios ejemplos en la sección de referencias bibliográficas, al final del texto.

Cuando se inserte una ecuación en la sección de modelos econométricos, deben describir las variables importantes: la variable dependiente y la variable o variables independientes clave. Para enfocarse en una sola variable independiente se puede escribir una ecuación como

$$GPA = \beta_0 + \beta_1 alcohol + \boldsymbol{x\delta} + u$$

o

$$\log(wage) = \beta_0 + \beta_1 educ + \boldsymbol{x\delta} + u,$$

donde la notación $\boldsymbol{x\delta}$ es una abreviatura de las otras variables explicativas. En este punto sólo se necesita describirlas de manera general; pueden describirse específicamente en la sección de datos de una tabla. Por ejemplo, en un estudio de los factores que afectan el sueldo de un director general, se puede incluir una tabla como la 19.1.

TABLA 19.1

Descripciones de variables

salary	sueldo anual (incluidos los bonos) en 1990 (en miles)
sales	ventas de la empresa en 1990 (en millones)
roe	rendimiento promedio sobre el capital, 1988-1990 (en porcentaje)
pcsal	cambio porcentual en el sueldo, 1998-1990
pcroe	cambio porcentual del rendimiento sobre el capital, 1998-1990
indust	= 1 si es una empresa industrial, 0 de otra manera
finance	= 1 si es una compañía financiera, 0 de otra manera
consprod	= 1 si es una empresa de productos de consumo, 0 de otra manera
util	= 1 si es una empresa de servicios públicos, 0 de otra manera
ceoten	número de años como director general de la empresa

En la tabla 19.2 se presenta un resumen de cómo se pueden configurar las estadísticas mediante la base de datos 401K.RAW, que se usó para estudiar los factores que afectan la participación en los planes de pensión 401(k).

TABLA 19.2

Resumen de estadísticas

Variable	Media	Desviación estándar	Mínimo	Máximo
prate	.869	.167	.023	1
mrate	.746	.844	.011	5
employ	4,621.01	16,299.64	53	443,040
age	13.14	9.63	4	76
sole	.415	.493	0	1
Número de observaciones = 3,784				

En la sección de resultados se pueden escribir las estimaciones en forma de ecuación, como se suele hacer, o en una tabla. En especial, cuando se han estimado varios modelos con diferentes conjuntos de variables explicativas, las tablas son muy útiles. Si se escriben las estimaciones como una ecuación, por ejemplo,

$$\widehat{\log(salary)} = \underset{(0.93)}{2.45} + \underset{(.115)}{.236}\log(sales) + \underset{(.003)}{.008}\, roe + \underset{(.028)}{.061}\, ceoten$$
$$n = 204,\ R^2 = .351,$$

se debe indicar, cerca de la primera ecuación, que los errores estándar están entre paréntesis. Es aceptable reportar el estadístico t para probar H_0: $\beta_j = 0$, o sus valores absolutos, pero es más importante indicar qué se está haciendo.

Si se reportan sus resultados en forma de tabla, se debe asegurar que las variables dependientes y las independientes están indicadas con claridad. Nuevamente se debe indicar si los errores estándar o los estadísticos t están bajo los coeficientes (este último se prefiere). Algunos autores gustan de utilizar asteriscos para indicar la significancia estadística en diferentes niveles de significancia (por ejemplo, una estrella indica significancia de 5%, dos, significancia a 10%, pero no a 5%, y así sucesivamente). No es necesario si se analiza de forma cuidadosa la significancia de las variables explicativas en el texto.

Un ejemplo de tabla con los resultados se muestra en la tabla 19.3.

Los resultados serán más fáciles de leer e interpretar si se eligen las unidades de sus variables dependientes e independientes de manera que los coeficientes no sean demasiado grandes o demasiado pequeños. Nunca se deben reportar números como 1.051e−007 o 3.524e+006 para sus coeficientes o errores estándar, y no se debe utilizar notación científica. Si los coeficientes

TABLA 19.3

Resultados MCO. Variable dependiente: tasa de participación

Variables independientes	(1)	(2)	(3)
mrate	.156 (.012)	.239 (.042)	.218 (.342)
mrate^2	—	−.087 (.043)	−.096 (.073)
log(*emp*)	−.112 (.014)	−.112 (.014)	−.098 (.111)
log(*emp*)^2	.0057 (.0009)	.0057 (.0009)	.0052 (.0007)
age	.0060 (.0010)	.0059 (.0010)	.0050 (.0021)
age^2	−.00007 (.00002)	−.00007 (.00002)	−.00006 (.00002)
sole	−.0001 (.0058)	.0008 (.0058)	.0006 (.0061)
constant	1.213 (.051)	.198 (.052)	.085 (.041)
¿binarias de la industria?	no	no	sí
Observaciones *R*-cuadrada	3,784 .143	3,784 .152	3,784 .162

Nota: Las cantidades entre paréntesis bajo las estimaciones son los errores estándar.

son extremadamente grandes o pequeños, se deben volver a escalar las variables dependientes o independientes, como se analizó en el capítulo 6. Se debe limitar el número de dígitos reportados después del punto decimal. Por ejemplo, si el paquete de regresión estima que un coeficiente es de .54821059, se debe reportar esto en el papel como .548 o incluso .55.

Como regla general, los comandos que utiliza el paquete de econometría en particular suelen producir resultados que no deben aparecer en el papel; sólo los resultados son importantes. Si algún comando especial se usó para realizar cierto método de estimación, éste se puede referir a un apéndice. Un apéndice también es un buen lugar para incluir los resultados adicionales que sustentan el análisis pero que no son fundamentales para él.

RESUMEN

En este capítulo se analizaron los componentes de un estudio empírico exitoso y se han dado sugerencias para mejorar la calidad de un análisis. Finalmente, el éxito de cualquier estudio depende sobre todo del cuidado y el esfuerzo que se le dedique.

TÉRMINOS CLAVE

Análisis de error de especificación
Análisis de sensibilidad
Archivo de texto (ASCII)
Bases de datos en línea
Editor de texto
Hoja de cálculo
Internet
Minería de datos
Servicios de búsqueda en línea

MUESTRA DE PROYECTOS EMPÍRICOS

A través de este libro se vieron ejemplos de análisis econométricos que provinieron o fueron inspirados en trabajos publicados. Cabe esperar que éstos hayan dado una idea acerca del alcance del análisis empírico. Se incluye la siguiente lista como preguntas de ejemplos adicionales que otros encontraron o, posiblemente, encontrarán interesantes. Éstas intentan estimular su imaginación; no se intentó abundar en los detalles de modelos específicos, de requisitos de datos o métodos de estimación alternos. Debe ser posible completar estos proyectos en un curso.

1. Aplique una encuesta en su campus escolar para responder una pregunta de interés para la universidad. Por ejemplo, ¿cuál es el efecto de trabajar sobre las calificaciones universitarias? Se puede preguntar a los estudiantes sus calificaciones promedio del bachillerato, las calificaciones promedio de las universidades, sus calificaciones en los exámenes de ingreso a la universidad o de aptitudes académicas, las horas trabajadas por semana, la participación en actividades deportivas, materias principales cursadas, sexo, raza, etc. Después, usar estas variables para explicar un modelo que explique el promedio de calificaciones. ¿Cuál es el efecto, de haberlo, de otra hora trabajada por semana sobre las calificaciones promedio? Una cuestión de interés es que las horas trabajadas pueden ser endógenas: podrían correlacionarse con factores inobservables que afecten las calificaciones promedio, o bien los promedios más bajos podrían hacer que los estudiantes trabajaran más.

   Una mejor aproximación sería recabar resultados del promedio de las calificaciones universitarias acumuladas antes del semestre y, luego, obtener el promedio del semestre más reciente, junto con las horas trabajadas durante ese periodo y las otras variables. Así, el promedio de calificaciones podría emplearse como control (variable explicativa) en la ecuación.
2. Existen muchas variantes del tema anterior. Se pueden estudiar los efectos del consumo de drogas y alcohol, o de vivir en una fraternidad, sobre el promedio de calificaciones. Quizá se quisieran controlar muchas variables del historial familiar, así como las variables del rendimiento académico anterior.
3. ¿Las leyes de control de armas, a nivel de ciudad, reducen los delitos violentos? Tales preguntas pueden ser difíciles de contestar con un solo corte transversal, debido a que las leyes de la ciudad y del estado suelen ser endógenas. [Vea Kleck y Patterson (1993) para un ejemplo. Ellos utilizaron datos de corte transversal y métodos de variables instrumentales, pero sus VI son cuestionables.] Los datos de panel pueden ser muy útiles para inferir la causalidad en estos contextos. Cuando menos, se podría controlar el índice delictivo del año anterior.
4. Low y McPheters (1983) usaron datos de corte transversal urbanos sobre tasas salariales y estimaciones de riesgo de muerte de los oficiales de policía junto con otros controles. La idea

es determinar si el trabajo de los oficiales de policía está compensado por exponerse a un riesgo más alto de lesión o muerte en el trabajo.

5. ¿Las leyes de consentimiento de los padres aumentan la tasa de embarazos en adolescentes? Es posible utilizar datos a nivel estatal para esto: una serie de tiempo para un determinado estado o, mejor aún, un conjunto de datos de panel de los estados. ¿Las mismas leyes reducen las tasas de aborto entre los adolescentes? El *Statistical Abstract of the United States* contiene todo tipo de datos a nivel estatal. Levine, Trainor y Zimmerman (1996) estudiaron los efectos de las restricciones en el financiamiento para el aborto sobre resultados similares. Otros factores, como el acceso a servicios de práctica de aborto, pueden afectar el embarazo adolescente y las tasas de aborto.
6. ¿Los cambios en las leyes de tránsito afectan las muertes en accidentes de tránsito? McCarthy (1994) contiene un análisis de datos de series de tiempo mensuales para el estado de California. Se puede usar un conjunto de variables binarias para indicar los meses en los cuales ciertas leyes estaban vigentes. El archivo TRAFFIC2.RAW contiene los datos que utilizó McCarthy. Una alternativa es obtener un conjunto de datos de panel sobre los estados de Estados Unidos, donde usted puede explotar la variación entre las leyes de un estado y a lo largo del tiempo. (Vea el archivo TRAFFIC1.RAW.)

   Mullahy y Sindelar (1994) usaron datos a nivel individual en correspondencia con las leyes estatales e impuestos sobre el alcohol para estimar los efectos de las leyes e impuestos sobre la probabilidad de manejar en estado de ebriedad.
7. ¿En el mercado crediticio se discrimina a las personas de color? Hunter y Walker (1996) estudiaron esta cuestión; de hecho, se utilizan sus datos en los ejercicios para computadora C7.8 y C17.2.
8. ¿Hay una prima matrimonial para los atletas profesionales? Korenman y Neumark (1991) encontraron una prima salarial significativa para hombres casados, después de utilizar una variedad de métodos econométricos, pero su análisis es limitado, pues no pudieron observar la productividad de forma directa. (Además, Korenman y Neumark usaron hombres con una variedad de ocupaciones.) Los atletas profesionales constituyen un grupo interesante en el cual estudiar la prima matrimonial, debido a que fácilmente se pueden recabar datos sobre varias medidas de productividad, además del salario. La base de datos NBASAL.RAW, en los jugadores de la *National Basketball Association* (NBA), es un ejemplo. De cada jugador se tiene información de los puntos anotados, recuperaciones, asistencias, tiempo de juego y demografía. Como en el ejercicio para computadora C6.9, se puede usar el análisis de regresión múltiple para probar si las medidas de productividad difieren por estatus marital. También se pueden utilizar este tipo de datos para probar si a los hombres casados se les paga más después de que se consideran las diferencias en productividad. (Por ejemplo, los dueños de la NBA pueden pensar que los hombres casados aportan estabilidad al equipo o que beneficiarán la imagen del mismo.) Para los deportes individuales, como el golf y el tenis, las ganancias anuales reflejan directamente la productividad. Tales datos, junto con la edad y la experiencia, son relativamente fáciles de recabar.
9. Responda esta pregunta: ¿los fumadores son menos productivos? Una variante sería: ¿los trabajadores que fuman se ausentan por enfermedad más días (si todo lo demás se mantiene igual)? Mullahy y Portney (1990) usan datos a nivel individual para evaluar esta pregunta. Usted podría utilizar datos, por ejemplo, a nivel metropolitano. La productividad promedio en manufactura puede relacionarse con el porcentaje de trabajadores de manufactura que fuman. Otras variables, como el promedio de educación del trabajador, el capital por trabajador y el tamaño de la ciudad (quizá se piense en otros más), deben controlarse.
10. ¿Los salarios mínimos alivian la pobreza? Puede usar datos estatales o nacionales para responder esta pregunta. La idea es que el salario mínimo varía entre los diferentes estados, puesto que algunos estados tienen mínimos más altos que el mínimo federal. Además, se presentan cambios con el tiempo en el mínimo nominal dentro de un estado, algunos debido

a cambios en el nivel federal y algunos otros debidos a cambios en el nivel estatal. Neumark y Wascher (1995) usaron un conjunto de datos de panel sobre los estados para estimar los efectos del salario mínimo en las tasas de empleo de los trabajadores jóvenes, así como en las tasas de matriculación escolar.

11. ¿Qué factores afectan el rendimiento estudiantil en las escuelas públicas? Es muy fácil obtener datos a nivel escolar o, al menos, datos a nivel distrital en la mayoría de los estados. ¿El gasto por estudiante importa? ¿Las proporciones entre el número de estudiantes y profesores tienen algún efecto? Es difícil estimar los efectos *ceteris paribus* debido a que el gasto está relacionado con otros factores, como los ingresos familiares o los índices de pobreza. La base de datos MEAP93.RAW para el bachillerato de Michigan contiene una medida de las tasas de pobreza. Otra posibilidad es utilizar datos de panel o al menos controlar la medida del desempeño del año anterior (tal como las calificaciones promedio en una prueba o el porcentaje de estudiantes que aprueban un examen).

    Se pueden estudiar factores menos obvios que afectan el desempeño de los estudiantes. Por ejemplo, después de controlar el ingreso, ¿la estructura familiar importa? Quizá las familias con dos padres, pero con sólo uno que trabaje, tiene un efecto positivo sobre el rendimiento escolar. (Puede haber al menos dos canales: los padres pasan más tiempo con los niños y quizá también participen como voluntarios en la escuela.) ¿Qué hay del efecto de las familias con sólo un padre, en el control del ingreso y otros factores? También se pueden fusionar los datos del censo para un año o dos, con los datos del distrito escolar.

    ¿Las escuelas públicas con más escuelas privadas cercanas educan mejor a sus estudiantes debido a la competencia? Existe una cuestión delicada de simultaneidad aquí debido a que las escuelas privadas quizás están ubicadas en áreas donde las escuelas públicas ya son pobres. Hoxby (1994) usó un método de variables instrumentales, donde las porciones poblacionales de varias religiones eran las VI del número de escuelas privadas.

    Rouse (1998) estudió una cuestión diferente: ¿los estudiantes que pudieron asistir a una escuela privada debido al programa de cupones de Milwaukee tuvieron un rendimiento mejor que los que no pudieron? Ella utilizó datos de panel y fue capaz de controlar un efecto estudiantil inobservable.

12. ¿El exceso de rendimientos sobre una acción, o el índice accionario, se puede predecir por la proporción rezagada entre precio y dividendos? ¿O por las tasas de interés rezagadas o una política monetaria semanal? Sería interesante elegir un índice accionario extranjero o uno de los índices estadounidenses menos conocidos. Cochrane (1997) ofrece una encuesta interesante de las teorías recientes y los resultados empíricos para explicar el exceso de rendimientos accionarios.

13. ¿Existe discriminación racial en el mercado de tarjetas de béisbol? Esto implica relacionar los precios de las tarjetas de béisbol con los factores que deben afectar sus precios, como estadísticas de carreras, si el jugador está en el Salón de la Fama, etc. Si todos los demás factores se mantienen fijos, ¿las tarjetas de los jugadores de color o hispanos se venden por debajo de su precio?

14. Se puede probar si el mercado de las apuestas deportivas es eficiente. Por ejemplo, ¿el rango de posibles resultados en los juegos de fútbol o básquetbol contienen toda la información útil para hacer una apuesta? La base de datos PNTSPRD.RAW contiene información sobre juegos de básquetbol universitario varonil. La variable de resultado es binaria. ¿El rango de resultados está cubierto o no? Entonces, es posible encontrar información conocida antes de que cada juego se realice, con el fin de predecir si el rango de resultados posibles está cubierto. (¡Buena suerte!)

15. ¿Qué efecto, si lo hubiera, tiene el éxito en las actividades deportivas universitarias, sobre otros aspectos de la universidad (solicitudes, calidad de los estudiantes, calidad de los departamentos no deportivos)? McCormick y Tinsley (1987) observaron los efectos del éxito deportivo de las principales universidades en los cambios en las puntuaciones del SAT de los estudiantes

de nuevo ingreso. Aquí, la sincronización de los acontecimientos es importante: supuestamente, el éxito pasado más reciente afecta las solicitudes actuales y la calidad del estudiante. Se deben controlar muchos otros factores, como el costo de la matrícula y las medidas de la calidad escolar, para hacer que el análisis sea convincente pues, sin controlar otros factores, existe una correlación negativa entre el desempeño académico y el desempeño atlético.

Una variante es elegir a los rivales naturales en fútbol o básquetbol varonil para buscar diferencias entre cada escuela en función de qué escuela ganó el partido de fútbol o uno o más juegos de básquetbol. ATHLET1.RAW y ATHLET2.RAW son pequeñas bases de datos que se podrían ampliar y actualizar.

16. Recabe las tasas de asesinatos de una muestra de países (por ejemplo, de los *FBI Uniform Crime Reports*) para dos años. Elija el último año de manera que las variables demográficas y económicas sean fáciles de obtener del *County and City Data Book*. Es posible obtener el número total de personas que están en espera del patíbulo, más las ejecuciones de los años intermedios a nivel del condado o municipio. Si los años son de 1990 a 1985, quizás estima

$$mrdrte_{90} = \beta_0 + \beta_1 mrdrte_{85} + \beta_2 executions + otros\ factores,$$

donde el interés radica en el coeficiente de las *ejecuciones*. La tasa de asesinatos rezagada y otros factores sirven como controles.

Otros factores pueden actuar para disuadir la comisión de delitos. Por ejemplo, Cloninger (1991) presentó un análisis de corte transversal de los efectos de la política de pena de muerte sobre la tasa de delitos.

Desde otro punto de vista, ¿qué factores afectan la tasa de delitos en los campus universitarios? ¿La fracción de estudiantes que vive en fraternidades o hermandades femeninas ejerce algún efecto? ¿El tamaño de la fuerza policiaca o la clase de vigilancia empleada importa? (Se debe tener cuidado aquí al inferir una causalidad.) ¿Tener un programa de acompañante ayuda a reducir los delitos? ¿Qué hay de la tasa de delitos en las comunidades cercanas? Recientemente se exige a los colegios y universidades que reporten sus estadísticas delictivas; en años previos, el reporte era voluntario.

17. ¿Qué factores afectan la productividad manufacturera a nivel estatal? Además de los niveles de capital y educación de los trabajadores, podría estudiar el grado de sindicalización. Un análisis de datos de panel podría resultar más convincente aquí, mediante dos años de censos (por ejemplo, de 1980 y 1990). Clark (1984) ofrece un análisis de la forma en que la sindicalización afecta el desempeño de la empresa y su productividad. ¿Qué otras variables podrían explicar la productividad?

Los datos a nivel de empresa se pueden obtener de *Compustat*. Por ejemplo, si todos los demás factores se mantienen fijos, ¿los cambios en la sindicalización afectarán el precio de las acciones de una empresa?

18. Use los datos a nivel estatal o de condado o, de ser posible, datos a nivel de distrito escolar para considerar factores que afecten el gasto educativo por alumno. Una pregunta interesante es: Con otros factores iguales (como los niveles de educación y de ingreso de los residentes), ¿los distritos con un porcentaje mayor de personas de edad avanzada gastan menos en sus escuelas? Los datos del censo pueden igualarse con los del gasto por distrito escolar para obtener un corte transversal muy grande. El Departamento Estadounidense de Educación compila tales datos.

19. ¿Cuáles son los efectos de las regulaciones estatales, como leyes que imponen la obligación de portar casco cuando se viaja en motocicletas, sobre los decesos en motocicleta? O, ¿las diferencias en las leyes que regulan la navegación en bote, como edad mínima del conductor, ayudan a explicar las tasas en los accidentes en bote? El Departamento Estadounidense de Transporte compila esta información. Un análisis de datos de panel parece lo más recomendable aquí.

20. ¿Qué factores afectan el crecimiento de la producción? Dos factores de interés son la inflación y la inversión [por ejemplo, Blomström, Lipsey y Zejan (1996)]. Quizá se podrían utilizar datos de series de tiempo de un país que encuentre interesante. O, es posible usar un corte transversal de países, como en De Long y Summers (1991). Friedman y Kuttner (1992) encontraron evidencia de que, al menos en la década de los ochenta, el diferencial entre la tasa de documentos comerciales y la tasa de los bonos de del Tesoro afectaron la producción real.
21. ¿Cuál es el comportamiento de las fusiones en la economía estadounidense (o de alguna otra economía)? Shughart y Tollison (1984) caracterizan (el logaritmo de) las fusiones anuales en la economía estadounidense como una caminata aleatoria al mostrar que la diferencia entre los logaritmos —en términos generales, la tasa de crecimiento— es impredecible dadas las tasas de crecimiento pasadas. ¿Esto aún es válido? ¿Aplica en diversas industrias? ¿Qué mediciones de la actividad económica se pueden usar para pronosticar las fusiones?
22. ¿Qué factores podrían explicar las diferencias raciales, y de género, en el empleo y los salarios? Por ejemplo, Holzer (1991) revisó la evidencia de "la hipótesis de desigualdad espacial" para explicar las diferencias en las tasas de empleo entre negros y blancos. Korenman y Neumark (1992) examinaron los efectos de la crianza infantil en los salarios de las mujeres, mientras que Hersch y Stratton (1997) observaron los efectos de las responsabilidades en el hogar sobre las mujeres y los hombres.
23. Obtenga datos mensuales o trimestrales sobre las tasas de empleo de adolescentes, el salario mínimo y los factores que afectan el empleo adolescente para estimar los efectos del salario mínimo sobre el empleo de los adolescentes. Solon (1985) utilizó datos estadounidenses trimestrales, mientras que Castillo-Freeman y Freeman (1992) emplearon datos anuales sobre Puerto Rico. Podría ser de utilidad analizar datos de series de tiempo en un estado con bajos salarios en Estados Unidos, donde los cambios en el salario mínimo tiendan a tener un efecto mayor.
24. A nivel de ciudad, estime un modelo de series de tiempo para el delito. Un ejemplo es Cloninger y Sartorius (1979). Como giro reciente, puede estimar los efectos de la vigilancia vecinal y los programas de básquetbol de media noche, políticas relativamente nuevas para la lucha contra la delincuencia. Puede ser engañoso inferir causalidad aquí. Incluir una variable dependiente rezagada puede ser útil. Puesto que se están usando datos de series de tiempo, se debe estar consciente del problema de regresión espuria.

    Grogger (1990) usó datos sobre recuentos de homicidios diarios para estimar los efectos disuasivos de la pena de muerte. ¿Podría haber otros factores, como las noticias sobre la respuesta letal de parte de la policía, que ejerzan un efecto en el número de delitos cotidianos?
25. ¿Existen efectos agregados de productividad del empleo de computadoras? Es necesario obtener datos de series de tiempo, quizás a nivel nacional sobre la productividad, el porcentaje de empleados que utiliza las computadoras, y otros factores. ¿Qué hay del gasto (probablemente como una fracción de las ventas totales) en la investigación y el desarrollo? ¿Qué factores sociológicos (por ejemplo, el consumo de alcohol o las tasas de divorcio) podrían afectar la productividad?
26. ¿Qué factores afectan los salarios del director general? Los archivos CEOSAL1.RAW y CEOSAL2.RAW son datos que contienen varias medidas de desempeño de una empresa, así como información sobre la antigüedad y la educación. Desde luego, se pueden actualizar estos archivos y buscar otros factores interesantes. Rose y Shepard (1997) consideraron la diversificación de la empresa como un determinante importante en la compensación del director general.
27. ¿Las diferencias en los códigos fiscales de los estados afectan la cantidad de inversión directa? Hines (1996) estudió los efectos de los impuestos corporativos estatales, junto con la capacidad de aplicar créditos fiscales extranjeros, sobre la inversión extranjera en Estados Unidos.
28. ¿Qué factores influyen en los resultados electorales? ¿El gasto importa? ¿Importan los votos sobre cuestiones específicas? ¿El estado de la economía local importa? Vea, por ejemplo,

Levitt (1994) y las bases de datos VOTE1.RAW y VOTE2.RAW. Fair (1996) desarrolló un análisis de series de tiempo de las elecciones presidenciales estadounidenses.

29. Pruebe si las tiendas o restaurantes practican la discriminación de precios con base en la raza o la etnicidad. Graddy (1997) usó datos de los restaurantes de comida rápida en Nueva Jersey y Pensilvania, junto con características a nivel de código postal, para ver si los precios variaban según las características de la población local. Encontró que los precios de artículos estándar, como las bebidas refrescantes, aumentaban cuando la fracción de residentes de color aumentaba. (Sus datos están contenidos en el archivo DISCRIM.RAW.) Es posible recabar datos similares en su área local al entrevistar a empleados de tiendas y restaurantes para saber los precios de los artículos comunes, y compararlos con los datos del censo reciente. Vea el trabajo de Graddy para detalles sobre su análisis.
30. Haga un estudio de "auditoría" para probar la discriminación por raza o género en las contrataciones. (Un estudio de esta naturaleza se describió en el ejemplo C.3 del apéndice C.) Haga que parejas de amigos igualmente calificados, por ejemplo, un hombre y una mujer, soliciten trabajo para puestos de trabajo en bares o restaurantes locales. Se les puede dar currículos falsos que indiquen la misma experiencia y antecedentes, y donde la única diferencia sea el sexo (o la raza). Después, puede dar seguimiento para saber quién obtiene las entrevistas y las ofertas de trabajo. Neumark (1996) describió uno de esos estudios realizado en Filadelfia. Una variante sería probar si el atractivo físico general o una característica específica, como padecer de obesidad o tener tatuajes visibles o perforaciones corporales, tiene algún impacto sobre las decisiones de contratación. Tal vez se quiera usar el mismo sexo en las parejas que elija, y puede no ser fácil conseguir voluntarios para tal estudio.

LISTA DE PUBLICACIONES

A continuación se presenta una lista parcial de publicaciones conocidas que contienen investigaciones empíricas en los negocios, la economía y otras ciencias sociales. En Internet puede hallarse una lista más completa de publicaciones en http://www.econolit.org.

American Economic Review
American Journal of Agricultural Economics
American Political Science Review
Applied Economics
Brookings Papers on Economic Activity
Canadian Journal of Economics
Demography
Economic Development and Cultural Change
Economic Inquiry
Economica
Economics Letters
Empirical Economics
Federal Reserve Bulletin
International Economic Review
International Tax and Public Finance
Journal of Applied Econometrics
Journal of Business and Economic Statistics
Journal of Development Economics
Journal of Economic Education

Journal of Empirical Finance
Journal of Environmental Economics and Management
Journal of Finance
Journal of Health Economics
Journal of Human Resources
Journal of Industrial Economics
Journal of International Economics
Journal of Labor Economics
Journal of Monetary Economics
Journal of Money, Credit and Banking
Journal of Political Economy
Journal of Public Economics
Journal of Quantitative Criminology
Journal of Urban Economics
National Bureau of Economic Research Working Papers Series
National Tax Journal
Public Finance Quarterly
Quarterly Journal of Economics
Regional Science & Urban Economics
Review of Economic Studies
Review of Economics and Statistics

FUENTES DE DATOS

En todo el mundo existen numerosas fuentes de datos. Los gobiernos de la mayoría de los países compilan una gran cantidad de datos; algunas fuentes generales y de fácil acceso en Estados Unidos, como *Economic Report of the President*, el *Statistical Abstract of the United States* y el *County and City Data Book*, ya se mencionaron. Los datos financieros internacionales sobre numerosos países se publican cada año en *International Financial Statistics*. Varias revistas, como *BusinessWeek* y *U.S. News and World Report*, suelen publicar estadísticas, como sueldos de directores generales y desempeño de la empresa o califican programas académicos, que son nuevos y se pueden usar en un análisis econométrico.

En lugar de intentar ofrecer aquí una lista, se dan algunas direcciones de Internet que son fuentes exhaustivas para los economistas. Un sitio muy útil para los economistas, llamado *Resources for Economists on the Internet*, lo mantiene Bill Goffe en SUNY, Oswego. La dirección es

http://www.rfe.org.

Este sitio ofrece vínculos a revistas, fuentes de datos y listas de economistas profesionales y académicos. Es muy fácil de usar.

La sección de estadísticas económicas y de negocios de la *American Statistical Association* contiene una lista muy detallada de las fuentes de datos y ofrece vínculos a ellas. La dirección es

http://www.econ-datalinks.org.

Además, la *Journal of Applied Econometrics* y la *Journal of Business and Economic Statistics* tienen archivos de datos que contienen bases de datos usados en la mayoría de los documentos publicados en las revistas durante varios años. Si usted encuentra una base de datos que le interese, esta es una buena forma de empezar, puesto que gran parte del trabajo de limpieza y formateo ya se hizo. La desventaja es que algunas de estas bases de datos se utilizan en análisis econométricos que son más avanzados que lo que se ha aprendido en este libro. Por otra parte, suele ser útil estimar modelos más simples usando métodos econométricos de comparación.

Muchas universidades como las de California-Berkeley, Michigan y de Maryland, mantienen bases de datos muy amplias, así como vínculos a una variedad de ellos. Su propia biblioteca quizá contenga un amplio conjunto de vínculos a bases de datos comerciales, económicas y de otras ciencias sociales. El *National Bureau of Economic Research* publica bases de datos que utilizan algunos de sus investigadores. Los gobiernos federales y estatales ahora publican una gran cantidad de datos a la que se puede tener acceso a través de Internet. Los datos del censo están disponibles en el *Census Bureau* estadounidense. (Dos publicaciones útiles son el *Economic Census*, publicado en años que terminan con dos y siete, y el *Census of Population and Housing*, publicado al principio de cada década.) Otras agencias, como el Departamento Estadounidense de Justicia, también pone datos a disposición del público.

Apéndice A

Herramientas matemáticas básicas

En este apéndice se cubren algunas bases matemáticas necesarias en el análisis econométrico. Se resumen las diferentes propiedades del operador de suma, se estudian las propiedades de las ecuaciones lineales y algunas no lineales; se repasan las proporciones y los porcentajes. También se presentan algunas funciones especiales que suelen presentarse en econometría aplicada, incluidas las funciones cuadráticas y el logaritmo natural. Las primeras cuatro secciones requieren sólo habilidades algebraicas básicas. La sección A.5 contiene un breve repaso del cálculo diferencial; aunque los conocimientos de cálculo no son necesarios para comprender la mayor parte del libro, éstos se utilizan en algunos apéndices de final de capítulo, y en varios de los capítulos más avanzados de la parte 3.

A.1 El operador de suma y la estadística descriptiva

El **operador de suma** es una abreviatura útil para manejar expresiones que comprenden sumas de grandes cantidades de números, y desempeña una función clave en estadística y en el análisis econométrico. Si $\{x_i: i = 1, \ldots, n\}$ denota una secuencia de n números, entonces se escribe la suma de estos números como

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n. \qquad \textbf{A.1}$$

Con esta definición, se demuestra fácilmente que el operador de suma tiene las siguientes propiedades:

Propiedad de la suma 1: Para toda constante c,

$$\sum_{i=1}^{n} c = nc. \qquad \textbf{A.2}$$

Propiedad de la suma 2: Para toda constante c,

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i. \qquad \textbf{A.3}$$

Propiedad de la suma 3: Si $\{(x_i, y_i): i = 1, 2, \ldots, n\}$ es un conjunto de n pares de números, y a y b son constantes, entonces

$$\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i. \tag{A.4}$$

También es importante tener en cuenta algunas cuestiones que no son válidas con el operador de suma. Sea $\{(x_i, y_i): i = 1, 2, \ldots, n\}$ nuevamente un conjunto de n pares de números con $y_i \neq 0$ para cada i. Entonces,

$$\sum_{i=1}^{n} (x_i/y_i) \neq \left(\sum_{i=1}^{n} x_i\right) \Big/ \left(\sum_{i=1}^{n} y_i\right).$$

En otras palabras, la suma de las razones no es la razón de las sumas. En el caso $n = 2$, la aplicación del álgebra elemental conocida también revela esta desigualdad: $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. Asimismo, la suma de los cuadrados no es el cuadrado de la suma: $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$, salvo en casos especiales. Que estas dos cantidades no sean generalmente iguales es fácil de constatar cuando $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Dados n números $\{x_i: i = 1, \ldots, n\}$, se calcula su **promedio** o *media* sumándolos, y después dividiendo el resultado entre n:

$$\bar{x} = (1/n) \sum_{i=1}^{n} x_i. \tag{A.5}$$

Cuando las x_i son una muestra de datos sobre una variable determinada (como los años de educación), se suele llamar a esto el *promedio* (o *media*) *muestral* para enfatizar que se calcula a partir de un conjunto determinado de datos. El promedio muestral es un ejemplo de una **estadística descriptiva**; en este caso, el estadístico describe la tendencia central del conjunto de puntos x_i.

Existen algunas propiedades básicas, importantes de entender, acerca de los promedios. Primero, suponga que se toma cada observación sobre x y se resta el promedio: $d_i \equiv x_i - \bar{x}$ (la "d" aquí representa la *desviación* a partir del promedio). Entonces, la suma de estas desviaciones siempre es cero:

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Se resume esto como

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \tag{A.6}$$

Un simple ejemplo numérico muestra cómo funciona esto. Suponga que $n = 5$ y $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$ y $x_5 = 5$. Por tanto, $\bar{x} = 2$, y la muestra centrada es $\{4, -1, -4, -2, 3\}$. Al sumar los datos se obtiene cero, que es justo lo que indica la ecuación (A.6).

En el análisis de la regresión lineal del capítulo 2, se necesitaba conocer algunos hechos algebraicos adicionales, que comprenden las desviaciones de los promedios muestrales. Uno importante es que la suma de las desviaciones cuadradas es la suma de las x_i cuadradas menos n veces el cuadrado de $\bar{x}$:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2. \tag{A.7}$$

Esto puede mostrarse mediante las propiedades básicas del operador de suma:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\\ &= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n(\bar{x})^2\\ &= \sum_{i=1}^{n}x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2.\end{aligned}$$

Dado un conjunto de datos sobre dos variables, $\{(xi, yi): i = 1, 2, \ldots, n\}$, se demuestra también que

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}x_i(y_i - \bar{y})\\ &= \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}x_iy_i - n(\bar{x}\cdot\bar{y});\end{aligned} \qquad \textbf{A.8}$$

ésta es una generalización de la ecuación (A.7). (Ahí, $y_i = x_i$ para toda i.)

El promedio es la medida de la tendencia central, eje de la mayor parte de este libro. No obstante, algunas veces es útil usar la **mediana** (o *mediana muestral*) para describir el valor central. Para obtener la mediana de n números $\{x_1, \ldots, x_n\}$, primero se ordenan los valores de x_i del menor al mayor. Entonces, si n es impar, la mediana muestral es el número intermedio de las observaciones ordenadas. Por ejemplo, dados los números $\{-4, 8, 2, 0, 21, -10, 18\}$, el valor de la mediana es 2 (porque la secuencia ordenada es $\{-10, -4, 0, 2, 8, 18, 21\}$). Si se cambia el número más grande en esta lista, 21, al doble de su valor, 42, la mediana aún será 2. Por otra parte, el promedio muestral aumentaría de 5 a 8, un importante cambio. En general, la mediana es menos sensible que el promedio a los cambios en los valores extremos (ya sean grandes o pequeños) en una lista de números. Ésta es la razón de que los "ingresos medianos" o "valores medianos de la vivienda" suelen ser los que se reportan y no los promedios cuando se resume el ingreso o los valores de las viviendas en una ciudad o país.

Si n es par, no hay una forma única de definir la mediana debido a que existen dos números en el centro. Por lo general, la mediana se define como el promedio de los dos valores intermedios (una vez más, después de ordenar los números del menor al mayor). Mediante esta regla, la mediana para el conjunto de números $\{4, 12, 2, 6\}$ sería $(4 + 6)/2 = 5$.

A.2 Propiedades de las funciones lineales

Las funciones lineales desempeñan un importante papel en econometría, pues son fáciles de interpretar y manipular. Si x y y son dos variables relacionadas por

$$y = \beta_0 + \beta_1 x, \qquad \textbf{A.9}$$

entonces se dice que y es una **función lineal** de x y que β_0 y β_1 son dos parámetros (números) que describen esta relación. El **intercepto** es β_0 y la **pendiente** es β_1.

La característica peculiar de una función lineal es que el cambio en y siempre es β_1 veces el cambio en x:

$$\Delta y = \beta_1 \Delta x, \qquad \textbf{A.10}$$

donde Δ denota el "cambio". En otras palabras, **el efecto marginal** de x sobre y es constante e igual a β_1.

Ejemplo A.1

[Función lineal del gasto en vivienda]

Suponga que la relación entre *housing* (gasto en vivienda mensual) e *income* (ingreso mensual) es

$$vivienda = 164 + .27\ ingreso. \qquad \textbf{A.11}$$

Entonces, por cada dólar adicional de ingreso, se gastan 27 centavos en vivienda. Si el ingreso familiar aumenta en \$200, entonces el gasto en vivienda aumenta en (.27)200 = \$54. Esta función se grafica en la figura A.1.

Con base en la ecuación (A.11), una familia sin ingreso gasta \$164 en vivienda, lo cual, por supuesto, no puede ser literalmente cierto. Para niveles bajos de ingreso, esta función lineal no describiría bien la relación entre vivienda e ingreso, lo cual es la razón por la que eventualmente sería necesario utilizar otros tipos de funciones para describir tales relaciones.

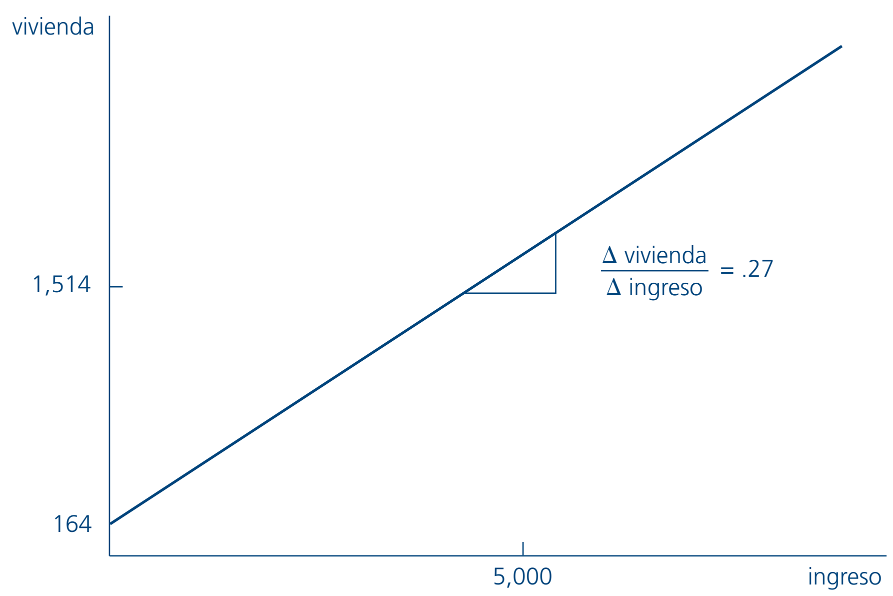
En (A.11) la *propensión marginal al consumo* (PMC) en vivienda a partir del ingreso es .27. Esto es diferente a la *propensión promedio al consumo* (PPC), la cual es

$$\frac{vivienda}{ingreso} = 164/income + .27.$$

La PPC no es constante; siempre es mayor que la PMC y se acerca a ésta a medida que el ingreso aumenta.

FIGURA A.1

Gráfica de *vivienda* = 164 + .27 *ingreso*.



Las funciones lineales se definen fácilmente para más de dos variables. Suponga que y está relacionada con dos variables, x_1 y x_2, en la forma general

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{A.12}$$

Es muy difícil visualizar esta función debido a que su gráfica es tridimensional. No obstante, β_0 aún es el intercepto (el valor de y cuando $x_1 = 0$ y $x_2 = 0$) y β_1 y β_2 miden pendientes particulares. A partir de (A.12), el cambio en y para los cambios dados en x_1 y x_2, es

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \tag{A.13}$$

Si x_2 no cambia, es decir, $\Delta x_2 = 0$, entonces se tiene

$$\Delta y = \beta_1 \Delta x_1 \text{ si } \Delta x_2 = 0,$$

así que β_1 es la pendiente de la relación en la dirección de x_1:

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ si } \Delta x_2 = 0.$$

Debido a que mide cómo cambia y respecto a x_1, manteniendo fija a x_2, β_1 suele llamarse el **efecto parcial** de x_1 sobre y. Puesto que el efecto parcial implica mantener los demás factores fijos, está estrechamente vinculado con la noción de mantener todo lo demás constante (***ceteris paribus***). El parámetro β_2 tiene una interpretación similar: $\beta_2 = \Delta y/\Delta x_2$ si $\Delta x_1 = 0$, así que β_2 es el efecto parcial de x_2 sobre y.

Ejemplo A.2

[Demanda de discos compactos]

Para los estudiantes universitarios, suponga que la cantidad demandada de discos compactos cada mes (*quantity*) está relacionada con el precio (*price*) de los discos compactos y el ingreso discrecional mensual (*income*) por

$$quantity = 120 - 9.8\ price + .03\ income,$$

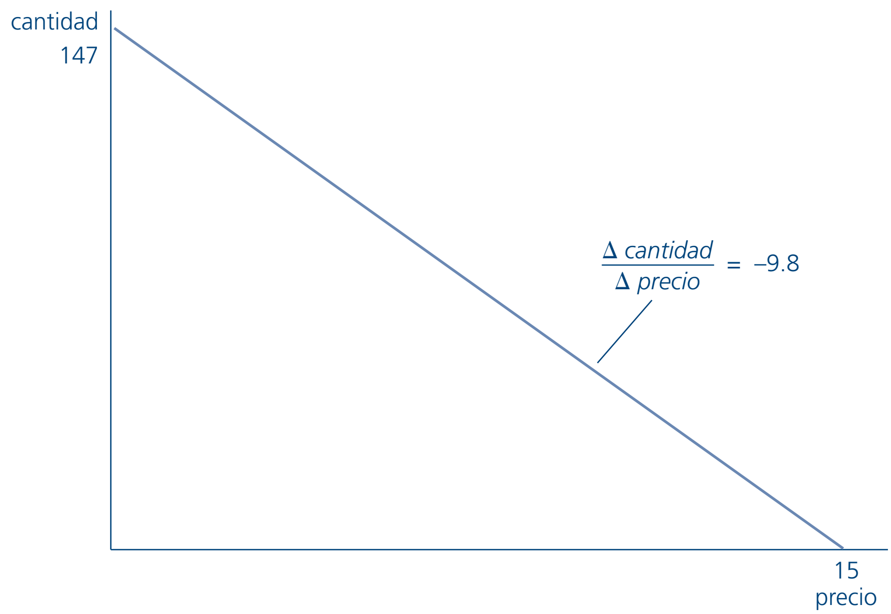
donde *price* es igual a los dólares por disco e *income* se mide en dólares. La *curva de demanda* es la relación entre *quantity* y *price*, manteniendo *income* (y otros factores) fijo. Ésta se grafica, en dos dimensiones, en la figura A.2 a un nivel de ingreso de $900. La pendiente de la curva de la demanda, −9.8, es el *efecto parcial* del precio sobre la cantidad: manteniendo el ingreso fijo, si el precio de los discos compactos aumenta un dólar, entonces la cantidad demandada disminuye 9.8. (Se abstrae del hecho de que los discos compactos sólo se pueden comprar en cantidades discretas.) Un incremento en el ingreso simplemente desplaza hacia arriba la curva de demanda (cambia el intercepto), pero la pendiente sigue siendo la misma.

A.3 Proporciones y porcentajes

Las proporciones y los porcentajes desempeñan un papel tan importante en la economía aplicada que es necesario sentirse muy cómodo trabajando con ellos. Numerosas cantidades reportadas en la prensa popular están en forma de porcentajes; algunos ejemplos son las tasas de interés, las tasas de desempleo y las tasas de graduación de bachillerato.

FIGURA A.2

Gráfica de *cantidad* = 120 − 9.8 *precio* + .03 *ingreso*, con *ingreso* fijo a $900.



Una habilidad importante es poder convertir las proporciones en porcentajes y viceversa. Un porcentaje se puede obtener fácilmente si se multiplica por 100 una proporción. Por ejemplo, si la proporción de adultos en un país con grado de bachillerato es de .82, entonces se dice que 82% de los adultos tienen un grado de bachillerato. Otra forma de entender los porcentajes y las proporciones es que una proporción es una forma decimal de un porcentaje. Por ejemplo, si la tasa impositiva marginal para una familia que gana $30,000 al año se reporta como 28%, entonces la proporción del siguiente dólar de ingreso que se paga en impuestos sobre la renta es .28 (o 28¢).

Cuando se utilizan porcentajes, con frecuencia es necesario convertirlos a la forma decimal. Por ejemplo, si el impuesto estatal sobre las ventas es 6% y se gastan $200 en un artículo gravable, entonces el impuesto que se paga sobre las ventas es = $12. Si el rendimiento anual de un certificado de depósito (CD) es de 7.6% y se invierten $3,000 en tal CD al principio del año, entonces los intereses serán de 3,000 (.076) = $228. Por más que quisiéramos, los intereses no se obtienen multiplicando 3,000 por 7.6.

Se debe tener cuidado con las proporciones, que en ocasiones, están reportadas de manera equivocada, como porcentajes, en los medios de comunicación. Si se lee: "El porcentaje de estudiantes de bachillerato que consumen alcohol es de .57", se sabe que esto en realidad significa 57% (no sólo la mitad de un porcentaje, como implica literalmente el enunciado). Los entusiastas del voleibol universitario quizá estén familiarizados con los reportes noticiosos que afirman "su porcentaje de boleo fue de .372". Esto significa que su porcentaje de boleo fue en realidad de 37.2%.

En econometría, con frecuencia se miden los *cambios* en diferentes cantidades. Sea x alguna variable, como el ingreso de un individuo, el número de delitos cometidos en una comunidad o las utilidades de una empresa. Sean x_0 y x_1 dos valores para x: x_0 es el valor inicial y x_1 es el valor subsiguiente. Por ejemplo, x_0 podría ser el ingreso anual de un individuo en 1994 y x_1 el ingreso del mismo individuo en 1995. El **cambio proporcional** en x cuando varía de x_0 a x_1, en ocasiones recibe el nombre de **cambio relativo,** simplemente es

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \tag{A.14}$$

si se supone por supuesto que $x_0 \neq 0$. En otras palabras, para obtener el cambio proporcional, simplemente se divide el cambio en x entre su valor inicial. Ésta es una forma de estandarizar el cambio de manera que esté libre de unidades. Por ejemplo, si el ingreso de un individuo va de \$30,000 a \$36,000 anuales, entonces el cambio proporcional es de 6,000/30,000 = .20.

Es más común enunciar los cambios en términos de porcentajes. El **cambio porcentual** en x que va de x_0 a x_1 es simplemente 100 veces el cambio proporcional:

$$\%\Delta x = 100(\Delta x/x_0); \tag{A.15}$$

la notación "$\%\Delta x$" se lee como "el cambio porcentual en x". Por ejemplo, cuando el ingreso va de \$30,000 a \$33,750, el ingreso ha aumentado 12.5%; para obtener esto, simplemente se multiplica el cambio proporcional .125 por 100.

Una vez más, es necesario tener cuidado con los cambios proporcionales que se reportan como cambios porcentuales. En el ejemplo anterior, por ejemplo, reportar el cambio porcentual en el ingreso como .12 es incorrecto y podría ocasionar confusiones.

Cuando se observan cambios en cuestiones como cantidades de dólares o población, no hay ambigüedad en cuanto a qué significa el cambio porcentual. Por el contrario, interpretar cálculos de cambios porcentuales puede ser confuso cuando la variable de interés misma es un porcentaje, algo que sucede con frecuencia en economía y otras ciencias sociales. Para ilustrarlo, sea x el porcentaje de adultos con educación universitaria en una ciudad particular. Suponga que el valor inicial es $x_0 = 24$ (24% tiene educación universitaria) y el nuevo valor es $x_1 = 30$. Se pueden calcular dos cantidades para describir cómo ha cambiado el porcentaje de personas con educación universitaria. El primer cálculo es el cambio en x, Δx. En este caso, $\Delta x = x_1 - x_0 = 6$: el porcentaje de personas con alguna educación universitaria, ha aumentado seis *puntos porcentuales*. Por otra parte, se puede calcular el cambio porcentual en x mediante la ecuación (A.15): $\%\Delta x = 100[(30 - 24)/24] = 25$.

En este ejemplo, el cambio en puntos porcentuales y el cambio porcentual son muy diferentes. El **cambio en puntos porcentuales** es sólo el cambio que se presenta en los porcentajes. El cambio porcentual es el cambio relativo al valor inicial. En general, es necesario poner mucha atención al número que se está calculando. Un buen investigador hará una distinción perfecta entre estos dos cambios; por desgracia, en la prensa popular, así como en las investigaciones académicas, el tipo de cambio reportado suele ser confuso.

Ejemplo A.3

[Aumento en el impuesto a las ventas en Michigan]

En marzo de 1994, los votantes de Michigan aprobaron un incremento en el impuesto a las ventas, pasando de 4 a 6%. En la campaña política, los partidarios de esta medida se refirieron a ella como un incremento de dos puntos porcentuales, o un incremento de dos centavos sobre el dólar. Los oponentes al incremento fiscal lo denominaron como un incremento de 50% en la tasa fiscal a las ventas. Ambas afirmaciones son correctas, pues tan sólo son formas diferentes de medir el incremento. Naturalmente, cada grupo reportó la medida que favorecía más su posición.

Para una variable tal como el salario no tiene sentido hablar de un "cambio de punto porcentual", pues el salario no se mide como porcentaje. Se puede describir una variación en el salario en términos de dólares o porcentajes.

A.4 Algunas funciones especiales y sus propiedades

En la sección A.2, se revisaron las propiedades básicas de las funciones lineales. También se indicó una característica importante de las funciones como $y = \beta_0 + \beta_1 x$: un cambio unitario en x genera el *mismo* cambio en y, sin importar el valor inicial de x. Como se observó antes, es lo mismo al decir que el efecto marginal de x en y es constante, algo que no es realista para muchas relaciones económicas. Por ejemplo, la importante noción económica de *rendimientos marginales decrecientes* no es consistente con una relación lineal.

Para representar una variedad de fenómenos económicos, es necesario estudiar varias funciones no lineales. Una **función no lineal** se caracteriza por el hecho de que el cambio en y para un cambio dado en x depende del valor inicial de x. Ciertas funciones no lineales aparecen, con frecuencia, en la economía empírica, así que es importante saber cómo interpretarlas. Una comprensión detallada de las funciones no lineales pertenece al ámbito del cálculo. Aquí, simplemente se resumirán algunos de los aspectos más importantes de las funciones y se dejarán los detalles de algunas derivaciones para la sección A.5.

Funciones cuadráticas

Una forma simple de capturar los rendimientos decrecientes es sumar un término cuadrático a una relación lineal. Considere la ecuación

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \qquad \textbf{A.16}$$

donde β_0, β_1 y β_2 son parámetros. Cuando $\beta_1 > 0$ y $\beta_2 < 0$, la relación entre y y x tiene la forma parabólica como se muestra en la figura A.3, donde $\beta_0 = 6$, $\beta_1 = 8$ y $\beta_2 = -2$.

Cuando $\beta_1 > 0$ y $\beta_2 < 0$, se puede demostrar (mediante el cálculo, en la siguiente sección) que el *máximo* de la función tiene lugar en el punto

$$x^* = \beta_1/(-2\beta_2). \qquad \textbf{A.17}$$

Por ejemplo, si $y = 6 + 8x - 2x^2$ (así que $\beta_1 = 8$ y $\beta_2 = -2$), entonces el valor más grande de y tiene lugar en $x^* = 8/4 = 2$ y este valor es $6 + 8(2) - 2(2)^2 = 14$ (véase la figura A.3).

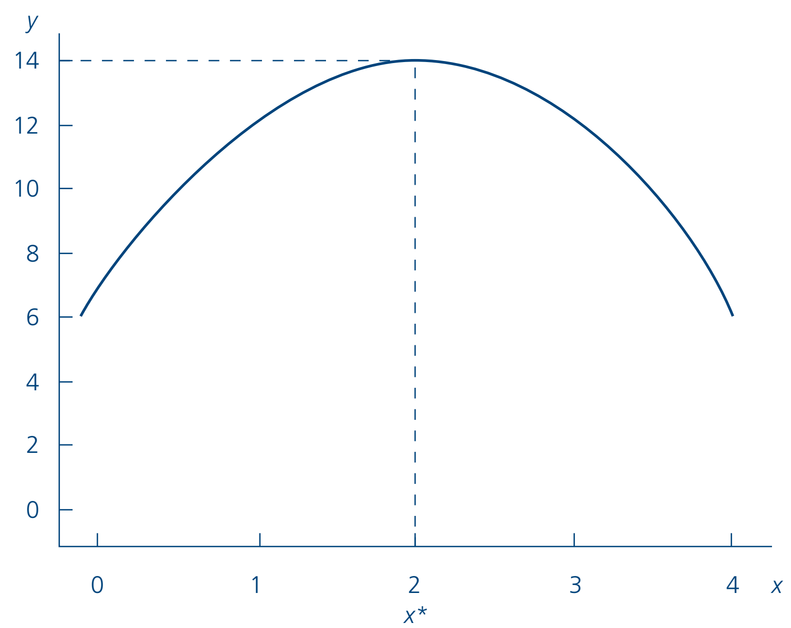
El hecho de que la ecuación (A.16) implique un **efecto marginal decreciente** de x en y se puede ver fácilmente en esta gráfica. Suponga que se comienza con un valor pequeño de x y después x aumenta en cierta cantidad, por ejemplo c. Esto tiene un efecto mayor en y que si se comenzara con un valor más alto de x y se aumentara x en la misma cantidad c. De hecho, una vez que $x > x^*$, un incremento en x hace que realmente y disminuya.

La afirmación de que x tiene un efecto marginal decreciente sobre y es lo mismo que decir que la pendiente de la función en la figura A.3 disminuye a medida que x aumenta. Aunque esto es evidente al observar la gráfica, se suele cuantificar la rapidez con que cambia la pendiente. Una aplicación del calculo da la pendiente aproximada de la función cuadrática como

$$slope = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x, \qquad \textbf{A.18}$$

FIGURA A.3

Gráfica de $y = 6 + 8x - 2x^2$.



para cambios "pequeños" en x. [El lado derecho de la ecuación (A.18) es la **derivada** de la función en la ecuación (A.16) respecto a x.] Otra forma de escribir esto es

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \text{ para } \Delta x \text{ "pequeñas".} \quad \textbf{A.19}$$

Para ver si funciona bien esta aproximación, considere nuevamente la función $y = 6 + 8x - 2x^2$. Entonces, con base en la ecuación (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Ahora suponga que se empieza en $x = 1$ y x se sustituye por $\Delta x = .1$. Utilizando (A.19), $\Delta y \approx (8 - 4)(.1) = .4$. Por supuesto, se puede calcular el cambio exactamente si se encuentran los valores de y cuando $x = 1$ y $x = 1.1$: $y_0 = 6 + 8(1) - 2(1)^2 = 12$ y $y_1 = 6 + 8(1.1) - 2(1.1)^2 = 12.38$, de manera que el cambio exacto en y es .38. La aproximación es muy cercana en este caso.

Ahora suponga que se comienza en $x = 1$ pero que x se sustituye por una cantidad mayor: $\Delta x = .5$. Entonces, la aproximación resulta $\Delta y \approx 4(.5) = 2$. El cambio exacto se determina al hallar la diferencia en y cuando $x = 1$ y $x = 1.5$. El valor anterior de y fue de 12 y el valor posterior es $6 + 8(1.5) - 2(1.5)^2 = 13.5$, así que el cambio real es 1.5 (no 2). La aproximación empeora en este caso debido a que el cambio en x es mayor.

La ecuación (A.19) se puede utilizar en numerosas aplicaciones para calcular el efecto marginal aproximado de x en y para cualquier valor inicial de x y cambios pequeños. Y si fuera necesario, siempre se podrá calcular el cambio exacto.

Ejemplo A.4

[Una función cuadrática del salario]

Suponga que la relación entre salarios por hora (*wage*) y años de experiencia (*exper*) está dada por

$$wage = 5.25 + .48\ exper - .008\ exper^2. \tag{A.20}$$

Esta función tiene la misma forma general que aquella de la figura A.3. Mediante la ecuación (A.17), *exper* tiene un efecto positivo sobre el salario hasta que se llega a un punto coyuntural, $exper^* = .48/[2(.008)] = 30$. El primer año de experiencia vale aproximadamente .48 ó 48 centavos [vea (A.19) con $x = 0$, $\Delta x = 1$]. Cada año adicional de experiencia aumenta el salario menos que el año anterior, lo cual refleja un rendimiento marginal decreciente respecto a la experiencia. A los 30 años, un año adicional de experiencia en realidad ocasionaría una disminución en el salario. Esto no es muy realista, pero es una de las consecuencias de emplear una función cuadrática para capturar un efecto marginal decreciente: en cierto punto, la función debe alcanzar un máximo y descender. Para fines prácticos, el punto donde esto sucede suele ser lo bastante grande como para carecer de importancia, pero no siempre.

La gráfica de la función cuadrática en (A.16) tiene forma de U si $\beta_1 < 0$ y $\beta_2 > 0$, en cuyo caso existe un rendimiento marginal creciente. El punto mínimo de la función es $-\beta_1/(2\beta_2)$.

Logaritmo natural

La función no lineal que desempeña el papel más importante en el análisis econométrico es el **logaritmo natural**. En este libro se denota el logaritmo natural simplemente como **función log**, como

$$y = \log(x). \tag{A.21}$$

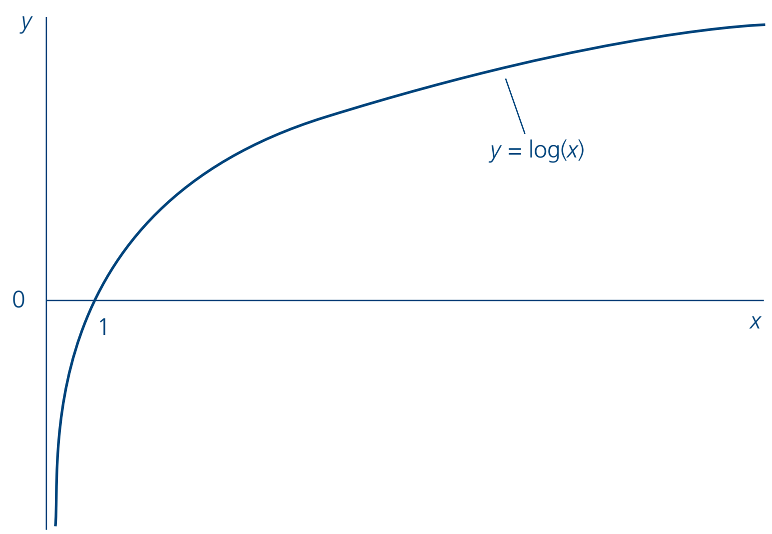
Existen diferentes símbolos del logaritmo natural, $\ln(x)$ o $\log_e(x)$ son los más comunes. Estas diferentes notaciones son útiles cuando se emplean logaritmos con bases diferentes. Para los fines de este libro, sólo es importante el logaritmo natural y, por tanto, $\log(x)$ denota en todo el libro al logaritmo natural. Esto corresponde con el uso en la notación de varios paquetes estadísticos, auque algunos emplean $\ln(x)$ [y la mayoría de las calculadoras utiliza $\ln(x)$]. Los economistas utilizan ambos, $\log(x)$ y $\ln(x)$, lo cual es útil saber cuando se leen trabajos sobre economía aplicada.

La función $y = \log(x)$ se define sólo para $x > 0$, y se grafica en la figura A.4. No es muy útil saber cómo se obtienen los valores de $\log(x)$. Para los fines de este libro, la función puede concebirse como una caja negra: se puede insertar cualquier $x > 0$ y obtener $\log(x)$ de una calculadora o computadora.

Se pueden observar varias cuestiones en la figura A.4. Primero, cuando $y = \log(x)$, la relación entre y y x presenta rendimientos marginales decrecientes. Una diferencia importante entre el log y la función cuadrática en la figura A.3, es que cuando $y = \log(x)$, el efecto de x en y nunca llega a ser negativo: la pendiente de la función se acerca cada vez más a cero a medida que x se hace mayor, pero la pendiente nunca alcanza el cero y, ciertamente, nunca se vuelve negativa.

FIGURA A.4

Gráfica de $y = \log(x)$.



También se puede observar lo siguiente en la figura A.4:

$$\log(x) < 0 \text{ para } 0 < x < 1$$
$$\log(1) = 0$$
$$\log(x) > 0 \text{ para } x > 1.$$

En particular, $\log(x)$ puede ser negativo o positivo. Algunas operaciones algebraicas útiles acerca de la función log son

$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2),\ x_1, x_2 > 0$$
$$\log(x_1/x_2) = \log(x_1) - \log(x_2),\ x_1, x_2 > 0$$
$$\log(x^c) = c\log(x),\ x > 0,\ c \text{ cualquier número.}$$

En ocasiones, será necesario depender de estas propiedades.

El logaritmo se puede utilizar para varias aproximaciones que surgen en las aplicaciones econométricas. Primero, $\log(1 + x) \approx x$ para $x \approx 0$. Se puede intentar esto con $x = .02$, .1 y .5 para ver qué cualidad de la aproximación se deteriora a medida que x crece. Aún más útil es el hecho de que la diferencia en los logaritmos puede utilizarse para aproximar los cambios proporcionales. Sean x_0 y x_1 valores positivos. Entonces, se puede demostrar (mediante cálculo) que

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \qquad \textbf{A.22}$$

para cambios pequeños en x. Si se multiplica la ecuación (A.22) por 100 y se escribe $\Delta\log(x) = \log(x_1) - \log(x_0)$, entonces

$$100 \cdot \Delta\log(x) \approx \%\Delta x \qquad \textbf{A.23}$$

para cambios pequeños en x. El significado de "pequeño" depende del contexto y se encontrarán varios ejemplos a lo largo del libro.

¿Por qué se debe aproximar el cambio mediante (A.23) cuando el cambio porcentual exacto es tan fácil de calcular? Por el momento, se explicará por qué la aproximación en (A.23) es útil en econometría. Primero, vea qué tan buena es la aproximación en dos ejemplos.

Primero, suponga que $x_0 = 40$ y $x_1 = 41$. Entonces, el cambio porcentual en x al pasar de x_0 a x_1 es 2.5% mediante $100(x_1 - x_0)/x_0$. Ahora, $\log(41) - \log(40) = .0247$ a cuatro dígitos, que cuando se multiplica por 100 es muy cercano a 2.5. La aproximación funciona muy bien. Ahora bien, considere un cambio mucho mayor: $x_0 = 40$ y $x_1 = 60$. El cambio porcentual exacto es 50%. No obstante, $\log(60) - \log(40) \approx .4055$, así que la aproximación resulta 40.55%, lo cual está mucho más alejado.

¿Por qué es útil la aproximación con (A.23) si sólo es satisfactoria para cambios pequeños? Para llegar a la respuesta, debe definirse la **elasticidad** de y respecto a x como

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}. \tag{A.24}$$

En otras palabras, la elasticidad de y respecto a x es el cambio porcentual en y cuando x aumenta 1%. Esta noción debe ser familiar derivada de los cursos de introducción a la economía.

Si y es una función lineal de x, $y = \beta_0 + \beta_1 x$, entonces la elasticidad es

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \tag{A.25}$$

lo cual claramente depende del valor de x. (Ésta es una generalización del resultado bien conocido de la teoría de la demanda básica: la elasticidad no es constante a lo largo de la curva de demanda.)

Las elasticidades son de importancia crítica en varias áreas de la economía aplicada, no sólo en la teoría de la demanda. En numerosas situaciones, es conveniente tener modelos de elasticidad constante y la función log permite especificar tales modelos. Si se emplea la aproximación con (A.23) tanto para x como para y, entonces la elasticidad es aproximadamente igual a $\Delta\log(y)/\Delta\log(x)$. Por tanto, un modelo de elasticidad constante se aproxima mediante la ecuación

$$\log(y) = \beta_0 + \beta_1 \log(x), \tag{A.26}$$

y β_1 es la elasticidad de y respecto a x (en el supuesto de que $x, y > 0$).

Ejemplo A.5

[Función de la demanda de elasticidad constante]

Si q es la cantidad demandada y p es el precio y esas variables están relacionadas por

$$\log(q) = 4.7 - 1.25 \log(p),$$

entonces, la elasticidad precio de la demanda es -1.25. Aproximadamente, un incremento de 1% en el precio ocasiona una disminución de 1.25% en la cantidad demandada.

Para los fines de este libro, el hecho de que β_1 en (A.26) sólo se aproxime a la elasticidad carece de importancia. De hecho, cuando la elasticidad se define mediante cálculo, como en la sección A.5, la definición es exacta. Para fines del análisis econométrico, (A.26) define un

modelo de elasticidad constante. Tales modelos desempeñan un importante papel en la economía empírica.

En el trabajo empírico suelen surgir otras posibilidades para emplear la función log. Suponga que $y > 0$ además de

$$\log(y) = \beta_0 + \beta_1 x. \tag{A.27}$$

Entonces, $\Delta\log(y) = \beta_1\Delta x$, así que $100\cdot\Delta\log(y) = (100\cdot\beta_1)\Delta x$. Se desprende que, cuando y y x están relacionadas por la ecuación (A.27),

$$\%\Delta y \approx (100\cdot\beta_1)\Delta x. \tag{A.28}$$

Ejemplo A.6

[Ecuación logarítmica del salario]

Suponga que el salario por hora (*wage*) y los años de educación (*educ*) están relacionados por

$$\log(wage) = 2.78 + .094\ educ.$$

Entonces, mediante la ecuación (A.28),

$$\%\Delta wage \approx 100(.094)\ \Delta educ = 9.4\ \Delta educ.$$

Se sigue que un año más de educación aumenta el salario por hora cerca de 9.4%.

Por lo general, la cantidad $\%\Delta y/\Delta x$ recibe el nombre de **semielasticidad** de y respecto a x. La semielasticidad es el cambio porcentual en y cuando x aumenta una *unidad*. Lo que se acaba de mostrar es que, en el modelo (A.27), la semielasticidad es constante e igual a $100\cdot\beta_1$. En el ejemplo A.6, podría ser conveniente resumir la relación entre salarios y educación al decir que un año más de educación, comenzando desde cualquier cantidad de educación, aumenta el salario aproximadamente 9.4%. Ésta es la razón de que tales modelos sean tan importantes en economía.

Otra relación de interés en economía aplicada es

$$y = \beta_0 + \beta_1\log(x), \tag{A.29}$$

donde $x > 0$. ¿Cómo se puede interpretar esta ecuación? Si se toma el cambio en y, se obtiene $\Delta y = \beta_1\Delta\log(x)$, la cual se puede reescribir como $\Delta y = (\beta_1/100)[100\cdot\Delta\log(x)]$. Por tanto, al emplear la aproximación con (A.23) se tiene

$$\Delta y \approx (\beta_1/100)(\%\Delta x). \tag{A.30}$$

En otras palabras, $\beta_1/100$ es el cambio unitario en y cuando x aumenta 1%.

Ejemplo A.7

[Función de la oferta de mano de obra]

Suponga que la oferta de mano de obra de un trabajador puede describirse como

$$hours = 33 + 45.1\ \log(wage),$$

donde *wage* es el salario por hora y *hours* son las horas trabajadas por semana. Entonces, con base en (A.30),

$$\Delta hours \approx (45.1/100)(\%\Delta wage) = .451\ \%\Delta wage.$$

En otras palabras, un incremento de 1% en el salario aumenta las horas semanales trabajadas por alrededor de .45, o ligeramente menos que media hora. Si el salario aumenta 10%, entonces $\Delta hours = .451(10) = 4.51$, o aproximadamente cuatro y media horas. Esta aproximación no se emplearía para cambios porcentuales mucho mayores en los salarios.

La función exponencial

Antes de dejar esta sección, es necesario analizar una función especial que está relacionada con el logaritmo. Como motivación, considere la ecuación (A.27). Ahí, $\log(y)$ es una función lineal de x. Pero, ¿cómo se calcula la misma y en función de x? La respuesta la da la **función exponencial**.

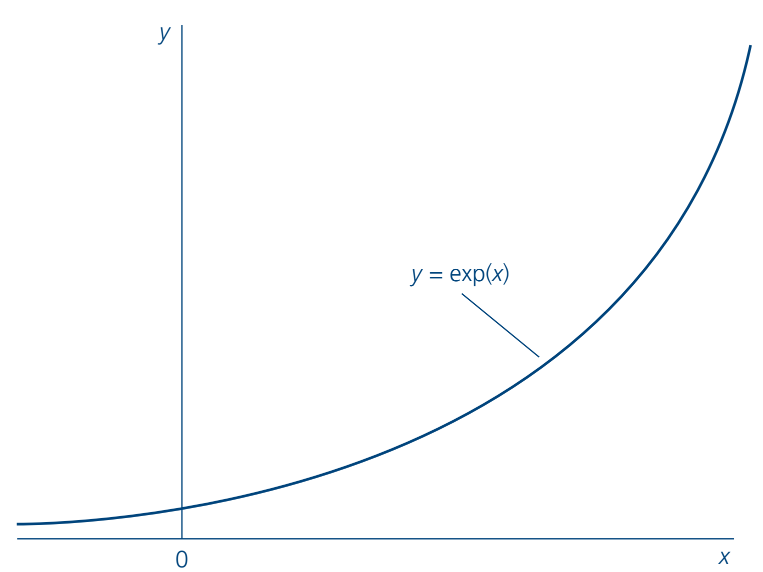
Se escribirá la función exponencial como $y = \exp(x)$, la cual se grafica en la figura A.5. Ahí se observa que $\exp(x)$ está definida para todo valor de x y siempre es mayor que cero. En ocasiones, la función exponencial se escribe como $y = e^x$, pero no se usará esta notación. Dos importantes valores de la función exponencial son: $\exp(0) = 1$ y $\exp(1) = 2.7183$ (a cuatro lugares decimales).

La función exponencial es la inversa de la función logaritmo de la siguiente manera: $\log[\exp(x)] = x$ para toda x y $\exp[\log(x)] = x$ para $x > 0$. En otras palabras, el logaritmo "cancela" la exponencial y viceversa. (Ésta es la razón de que la función exponencial en ocasiones reciba el nombre de función *antilog*.) En particular, observe que $\log(y) = \beta_0 + \beta_1 x$ es equivalente a

$$y = \exp(\beta_0 + \beta_1 x).$$

FIGURA A.5

Gráfica de $y = \exp(x)$.



Si $\beta_1 > 0$, la relación entre x y y tiene la misma forma que en la figura A.5. Por tanto, si $\log(y) = \beta_0 + \beta_1 x$ con $\beta_1 > 0$, entonces x tiene un efecto marginal *creciente* sobre y. En el ejemplo A.6, esto significa que otro año de educación genera un cambio mayor en el salario que el año anterior de educación.

Dos hechos útiles referentes a la función exponencial son $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ y $\exp[c \cdot \log(x)] = x^c$.

A.5 Cálculo diferencial

En la sección anterior se indicaron varias aproximaciones que tienen su fundamento en el cálculo. Sea $y = f(x)$ para alguna función f. Entonces, para los cambios pequeños en x,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \tag{A.31}$$

donde df/dx es la derivada de la función f evaluada en el punto x_0. También la derivada se escribe como dy/dx.

Por ejemplo, si $y = \log(x)$, entonces $dy/dx = 1/x$. Mediante (A.31), con dy/dx evaluada en x_0, se tiene $\Delta y \approx (1/x_0)\Delta x$, o $\Delta\log(x) \approx \Delta x/x_0$, que es la aproximación dada en (A.22).

En econometría aplicada es útil recordar las derivadas de algunas funciones, pues se emplea la derivada para definir la pendiente de una función en cierto punto. Entonces se puede utilizar (A.31) para encontrar el cambio aproximado en y para pequeños cambios en x. En el caso lineal, la derivada simplemente es la pendiente de la recta, como se esperaría: si $y = \beta_0 + \beta_1 x$, entonces $dy/dx = \beta_1$.

Si $y = x^c$, entonces $dy/dx = cx^{c-1}$. La derivada de una suma de dos funciones es la suma de las derivadas: $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. La derivada de una constante multiplicada por cualquier función es la misma constante multiplicada por la derivada de la función: $d[cf(x)]/dx = c[df(x)/dx]$. Estas sencillas reglas permiten encontrar la derivada de funciones más complicadas. Otras reglas, como las del producto, cociente y de la cadena, les serán conocidas a aquellos que hayan tomado cursos de cálculo, pero aquí no se repasarán.

Algunas funciones que suelen utilizarse en economía junto con sus derivadas son

$$y = \beta_0 + \beta_1 x + \beta_2 x^2;\ dy/dx = \beta_1 + 2\beta_2 x$$
$$y = \beta_0 + \beta_1/x;\ dy/dx = -\beta_1/(x^2)$$
$$y = \beta_0 + \beta_1\sqrt{x};\ dy/dx = (\beta_1/2)x^{-1/2}$$
$$y = \beta_0 + \beta_1\log(x);\ dy/dx = \beta_1/x$$
$$y = \exp(\beta_0 + \beta_1 x);\ dy/dx = \beta_1\exp(\beta_0 + \beta_1 x).$$

Si $\beta_0 = 0$ y $\beta_1 = 1$ en esta última expresión se obtiene $dy/dx = \exp(x)$, cuando $y = \exp(x)$.

En la sección A.4, se observa que la ecuación (A.26) define un modelo de elasticidad constante con la ayuda del cálculo. La definición de la elasticidad en cálculo es $(dy/dx) \cdot (x/y)$. Se puede demostrar con las propiedades de logaritmos y exponentes que, cuando (A.26) aplica, $(dy/dx) \cdot (x/y) = \beta_1$.

Cuando y es una función de variables múltiples, la noción de **derivada parcial** cobra importancia. Suponga que

$$y = f(x_1, x_2). \tag{A.32}$$

Entonces, existen dos derivadas parciales, una respecto a x_1 y otra respecto a x_2. La derivada parcial de y respecto a x_1, denotada aquí como $\partial y/\partial x_1$, es sólo la derivada usual de (A.32) respecto a x_1, donde x_2 se trata como una *constante*. Asimismo, $\partial y/\partial x_2$ sólo es la derivada de (A.32) respecto a x_2, manteniendo a x_1 fija.

Las derivadas parciales son útiles por la misma razón que las derivadas ordinarias. Se puede aproximar el cambio en y como

$$\Delta y \approx \frac{\partial y}{\partial x_1} \cdot \Delta x_1, \text{ manteniendo } x_2 \text{ fija.} \qquad \text{A.33}$$

Por tanto, el cálculo permite definir los efectos parciales de los modelos no lineales tal como se podría en los modelos lineales. De hecho, si

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

entonces

$$\frac{\partial y}{\partial x_1} = \beta_1, \; \frac{\partial y}{\partial x_2} = \beta_2.$$

Éstos se pueden reconocer como los efectos parciales definidos en la sección A.2.

Un ejemplo más complicado es

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1 \cdot x_2. \qquad \text{A.34}$$

Ahora la derivada de (A.34) respecto a x_1 (si x_2 se trata como una constante) es simplemente

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2;$$

observe cómo depende de x_1 y x_2. La derivada de (A.34) respecto a x_2 es $\partial y/\partial x_2 = -3 + 7x_1$, así que esto sólo depende de x_1.

Ejemplo A.8

[Función del salario con interacción]

Una función que relaciona los salarios (*wage*) con los años de educación (*educ*) y la experiencia, (*exper*), es

$$\begin{aligned} wage = {} & 3.10 + .41\, educ + .19\, exper - .004\, exper^2 \\ & + .007\, educ \cdot exper. \end{aligned} \qquad \text{A.35}$$

El efecto parcial de *exper* sobre *wage* es la derivada parcial de (A.35):

$$\frac{\partial wage}{\partial exper} = .19 - .008\, exper + .007\, educ.$$

Éste es el cambio aproximado en el salario debido a la experiencia cada vez mayor de un año al otro. Observe que este efecto parcial depende del nivel inicial de *exper* y *educ*. Por ejemplo, un trabajador que comenzó con $educ = 12$ y $exper = 5$, el siguiente año de experiencia aumenta el salario aproximadamente $.19 - .008(5) + .007(12) = .234$, o 23.4 centavos por hora. El cambio exacto puede determinarse al calcular (A.35) con $exper = 5$, $educ = 12$ y con $exper = 6$, $educ = 12$, y después obtener la diferencia. Ésta resulta ser .23, lo cual está muy cerca de la aproximación.

El cálculo diferencial desempeña un importante papel en las funciones de maximización y minimización de una o más variables. Si $f(x_1, x_2, \ldots, x_k)$ es una función diferenciable de k variables, entonces una condición necesaria para $x_1^*, x_2^*, \ldots, x_k^*$ para minimizar o maximizar f sobre todos los valores posibles de x_j es

$$\frac{\partial f}{\partial x_j}(x_1^*, x_2^*, \ldots, x_k^*) = 0, j = 1, 2, \ldots, k. \qquad \textbf{A.36}$$

En otras palabras, todas las derivadas parciales de f deben ser cero cuando se evalúan en x_h^*. Éstas reciben el nombre de *condiciones de primer orden* para maximizar o minimizar una función. En la práctica, se espera resolver la ecuación (A.36) para x_h^*. Entonces, se pueden emplear otros criterios para determinar si se debe maximizar o minimizar la función. Aquí no serán necesarios. [Vea Sydsaeter y Hammond (1995) para un análisis del cálculo multivariado y su uso para optimizar funciones.]

RESUMEN

Las herramientas matemáticas aquí revisadas son cruciales para comprender el análisis de regresión, la probabilidad y la estadística que se cubren en los apéndices B y C. El tema de las funciones no lineales, en especial de las funciones cuadráticas, logarítmicas y exponenciales, es clave para comprender la investigación de la economía aplicada moderna. El nivel de comprensión requerido de estas funciones no incluye conocimientos de cálculo, aunque éste es necesario para ciertas derivaciones.

TÉRMINOS CLAVE

Cambio en puntos porcentuales
Cambio porcentual
Cambio proporcional
Cambio relativo
Ceteris paribus
Derivada
Derivada parcial
Efecto marginal
Efecto marginal decreciente
Efecto parcial
Elasticidad
Estadística descriptiva
Función exponencial
Función lineal
Función log
Función no lineal
Intercepto
Logaritmo natural
Mediana
Modelo de elasticidad constante
Operador de suma
Pendiente
Promedio
Semielasticidad

PROBLEMAS

A.1 La tabla siguiente contiene los gastos mensuales en vivienda de 10 familias.

Familia	Gastos mensuales en vivienda (en dólares)
1	300
2	440
3	350

(*continúa*)

Familia	Gastos mensuales en vivienda (en dólares)
4	1,100
5	640
6	480
7	450
8	700
9	670
10	530

i) Calcule el promedio del gasto mensual en vivienda.
ii) Calcule la mediana del gasto mensual en vivienda.
iii) Si los gastos mensuales en vivienda se midieran en cientos de dólares, no en dólares, ¿cuáles serían la mediana y el promedio de esos gastos?
iv) Suponga que la familia número 8 aumenta su gasto mensual en vivienda a $900 dólares, pero que los gastos de las demás familias siguen siendo los mismos. Calcule el promedio y la mediana de los gastos en vivienda.

A.2 Suponga que la siguiente ecuación describe la relación entre el número promedio de inasistencias a clases durante un semestre (*missed*) y la distancia a la escuela (*distance*, medida en miles):

$$missed = 3 + 0.2\, distance.$$

i) Trace esta línea y asegúrese de nombrar los ejes. ¿Cómo interpreta el intercepto en esta ecuación?
ii) ¿Cuál es el número promedio de inasistencias a clases de alguien que vive a cinco millas de distancia?
iii) ¿Cuál es la diferencia en el número promedio de inasistencias a clases para alguien que vive a 10 millas de distancia y de alguien que vive a 20 millas de distancia?

A.3 En el ejemplo A.2, la cantidad de discos compactos se relacionó con el precio y el ingreso por $quantity = 120 - 9.8\, price + .03\, income$. ¿Cuál es la demanda de CD si $price = 15$ e $income = 200$? ¿Qué sugiere esto acerca de emplear las funciones lineales para describir las curvas de demanda?

A.4 Suponga que la tasa de desempleo en Estado Unidos va de 6.4% en un año a 5.6% al siguiente.
i) ¿Cuál es la disminución en puntos porcentuales en la tasa de desempleo?
ii) ¿En qué porcentaje disminuyó la tasa de desempleo?

A.5 Suponga que el rendimiento de conservar las acciones de una determinada empresa va de 15% en un año a 18% al siguiente. La mayoría de los accionistas afirma que “el rendimiento de las acciones aumentó sólo en 3%”, mientras que el director general afirma que “el rendimiento de las acciones de la empresa ha aumentado 20%”. Concilie estas discrepancias.

A.6 Suponga que la persona A gana $35,000 al año y la persona B gana $42,000.

i) Determine el porcentaje exacto por el cual el salario de la persona B excede al de la persona A.

ii) Ahora use la diferencia en los logaritmos naturales para obtener la diferencia porcentual aproximada.

A.7 Suponga que el siguiente modelo describe las relaciones entre salario anual (*salary*) y el número de años previos de experiencia en el mercado laboral (*exper*):

$$\log(salary) = 10.6 + .027\, exper.$$

i) ¿Cuál es el *salario* cuando $exper = 0$? ¿Cuando $exper = 5$? (*Sugerencia:* necesitará exponenciar.)

ii) Utilice la ecuación (A.28) para aproximar el incremento porcentual en *salary* cuando *exper* aumenta cinco años.

iii) Emplee los resultados de la parte i) para calcular la diferencia porcentual exacta en el salario cuando $exper = 5$ y $exper = 0$. Comente cómo se compara esto con la aproximación en la parte ii).

A.8 Sea *grthemp* el crecimiento proporcional en el empleo, a nivel nacional, de 1990 a 1995 y *salestax*, la tasa fiscal sobre ventas del país, expresado como una proporción. Interprete el intercepto y la pendiente en la ecuación

$$grthemp = .043 - .78\, salestax.$$

A.9 Suponga que el rendimiento de cierto cultivo, *yield* (en fanegas por acre), está relacionado con la cantidad de fertilizante, *fertilizer* (en libras por acre), como

$$yield = 120 + .19\sqrt{fertilizer}.$$

i) Grafique esta relación sustituyendo varios valores para *fertilizer*.

ii) Describa cómo se compara la forma de esta relación con una relación lineal entre *yield* y *fertilizer*.

Apéndice B

Fundamentos de probabilidad

En este apéndice se hará un repaso de los conceptos clave de la probabilidad. Los apéndices B y C pretenden servir como repaso, no como sustitutos de un curso de probabilidad y estadística. Sin embargo, en ellos se revisan todos los conceptos de probabilidad y estadística que se usan en este libro.

La probabilidad por sí misma es de interés para estudiantes de negocios, economía y otras ciencias sociales. Considere, por ejemplo, el caso de una aerolínea que tiene que decidir cuántas reservaciones aceptar para un vuelo en el que hay 100 asientos disponibles. Si hay menos de 100 personas que quieran hacer una reservación, todas serán aceptadas. Pero, ¿qué pasa si hay más de 100 personas que quieran una reservación? Una solución segura es aceptar como máximo 100 reservaciones. Sin embargo, dado que hay personas que hacen una reservación y después no la utilizan, existe la posibilidad de que aun cuando se hayan hecho 100 reservaciones, el avión no se llene. Esto ocasiona pérdidas a la aerolínea. Otra solución puede ser aceptar más de 100 reservaciones esperando que no todas sean utilizadas, de manera que la cantidad final de pasajeros sea lo más cercana posible a 100. Con esta política, la aerolínea corre el riesgo de tener que compensar a las personas que, debido al exceso de reservaciones, no puedan ser aceptadas para el vuelo.

La pregunta natural que surge en este contexto es: ¿se puede determinar la cantidad óptima (o mejor) de reservaciones que deba hacer la aerolínea? Este no es un problema trivial. Sin embargo, dada cierta información (sobre los costos de la aerolínea y qué tan frecuente es que las personas utilicen su reservación), puede emplearse la probabilidad básica para encontrar la solución a este caso.

B.1 Variables aleatorias y sus distribuciones de probabilidad

Suponga que lanza una moneda 10 veces y se cuenta la cantidad de ellas en que se obtiene cara. Esto es un ejemplo de un **experimento**. En general, un experimento es un procedimiento que puede, al menos en teoría, repetirse una cantidad infinita de veces y que tiene un conjunto bien definido de resultados. Lanzar una moneda puede, en principio, repetirse una y otra vez. Antes de lanzarla 10 veces, ya se sabe que la cantidad de caras que puede obtenerse será un número desde 0 hasta 10, de manera que los resultados de este experimento están bien definidos.

Una **variable aleatoria** es aquella que toma un valor numérico que será determinado por un experimento. En el ejemplo del lanzamiento de la moneda, la cantidad de caras en 10 lanzamientos es un ejemplo de una variable aleatoria. Antes de lanzar la moneda 10 veces, no se sabe en cuántas de ellas se obtendrá cara. Una vez lanzada 10 veces y contado el número de caras, se obtiene el valor de la variable aleatoria en este particular ensayo del experimento. En otro ensayo puede obtenerse otro resultado.

En el ejemplo antes citado de las reservaciones de la aerolínea, la cantidad de personas que utiliza su reservación es una variable aleatoria: antes de un determinado vuelo no se sabe cuántas personas utilizan su reservación.

Para analizar los datos recolectados, en negocios y en ciencias sociales, es importante tener un conocimiento básico sobre las variables aleatorias y sus propiedades. De acuerdo con la convención usual en probabilidad y estadística, en los apéndices B y C las variables aleatorias se denotarán empleando letras mayúsculas, por lo general W, X, Y y Z, y los valores particulares de las variables aleatorias empleando las correspondientes letras minúsculas, w, x, y y z. Por ejemplo, en el experimento del lanzamiento de una moneda, sea X la cantidad de caras obtenidas en los 10 lanzamientos de la misma. Entonces, X no está relacionada con ningún valor determinado, pero se sabe que X toma uno de los valores del conjunto $\{0, 1, 2, \ldots, 10\}$. Un valor particular es, por ejemplo, $x = 6$.

Para colecciones grandes de variables aleatorias se usan subíndices. Por ejemplo, si se obtienen los ingresos del último año de 20 hogares de Estados Unidos, elegidos aleatoriamente, estas variables aleatorias se denotarán $X_1, X_2, \ldots, X_{20}$; y los valores particulares se denotarán $x_1, x_2, \ldots, x_{20}$.

Como se establece en la definición, las variables aleatorias siempre están definidas para tomar valores numéricos, aun cuando describan eventos cualitativos. Por ejemplo, considere el lanzamiento de una moneda en el que los dos resultados son cara o cruz. Se puede definir una variable aleatoria como sigue: $X = 1$ si se obtiene cara y $X = 0$ si se obtiene cruz.

A una variable aleatoria que sólo puede tomar los valores 0 o 1 se le llama **variable aleatoria de Bernoulli** (o **binaria**). En la probabilidad básica, al evento $X = 1$ suele llamársele “éxito” y al evento $X = 0$ “fracaso”. Para una aplicación particular, la nomenclatura éxito-fracaso puede que no corresponda a nuestra noción de éxito y fracaso, sin embargo, es una terminología útil que será adoptada en este libro.

Variables aleatorias discretas

Una **variable aleatoria discreta** es una variable que sólo toma una cantidad finita o una cantidad infinita contable de valores. La noción de “infinita contable” significa que aunque la variable aleatoria pueda tomar una cantidad infinita de valores, éstos se pueden poner en correspondencia uno a uno con los enteros positivos. Como la diferencia entre “contable infinito” e “incontable infinito” es un poco sutil, en este libro se concentrará la atención en variables aleatorias discretas que tomen sólo una cantidad finita de valores. Larsen y Marx (1986, capítulo 3) proporcionan un estudio detallado de este tema.

Una variable aleatoria de Bernoulli es el ejemplo más sencillo de variable aleatoria discreta. Lo único que se necesita para describir por completo el comportamiento de una variable aleatoria de Bernoulli es la probabilidad de que la variable tome el valor uno. En el ejemplo del lanzamiento de la moneda, si ésta es “legítima”, entonces $P(X = 1) = 1/2$ (que se lee “la probabilidad de que X sea igual a uno es un medio”). Como las probabilidades deben sumar uno, también $P(X = 0) = 1/2$.

En las ciencias sociales, los científicos están interesados en algo más que el lanzamiento de una moneda, de manera que hay que considerar situaciones más generales. Considere, de nuevo, el ejemplo de la aerolínea que tiene que decidir cuántas reservaciones aceptar para un vuelo en el que hay 100 asientos disponibles. Este problema puede analizarse en el contexto de las variables aleatorias de Bernoulli de la manera siguiente: dado un cliente, tomado aleatoriamente, se define una variable aleatoria de Bernoulli como $X = 1$ si la persona utiliza su reservación y $X = 0$ si no la utiliza.

No hay razón para pensar que la probabilidad de que un determinado cliente utilice su reservación sea 1/2; en principio, esta probabilidad puede ser cualquier número entre cero y uno. Llamémosle θ a este número, de manera que

$$P(X = 1) = \theta \tag{B.1}$$

$$P(X = 0) = 1 - \theta. \tag{B.2}$$

Por ejemplo, si $\theta = .75$, entonces hay 75% de probabilidad de que el cliente utilice su reservación y 25% de que no la utilice. Intuitivamente se ve que el valor de θ es crucial para determinar la estrategia de reservaciones de la aerolínea. Los métodos para *estimar* θ, dados los datos históricos de las reservaciones en aerolíneas, son materia de la estadística matemática, tema que se verá en el apéndice C.

De manera más general, cualquier variable aleatoria discreta queda por completo descrita enumerando sus posibles valores así como las probabilidades de que tome cada uno de ellos. Si los valores X que puede tomar k son $\{x_1, \ldots, x_k\}$, entonces las probabilidades $p_1, p_2, \ldots, p_k$ se definen mediante

$$p_j = P(X = x_j),\ j = 1, 2, \ldots, k, \tag{B.3}$$

donde cada p_j está entre 0 y 1 y

$$p_1 + p_2 + \ldots + p_k = 1. \tag{B.4}$$

La ecuación (B.3) se lee: "la probabilidad de que X tome el valor x_j es igual a p_j."

Las ecuaciones (B.1) y (B.2) muestran que las probabilidades de éxito y fracaso de esta variable aleatoria de Bernoulli quedan totalmente determinadas por el valor de θ. Como las variables aleatorias de Bernoulli son tan frecuentes, se tiene una notación especial para ellas: $X \sim \text{Bernoulli}(\theta)$ que se lee "X tiene una distribución de Bernoulli con probabilidad de éxito igual a θ."

The **función de densidad de probabilidad** (**fdp**) de X resume la información concerniente a los valores que puede tomar X y a sus correspondientes probabilidades:

$$f(x_j) = p_j,\ j = 1, 2, \ldots, k, \tag{B.5}$$

con $f(x) = 0$ para toda x distinta de x_j para alguna j. En otras palabras, para cualquier número real x, $f(x)$ es la probabilidad de que la variable aleatoria X tome ese determinado valor. Cuando se tienen más de una variable para la fdp en cuestión se acostumbra usar subíndices: f_X es la fdp de X, f_Y es la fdp de Y, etcétera.

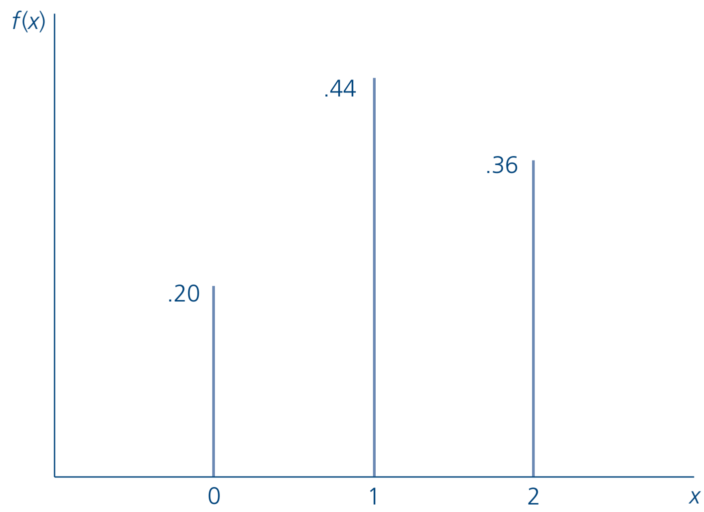
Dada la fdp de cualquier variable aleatoria discreta, es fácil calcular la probabilidad de cualquier evento relacionado con esa variable aleatoria. Por ejemplo, suponga que X es la cantidad de tiros libres anotados por un jugador de baloncesto en dos intentos, entonces X puede tomar los valores $\{0,1,2\}$. Suponga que la fdp de X está definida por

$$f(0) = .20,\ f(1) = .44 \text{ y } f(2) = .36.$$

Estas tres probabilidades suman uno, como debe ser. Usando esta fdp, la probabilidad de que el jugador anote *por lo menos* un tiro libre es: $P(X \geq 1) = P(X = 1) + P(X = 2) = .44 + .36 = .80$. En la figura B.1 se muestra la fdp de X.

FIGURA B.1

Fdp de la cantidad de tiros libres anotados en dos intentos.



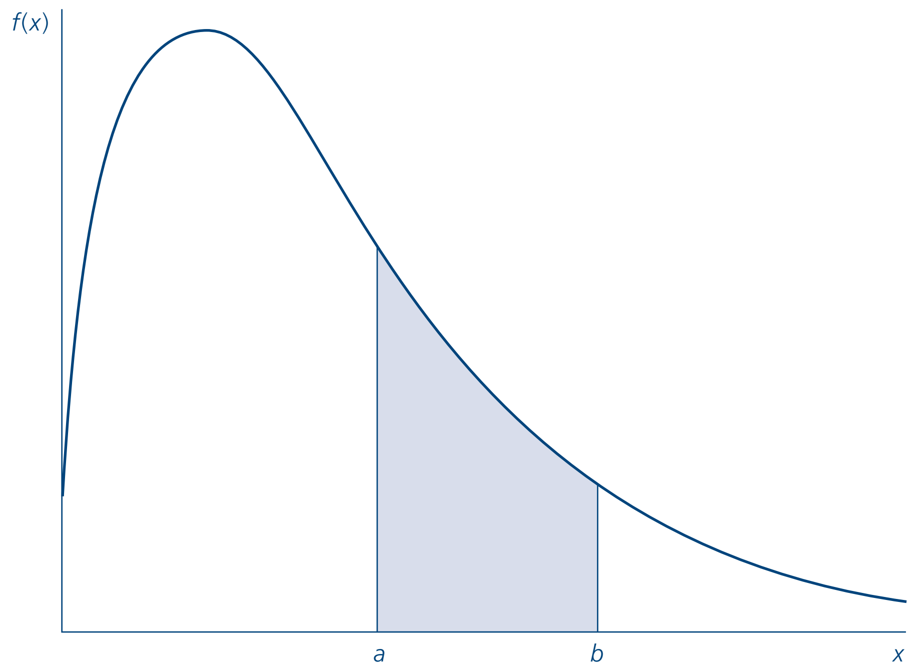
Variables aleatorias continuas

Una variable X es una **variable aleatoria continua** si la probabilidad de que la variable aleatoria tome cualquier valor real es *cero*. Esta definición es un poco contraintuitiva, ya que en cualquier aplicación se observará algún valor de la variable aleatoria. La idea es que los valores que puede tomar una variable aleatoria continua X son tantos que no es posible contarlos o hacerlos coincidir con los enteros positivos, de manera que la consistencia lógica indica que X puede tomar cada uno de estos valores con probabilidad cero. Aunque en la práctica las mediciones son siempre discretas, a las variables aleatorias que toman numerosos valores es mejor tratarlas como continuas. Por ejemplo, la medición más refinada del precio de un artículo se da en términos de centavos. Se puede pensar en enumerar por orden todos los valores de precio posibles (aun cuando esta lista quizá continue de forma indefinida), lo que técnicamente hace que el precio sea una variable aleatoria discreta. Sin embargo, son tantos los posibles valores de precio que usar la mecánica de las variables aleatorias discretas no es posible.

Para las variables aleatorias continuas también se puede definir una función de densidad de probabilidad y, como ocurre con las variables aleatorias discretas, la fdp proporciona información sobre los posibles valores de esta variable aleatoria. Sin embargo, dado que no tiene sentido analizar la posibilidad de que una variable aleatoria continua tome un determinado valor, la fdp de una variable de este tipo sólo se usa para calcular eventos que comprenden un rango de valores. Por ejemplo, si a y b son constantes y $a < b$, la probabilidad de que X se encuentre entre a y b, $P(a \leq X \leq b)$, es el *área* bajo la fdp entre los puntos a y b, como se muestra en la figura B.2. El lector familiarizado con el cálculo recordará que esta es la *integral* de la función f entre los puntos a y b. Toda el área bajo la fdp siempre debe ser igual a uno.

FIGURA B.2

Probabilidad de que *X* se encuentra entre los puntos *a* y *b*.



Para calcular probabilidades de variables aleatorias continuas, es más fácil emplear la **función de distribución acumulada (fda)**. Si X es cualquier variable aleatoria, entonces para cualquier número real x su fda está definida por

$$F(x) \equiv P(X \leq x). \tag{B.6}$$

En el caso de variables aleatorias discretas, (B.6) se obtienen sumando las fdp de todos los valores x_j tales que $x_j \leq x$. En el caso de variables aleatorias continuas, $F(x)$ es el área bajo la fdp, f, a la izquierda del punto x. Como $F(x)$ es simplemente una probabilidad, su valor estará siempre entre 0 y 1. Además, si $x_1 < x_2$, entonces $P(X \leq x_1) \leq P(X \leq x_2)$, es decir $F(x_1) \leq F(x_2)$. Esto significa que una fda es una función creciente (o por lo menos no decreciente) de x.

Dos propiedades importantes de las fda útiles para calcular probabilidades son las siguientes:

$$\text{Para todo número } c,\ P(X > c) = 1 - F(c). \tag{B.7}$$

$$\text{Para todo par de números } a < b,\ P(a < X \leq b) = F(b) - F(a). \tag{B.8}$$

En nuestro estudio de la econometría, las fda sólo se usarán para calcular probabilidades de variables aleatorias continuas, en cuyo caso no importa si las desigualdades empleadas en las fórmulas de probabilidad son desigualdades estrictas o no. Es decir, dada una variable aleatoria continua X,

$$\mathrm{P}(X \geq c) = \mathrm{P}(X > c), \tag{B.9}$$

y

$$\mathrm{P}(a < X < b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a < X \leq b). \tag{B.10}$$

Las ecuaciones (B.9) y (B.10) combinadas con las ecuaciones (B.7) y (B.8) expanden enormemente los cálculos de probabilidad que pueden hacerse usando las fda.

Para todas las distribuciones continuas importantes en la probabilidad y en la estadística, las funciones de distribución acumulada ya han sido tabuladas. La más conocida de estas distribuciones es la normal, que se verá junto con otras en la sección B.5.

B.2 Distribuciones conjuntas, distribuciones condicionales e independencia

En economía, suele interesar la ocurrencia de eventos en los que participan más de una variable aleatoria. Por ejemplo, en el caso antes citado de las reservaciones de una aerolínea, a esta le puede interesar la probabilidad de que una persona que haga una reservación utilice su reservación *y* además sea un viajero de negocios; este es un ejemplo de *probabilidad conjunta*. O puede ocurrir que a la aerolínea le interese la *probabilidad condicional* siguiente: dada la condición de que la persona es viajero de negocios, ¿cuál es la probabilidad de que utilice su reservación? En las dos subsecciones siguientes, se formalizarán los conceptos de distribución conjunta y distribución condicional, así como el importante concepto de *independencia* de las variables aleatorias.

Distribuciones conjuntas e independencia

Sean X y Y variables aleatorias discretas. Entonces (X,Y) tiene una **distribución conjunta**, descrita completamente por la *función de densidad de probabilidad conjunta* de (X,Y):

$$f_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y), \tag{B.11}$$

donde el lado derecho es la probabilidad de que $X = x$ y $Y = y$. Si X y Y son continuas, también se puede definir una fdp, pero estos detalles no se verán aquí, debido a que en este libro las fdp de variables aleatorias continuas no se emplean explícitamente.

Hay un caso en el que la fdp conjunta se obtiene con facilidad si se cuenta con las fdp de X y de Y. En particular, se dice que las variables aleatorias X y Y son independientes si y solo si,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{B.12}$$

para toda x y y, donde f_X es la fdp de X y f_Y es la fdp de Y. Cuando se tiene más de una variable aleatoria, a las fdp f_X y f_Y se les suele llamar *funciones de densidad de probabilidad marginal* para distinguirlas de la fdp conjunta $f_{X,Y}$. Esta definición de independencia es válida tanto para variables discretas como para variables continuas.

Para entender el significado de (B.12), lo más fácil es ver el caso discreto. Si X y Y son discretas, entonces (B.12) es lo mismo que

$$P(X = x, Y = y) = P(X = x)\, P(Y = y); \quad \boxed{\textbf{B.13}}$$

en otras palabras, la probabilidad de que $X = x$ y $Y = y$ es el producto de las probabilidades $P(X = x)$ y $P(Y = y)$. Una consecuencia de B.13 es que (cuando X y Y son independientes) las probabilidades conjuntas son bastante fáciles de calcular, ya que sólo es necesario conocer $P(X = x)$ y $P(Y = y)$.

Si las variables aleatorias no son independientes se dice que son *dependientes.*

Ejemplo B.1

[Lanzamiento de un tiro libre]

Imagine a un jugador de baloncesto que lanza dos tiros libres. Sea X una variable aleatoria de Bernoulli igual a uno si el jugador anota el primer tiro libre e igual a cero si no es así. Sea Y una variable aleatoria de Bernoulli igual a uno si el jugador anota el segundo tiro libre. Suponga que en tiros libres el jugador tiene una probabilidad de anotar de 80%, entonces $P(X = 1) = P(Y = 1) = .8$. ¿Cuál es la probabilidad de que el jugador anote los dos tiros libres?

Si X y Y son independientes, esta pregunta se responde fácilmente: $P(X = 1, Y = 1) = P(X = 1)\, P(Y = 1) = (.8)(.8) = .64$. Por tanto, la posibilidad de que anote los dos tiros libres es de 64%. Pero si la probabilidad de anotar el segundo tiro libre depende de si anota o no el primer tiro libre —es decir, si X y Y no son independientes— entonces este sencillo cálculo no es válido.

La independencia de variables aleatorias es un concepto muy importante. En la subsección siguiente se mostrará que si X y Y son independientes, conocer el valor de X no modifica las probabilidades de los valores posibles de Y, y viceversa. Un hecho útil acerca de la independencia es que si X y Y son independientes, y para funciones g y h cualesquiera se definen las nuevas variables aleatorias $g(X)$ y $h(Y)$, entonces las nuevas variables aleatorias también son independientes.

Pero no es necesario quedarse con dos variables. Si $X_1, X_2, \ldots, X_n$ son variables aleatorias discretas, entonces su fdp conjunta es $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Las variables aleatorias $X_1, X_2, \ldots, X_n$ son **variables aleatorias independientes** si y solo si su fdp conjunta es el producto de sus fdp individuales para toda $(x_1, x_2, \ldots, x_n)$. Esta definición de independencia también es válida para variables aleatorias continuas.

El concepto de independencia es muy importante en la probabilidad y la estadística en la obtención de algunas de las distribuciones clásicas. Ya antes se definió una variable aleatoria de Bernoulli como una variable aleatoria uno-cero que indica la ocurrencia o no de un evento. Con frecuencia, lo que interesa es la cantidad de éxitos en una sucesión de ensayos *independientes* de Bernoulli. El ejemplo más usual de estos ensayos es el lanzamiento de una moneda, una y otra vez. Como el resultado de cada lanzamiento no tiene nada que ver con los resultados de los otros lanzamientos, la independencia es un supuesto adecuado.

En situaciones más complicadas, la independencia suele ser una aproximación razonable. En el ejemplo de las reservaciones de la aerolínea, suponga que para un determinado vuelo la aerolínea acepta n reservaciones. Para $i = 1, 2, \ldots, n$, sea Y_i la variable aleatoria de Bernoulli que indica si el cliente i utiliza su reservación o no: $Y_i = 1$ si el cliente i utiliza su reservación y $Y_i = 0$ si no la utiliza. Sea θ la probabilidad de éxito (el cliente utiliza su reservación); para cada Y_i hay una distribución de Bernoulli(θ). Como aproximación, puede suponerse que las Y_i son independientes unas de otras, aunque en realidad puede que esto no sea verdad: algunas personas viajan en grupo, lo que significa que, el que una persona utilice o no su reservación no es completamente independiente de si las otras la utilizan. Como modelar este tipo de dependencia es complicado, se puede aceptar usar la independencia como aproximación.

La variable que interesa es la cantidad total de los n clientes que utiliza su reservación; llámese X a esta variable. Como cada Y_i es una unidad en el momento que la persona utiliza o no su reservación, se puede escribir $X = Y_1 + Y_2 + \ldots + Y_n$. Ahora, suponiendo que la probabilidad de éxito de cada Y_i sea θ y que las Y_i sean independientes, se pueden mostrar que X tiene una **distribución binomial**. Es decir, la función de densidad de probabilidad de X es

$$f(x) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}, \; x = 0, 1, 2, \ldots, n, \tag{B.14}$$

donde $\binom{n}{x} = \frac{n!}{x!(n - x)!}$, y para todo entero n, $n!$ (que se lee "n factorial") está definido como $n! = n\cdot(n - 1)\cdot(n - 2) \cdots 1$. Por definición, $0! = 1$. Si una variable aleatoria X tiene la fdp dada por (B.14), se escribe $X \sim$ Binomial (n, θ). La ecuación (B.14) se emplea para calcular $\mathrm{P}(X = x)$ para cualquier valor de x desde 0 hasta n.

Si el vuelo dispone de 100 asientos, a la aerolínea le interesa $\mathrm{P}(X > 100)$. Suponga que inicialmente $n = 120$, es decir, que la aerolínea ha aceptado 120 reservaciones, y que la probabilidad de que cada persona utilice su reservación es $\theta = .85$. Entonces, $\mathrm{P}(X > 100) = \mathrm{P}(X = 101) + \mathrm{P}(X = 102) + \ldots + \mathrm{P}(X = 120)$ y cada una de las probabilidades de esta suma se encuentra empleando la ecuación (B.14) con $n = 120$, $\theta = .85$ y el valor de x correspondiente (desde 101 hasta 120). Hacer estos cálculos a mano es bastante complicado, pero por fortuna muchos paquetes de estadística tienen comandos para calcular este tipo de probabilidades. En este caso, la probabilidad de que más de 100 personas utilicen su reservación es aproximadamente .659, que es tal vez un riesgo de exceso de reservaciones mayor al que la aerolínea estaría dispuesta a tolerar. En cambio, si la cantidad de reservaciones es 110, la probabilidad de que más de 100 pasajeros utilice su reservación es aproximadamente de sólo .024.

Distribuciones condicionales

En econometría, usualmente interesa saber cómo está relacionada una variable, a la que se le llamará Y, con otra u otras variables. Por ahora, suponga que sólo interesa el efecto de una variable, llámesele X. Lo más que se puede saber acerca de cómo afecta X a Y está contenido en la **distribución condicional** de Y dada X. Esta información está resumida en la *función de densidad de probabilidad condicional*, definida por

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) \tag{B.15}$$

para todos los valores de x tales que $f_X(x) > 0$. La interpretación de la ecuación (B.15) se ve más fácilmente cuando X y Y son discretas. Entonces,

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x), \tag{B.16}$$

donde el lado derecho se lee como "la probabilidad de que $Y = y$ dado que $X = x$". Cuando Y es continua, $f_{Y|X}(y|x)$ no se puede interpretar directamente como una probabilidad, por las razones antes vistas, sino que las probabilidades condicionales se encuentran calculando áreas bajo la fdp condicional.

Una característica importante de las distribuciones condicionales es que, si X y Y son variables aleatorias independientes, conocer el valor que toma X no dice nada acerca de la probabilidad de que Y tome diversos valores (y viceversa). Es decir, $f_{Y|X}(y|x) = f_Y(y)$ y $f_{X|Y}(x|y) = f_X(x)$.

Ejemplo B.2

[Lanzamiento de un tiro libre]

Considere nuevamente el ejemplo del basquetbolista, en el que se tienen dos tiros libres. Suponga que la densidad condicional es

$$f_{Y|X}(1|1) = .85, f_{Y|X}(0|1) = .15$$
$$f_{Y|X}(1|0) = .70, f_{Y|X}(0|0) = .30.$$

Esto significa que la probabilidad de que un jugador anote el segundo tiro libre depende de si anotó o no el primero: si anotó éste, la posibilidad de que anote el segundo es .85; si lo falló, la posibilidad de que anote el segundo es .70. Esto implica que X y Y *no* son independientes; son dependientes.

Se puede calcular $\mathrm{P}(X = 1, Y = 1)$ siempre y cuando se conozca $\mathrm{P}(X = 1)$. Suponga que la probabilidad de anotar el primer tiro libre es .8, es decir, $\mathrm{P}(X = 1) = .8$. Entonces, de acuerdo con la ecuación (B.15), se tiene

$$\mathrm{P}(X = 1, Y = 1) = \mathrm{P}(Y = 1|X = 1)\cdot \mathrm{P}(X = 1) = (.85)(.8) = .68.$$

B.3 Características de las distribuciones de probabilidad

En muchos casos, únicamente interesan aquí algunas características de las distribuciones de las variables aleatorias. Estas características de interés se pueden agrupar en tres categorías: medidas de tendencia central, medidas de variabilidad o dispersión y medidas de la relación entre dos variables aleatorias. La última de estas categorías se verá en la sección B.4.

Una medida de tendencia central: el valor esperado

El valor esperado es uno de los conceptos más importantes de la probabilidad que surgen en el estudio de la econometría. Si X es una variable aleatoria, el **valor esperado** (o la esperanza) de X, que se denota $\mathrm{E}(X)$ o, alguna veces, μX o simplemente μ, es un promedio ponderado de todos los posibles valores de X. Los pesos de ponderación están determinados por la función de densidad de probabilidad. Al valor esperado también se le suele llamar *media poblacional*, en especial cuando se quiere hacer énfasis en que X representa una variable poblacional.

La definición precisa de valor esperado es más sencilla de dar cuando X es una variable aleatoria discreta que toma un número finito de valores, por ejemplo, $\{x_1, \ldots, x_k\}$. Sea $f(x)$ la función de densidad de probabilidad de X. El valor esperado de X es el promedio ponderado

$$\mathrm{E}(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \quad \textbf{B.17}$$

Este valor es fácil de calcular cuando se conocen todos los valores de la fdp para cada uno de los valores que puede tomar X.

Ejemplo B.3

[Cálculo del valor esperado]

Suponga que X toma los valores -1, 0 y 2 con probabilidades 1/8, 1/2 y 3/8, respectivamente. Entonces

$$\mathrm{E}(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

Este ejemplo ilustra algo curioso acerca del valor esperado: el valor esperado de X puede ser un número que no sea ninguno de los posibles valores de X. Se sabe que X toma los valores -1, 0 y 2, y sin embargo, su valor esperado es 5/8. Esto hace al valor esperado inadecuado para resumir la tendencia central de ciertas variables aleatorias discretas; sin embargo cálculos como los recién mencionados suelen ser útiles, como se verá más adelante.

Si X es una variable aleatoria continua, entonces $\mathrm{E}(X)$ está definida como una integral:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx, \tag{B.18}$$

que se supone está bien definida. Esta integral también se interpreta como un promedio ponderado. En la mayoría de las distribuciones continuas más comunes, $\mathrm{E}(X)$ es un número cuyo valor es un valor posible de X. En este libro no será necesario calcular valores esperados usando la integración, pero sí se recurrirá a algunas de las fórmulas más conocidas de la probabilidad para los valores esperados de variables aleatorias especiales.

Dada una variable aleatoria X y una función $g(\cdot)$, se puede crear una nueva variable aleatoria $g(X)$. Por ejemplo, si X es una variable aleatoria, entonces también X^2 y $\log(X)$ (si $X > 0$) son variables aleatorias. El valor esperado de $g(X)$ es, otra vez, un simple promedio ponderado:

$$\mathrm{E}[g(X)] = \sum_{j=1}^{k} g(x_j)f_X(x_j) \tag{B.19}$$

o, para variables aleatorias continuas,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx. \tag{B.20}$$

Ejemplo B.4

[Valor esperado de X^2]

En el caso de la variable aleatoria del ejemplo B.3, sea $g(X) = X^2$. Entonces,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

En el ejemplo B.3, se obtuvo $\mathrm{E}(X) = 5/8$, de manera que $[\mathrm{E}(X)]^2 = 25/64$. Esto demuestra que $\mathrm{E}(X^2)$ *no* es igual a $[\mathrm{E}(X)]^2$. Efectivamente, si $g(X)$ no es una función lineal $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (salvo en casos muy especiales).

Si X y Y son variables aleatorias, entonces para toda función g $g(X,Y)$ es una variable aleatoria y por tanto, se puede definir su esperanza. Si tanto X como Y son discretas, y toman los valores $\{x_1, x_2, \ldots, x_k\}$ y $\{y_1, y_2, \ldots, y_m\}$, respectivamente, su valor esperado es

$$\mathrm{E}[g(X,Y)] = \sum_{h=1}^{k} \sum_{j=1}^{m} g(x_h,y_j)f_{X,Y}(x_h,y_j),$$

donde $f_{X,Y}$ es la fdp de (X,Y). En el caso de variables aleatorias continuas, la definición es más complicada, ya que se requiere la integración; pero no la necesitaremos aquí. La extensión a más de dos variables aleatorias es inmediata.

Propiedades de los valores esperados

En econometría, lo que más interesa no es el cálculo de los valores esperados de las distintas distribuciones; los principales cálculos ya se han hecho muchas veces y se puede confiar en ellos. Lo que se necesita es manipular algunos valores esperados empleando algunas reglas sencillas. Estas reglas son tan importantes que aquí se les han colocado etiquetas:

Propiedad E.1: Para toda constante c, $\mathrm{E}(c) = c$.

Propiedad E.2: Para todo par de constantes a y b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

Una consecuencia importantes de E.2 es que, si $\mu = \mathrm{E}(X)$, y se define una nueva variable aleatoria $Y = X - \mu$, entonces $\mathrm{E}(Y) = 0$; en E.2 hágase $a = 1$ y $b = -\mu$.

Como ejemplo de la propiedad E.2, sea X la temperatura medida en grados Celsius hacia la tarde de un día y en un lugar determinado; suponga que la temperatura esperada es $\mathrm{E}(X) = 25$. Si Y es la temperatura medida en grados Fahrenheit, entonces $Y = 32 + (9/5)X$. De acuerdo con la propiedad E.2, la temperatura esperada en grados Farenheit es $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

Por lo general, es fácil calcular el valor esperado de una función lineal de muchas variables aleatorias.

Propiedad E.3: Si $\{a_1, a_2, \ldots, a_n\}$ son constantes y $\{X_1, X_2, \ldots, X_n\}$ son variables aleatorias, entonces

$$\mathrm{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mathrm{E}(X_1) + a_2\mathrm{E}(X_2) + \ldots + a_n\mathrm{E}(X_n).$$

O, empleando la sumatoria,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i\mathrm{E}(X_i). \tag{B.21}$$

Como caso especial de esto, se tiene (con todas las $a_i = 1$)

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i). \tag{B.22}$$

de manera que el valor esperado de una suma es la suma de los valores esperados. Esta propiedad es usada con frecuencia para derivaciones en estadística matemática.

Ejemplo B.5

[Determinar ingresos esperados]

Sean X_1, X_2 y X_3 las cantidades de pizzas pequeñas, medianas y grandes vendidas en una pizzería en un día determinado. Éstas son variables aleatorias cuyos valores esperados son $\mathrm{E}(X_1) = 25$, $\mathrm{E}(X_2) = 57$ y

$E(X_3) = 40$. Los precios de las pizzas pequeñas, medianas y grandes son \$5.50, \$7.60 y \$9.15 respectivamente. Por tanto, el ingreso esperado por día de la venta de las pizzas es

$$\begin{aligned} E(5.50\, X_1 + 7.60\, X_2 + 9.15\, X_3) &= 5.50\, E(X_1) + 7.60\, E(X_2) + 9.15\, E(X_3) \\ &= 5.50(25) + 7.60(57) + 9.15(40) = 936.70, \end{aligned}$$

es decir, \$936.70. El ingreso real en un día determinado será, por lo general, distinto de este valor, pero este es el valor *esperado*.

La propiedad E.3 también puede usarse para mostrar que si $X \sim$ Binomial (n,θ), entonces $E(X) = n\theta$. Es decir, la cantidad esperada de éxitos en n ensayos de Bernoulli es simplemente la cantidad de ensayos multiplicada por la probabilidad de éxito en un ensayo particular. Esto se puede ver con facilidad escribiendo X como $X = Y_1 + Y_2 + \ldots + Y_n$, donde cada $Y_i \sim$ binomial(i,n). Entonces,

$$E(X) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

Esto se puede aplicar al ejemplo de las reservaciones de la aerolínea, donde ésta hace $n = 120$ reservaciones y la probabilidad de que el cliente utilice su reservación es $\theta = .85$. La cantidad *esperada* de personas que utilizará su reservación es 120 (.85) = 102. Por lo tanto, si se dispone de 100 asientos, la cantidad esperada de personas que usarán su reservación es demasiada; esto puede servir de orientación a la aerolínea para decidir si es buena idea hacer 120 reservaciones.

En realidad, lo que la aerolínea deberá hacer es definir una función de utilidad que tome en cuenta el ingreso neto por asiento vendido y el costo por pasajero al que se le niegue el embarque. Esta función de utilidad es aleatoria debido a que la cantidad de personas que realmente utilicen su reservación es aleatoria. Sea r el ingreso neto por pasajero. (Para simplificar este se puede considerar como el precio del boleto.) Sea c la compensación que se le debe a cada pasajero al que se le niega el embarque. Ni r ni c son aleatorios; se supone que la aerolínea conoce estas cantidades. Sea Y la ganancia del vuelo. Entonces, como se dispone de 100 asientos,

$$\begin{aligned} Y &= rX \text{ si } X \leq 100 \\ &= 100r - c(X - 100) \text{ si } X > 100. \end{aligned}$$

La primera ecuación da la ganancia si no se presentan más de 100 personas para el vuelo; la segunda ecuación da la ganancia si más de 100 personas quieren utilizar su reservación. (En el último caso, la ganancia neta de la venta de los boletos es $100r$, ya que se han vendido los 100 boletos y $c(X - 100)$ es el costo de hacer más de 100 reservaciones.) Empleando el hecho de que X tiene una distribución Binomial$(n, .85)$ donde n es la cantidad de reservaciones hecha, la ganancia esperada, $E(Y)$, puede determinarse como función de n (y de r y c). Calcular $E(Y)$ de forma directa quizá resulte difícil, pero empleando una computadora esta cantidad puede determinarse rápidamente. Una vez dados los valores de r y c, el valor de n que maximiza la ganancia esperada puede calcularse buscando entre los diferentes valores de n.

Otra medida de tendencia central: la mediana

El valor esperado es sólo una de las posibilidades para definir la tendencia central de una variable aleatoria. Otra medida de tendencia central es la **mediana**. Una definición general de la *mediana* es demasiado complicada para los propósitos de este libro. Si X es continua, entonces la mediana de X, llámese m, es el valor tal que una mitad del área bajo la curva de la fdp queda a la izquierda de m y la otra mitad del área queda a la derecha de m.

Si X es discreta y toma una cantidad finita de valores, la mediana se obtiene ordenando todos los posibles valores de X y seleccionando después el valor de en medio. Por ejemplo, si

X toma los valores $\{-4, 0, 2, 8, 10, 13, 17\}$, entonces la mediana de X es 8. Pero si X toma un número par de valores, hay en realidad dos valores de en medio, estos suelen promediarse para obtener un único valor para la mediana. Por tanto, si X toma los valores $\{-5,3,9,17\}$, entonces los valores medianos son 3 y 9; si se determina el promedio de estos dos valores se obtiene un valor mediano que es igual a 6.

En general, la mediana, que suele denotarse Med(X), y el valor esperado, E(X), suelen ser diferentes. Ninguno de los dos es "mejor" que el otro como medida de tendencia central; ambos son adecuados como medida del centro de la distribución de X. Hay un caso especial, en el que la mediana y el valor esperado (o media) son iguales. Si X tiene una **distribución simétrica** respecto al valor μ, entonces μ es tanto el valor esperado como la mediana. Matemáticamente, la condición es $f(\mu + x) = f(\mu - x)$ para toda x. En la figura B.3 se ilustra este caso.

Medidas de variabilidad: varianza y desviación estándar

Aunque la tendencia central de una variable aleatoria es valiosa, no da toda la información deseada acerca de la distribución de la variable aleatoria. En la figura B.4 se muestran las fdp de dos variables aleatorias que tienen la misma media. Se ve con claridad que la distribución de X está más estrechamente centrada en torno a la media que la distribución de Y. Se desea tener una manera sencilla de resumir las diferencias en la dispersión de las distribuciones.

Varianza

Dada una variable aleatoria X, sea $\mu = \text{E}(X)$. Hay varias maneras de medir qué tan lejos está X de su valor esperado, pero la forma más sencilla para manipular algebraicamente es el cuadrado

FIGURA B.3

Una distribución de probabilidad simétrica.

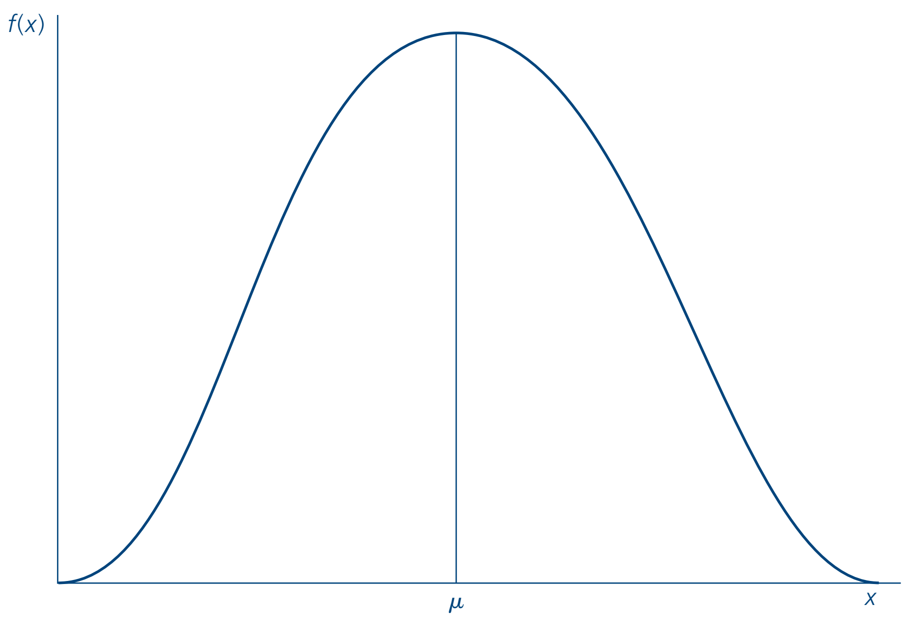
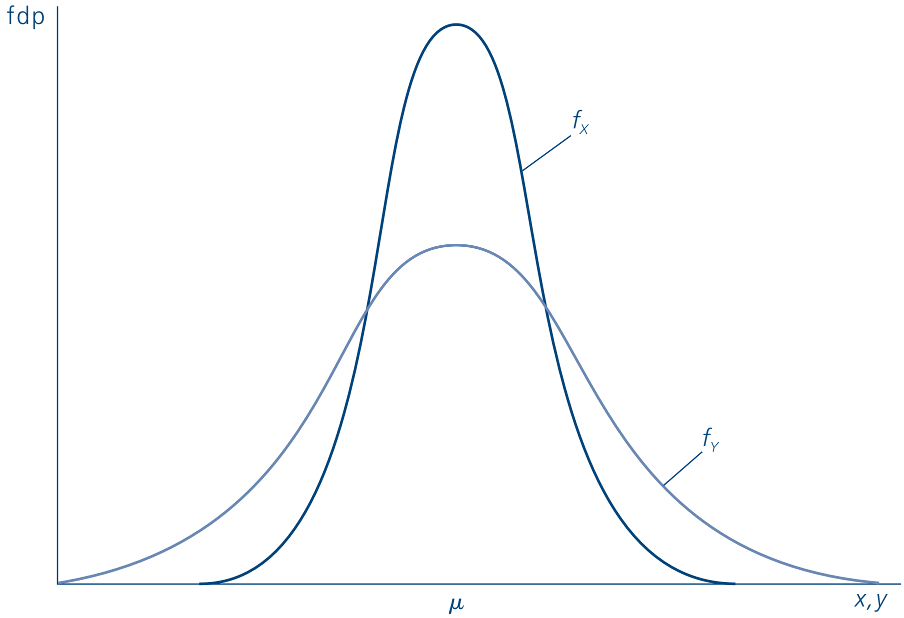


FIGURA B.4

Variables aleatorias con la misma media pero con distintas distribuciones.



de la diferencia, $(X - \mu)^2$. (Al elevar al cuadrado se elimina el signo de la distancia medida; el valor positivo que se obtiene corresponde a la noción intuitiva de distancia y trata de manera simétrica a valores mayores y menores que μ.) Esta distancia también es una variable aleatoria, pues cambia con cada valor que toma X. Así como se necesita un número que resuma la tendencia central de X, también se requiere uno que indique qué tan lejos, *en promedio*, está X de μ. Este número es la **varianza**, que da la distancia esperada de X a su media:

$$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mu)^2]. \qquad \textbf{B.23}$$

La varianza también suele denotarse como σ_X^2, o simplemente σ^2, cuando el contexto es claro. De acuerdo con la ecuación (B.23) se tiene que la varianza siempre es no negativa.

Es útil observar que

$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \qquad \textbf{B.24}$$

Cuando se usa, ya sea (B.23) o (B.24), no es necesario distinguir entre variables aleatorias discretas y variables aleatorias continuas; la definición de varianza es la misma en ambos casos. Lo más frecuente es que se calcule $\mathrm{E}(X)$, después $\mathrm{E}(X^2)$, y después se utiliza la fórmula (B.24). Por ejemplo, si $X \sim \mathrm{Bernoulli}(\theta)$, entonces $\mathrm{E}(X) = \theta$, y, como $X^2 = X$, $\mathrm{E}(X^2) = \theta$. Por tanto, de acuerdo con la ecuación (B.24 $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

A continuación se presentan dos propiedades importantes de la varianza.

Propiedad VAR.1: $\mathrm{Var}(X) = 0$ si y solo si, existe una constante c, tal que $\mathrm{P}(X = c) = 1$, en cuyo caso, $\mathrm{E}(X) = c$.

Esta propiedad indica que la varianza de cualquier constante es cero y que si una variable aleatoria tiene varianza cero, entonces es una constante.

Propiedad VAR.2: Para cualesquiera constantes a y b, $\text{Var}(aX + b) = a^2\text{Var}(X)$.

Esto significa que sumar una constante a una variable aleatoria no modifica la varianza, pero multiplicar una variable aleatoria por una constante aumenta la varianza en un factor igual al *cuadrado* de la constante. Por ejemplo, si *X denota temperatura en grados Celsius y* $Y = 32 + (9/5)X$ es temperatura en grados Fahrenheit, entonces $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

Desviación estándar

La **desviación estándar** de una variable aleatoria, que se denota $\text{sd}(X)$, es la raíz cuadrada positiva de la varianza: $\text{sd}(X) \equiv +\sqrt{\text{Var}(X)}$. La desviación estándar también se suele denotar σ_X, o simplemente σ, cuando se sobreentiende la variable aleatoria. De las propiedades VAR.1 y VAR.2 se deducen de inmediato dos propiedades de la desviación estándar.

Propiedad SD.1: Para toda constante c, $\text{sd}(c) = 0$.

Propiedad SD.2: Para todas las constantes a y b,

$$\text{sd}(aX + b) = |a|\text{sd}(X).$$

En particular, si $a > 0$, entonces $\text{sd}(aX) = a\cdot\text{sd}(X)$.

Esta última propiedad hace que resulte más natural trabajar con la desviación estándar que con la varianza. Por ejemplo, suponga que X es una variable aleatoria medida en miles de dólares, como por ejemplo, ingreso. Si se define $Y = 1{,}000X$, entonces Y es ingreso medido en dólares. Suponga que $\text{E}(X) = 20$ y que $\text{sd}(X) = 6$. Entonces $\text{E}(Y) = 1{,}000\text{E}(X) = 20{,}000$ y $\text{sd}(Y) = 1{,}000\cdot\text{sd}(X) = 6{,}000$, de manera que tanto el valor esperado como la desviación estándar aumentan en un mismo factor, 1,000. Si se calcula la varianza, se tiene $\text{Var}(Y) = (1{,}000)^2\text{Var}(X)$, de manera que la varianza de Y es un millón de veces más grande que la de X.

Estandarización de una variable aleatoria

Como aplicación de las propiedades de la varianza y de la desviación estándar –y un tema de interés práctico por sí mismo– suponga que dada una variable aleatoria X, se define una nueva variable aleatoria restándole su media μ y dividiendo entre su desviación estándar σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \boxed{\textbf{B.25}}$$

lo cual puede escribirse como $Z = aX + b$, donde $a \equiv (1/\sigma)$ y $b \equiv -(\mu/\sigma)$. Entonces, de acuerdo con la propiedad E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

De acuerdo con la propiedad VAR.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Por tanto, la variable aleatoria Z tiene media cero y varianza (y por tanto desviación estándar) igual a uno. A este proceso suele conocérsele como *estandarización* de la variable aleatoria X y a Z como **variable aleatoria estandarizada**. (En cursos de introducción a la estadística a veces es llamada *transformada-z* de X.) Es importante recordar que en el denominador de la ecuación (B.25) aparece la desviación estándar, no la varianza. Como se verá, esta transformación se usa con frecuencia en la inferencia estadística.

Como ejemplo concreto, considere que $\mathrm{E}(X) = 2$ y $\mathrm{Var}(X) = 9$. Entonces, el valor esperado de $Z = (X - 2)/3$ es cero y su varianza es uno.

Sesgo y curtosis

La versión estandarizada de una variable aleatoria puede utilizarse para definir otras características de la distribución de una variable aleatoria. Estas características se describen usando los llamados *momentos de orden superior*. Por ejemplo, el tercer momento de la variable aleatoria Z de la ecuación (B.25) se usa para determinar si la distribución es simétrica respecto a su media. Se puede escribir

$$\mathrm{E}(Z^3) = \mathrm{E}[(X - \mu)^3]/\sigma^3$$

Si X tiene una distribución simétrica respecto a μ, entonces Z tiene una distribución simétrica respecto a cero. (La división entre σ^3 no cambia si la distribución es simétrica o no.) Esto quiere decir que la densidad de Z en dos puntos z y $-z$ es la misma, lo que significa que, al calcular $\mathrm{E}(Z^3)$, los valores positivos de z^3 cuando $z > 0$ se compensan exactamente con los valores negativos $(-z)^3 = -z^3$. Por tanto, si X es simétrica respecto a cero, entonces $\mathrm{E}(Z) = 0$. Por lo general, $\mathrm{E}[(X - \mu)^3]/\sigma^3$ es considerada una medida del **sesgo** de la distribución de X. En la estadística se suele estimar $\mathrm{E}(Z^3)$ para determinar si la distribución de la población subyacente parece ser simétrica. (El ejercicio para computadora C5.4 del capítulo 5 proporciona un ejemplo.)

El cuarto momento de Z también puede ser informativo,

$$\mathrm{E}(Z^4) = \mathrm{E}[(X - \mu)^4]/\sigma^4.$$

Como $Z^4 \geq 0$, $\mathrm{E}(Z^4) \geq 0$ (y, en los casos interesantes, estrictamente mayor que cero). Sin un valor de referencia, es difícil interpretar los valores de $\mathrm{E}(Z^4)$, pero valores mayores significan que las colas de la distribución de X son más gruesas. El cuarto momento $\mathrm{E}(Z^4)$ se considera una medida de la **curtosis** de la distribución de X. En la sección B.5 se obtendrá el valor de $\mathrm{E}(Z^4)$ para una distribución normal.

B.4 Características de las distribuciones conjuntas y de las condicionales

Medidas de asociación: covarianza y correlación

Aunque la fdp conjunta de dos variables aleatorias describe completamente la relación entre ellas, es útil tener una medida resumida del promedio de la variación de estas dos variables aleatorias, una respecto a la otra. Como en el caso del valor esperado y de la varianza, éste es un solo número que resume algo acerca de toda una población, la que, en este caso, es la distribución conjunta de dos variables aleatorias.

Covarianza

Sean $\mu_X = \mathrm{E}(X)$ y $\mu_Y = \mathrm{E}(Y)$ y considere la variable aleatoria $(X - \mu_X)(Y - \mu_Y)$. Si X es mayor a su media y Y es mayor a su media, entonces $(X - \mu_X)(Y - \mu_Y) > 0$. Esto también es cierto si

$X < \mu_X$ y $Y < \mu_Y$. Por otro lado, si $X > \mu_X$ y $Y < \mu_Y$, o viceversa, entonces $(X - \mu_X)(Y - \mu_Y) < 0$. ¿Cómo puede indicar este producto algo acerca de la relación entre X y Y?

La **covarianza** entre dos variables aleatorias X y Y, llamada algunas veces *covarianza poblacional* para hacer énfasis en que se refiere a la relación entre dos variables que describen una población, está definida como el valor esperado del producto $(X - \mu_X)(Y - \mu_Y)$:

$$\mathrm{Cov}(X,Y) \equiv \mathrm{E}[(X - \mu_X)(Y - \mu_Y)], \tag{B.26}$$

que también suele denotarse σ_{XY}. Si $\sigma_{XY} > 0$, entonces, en promedio, cuando X es mayor a su media, también Y es mayor a su media. Si $\sigma_{XY} < 0$, entonces, en promedio, cuando X es mayor a su media Y es menor a su media.

Algunas expresiones útiles para calcular $\mathrm{Cov}(X,Y)$ son las siguientes:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}[(X - \mu_X)Y] \\ &= \mathrm{E}[X(Y - \mu_Y)] = \mathrm{E}(XY) - \mu_X\mu_Y.\end{aligned} \tag{B.27}$$

De (B.27) se sigue que si $\mathrm{E}(X) = 0$ o $\mathrm{E}(Y) = 0$, entonces $\mathrm{Cov}(X, Y) = \mathrm{E}(XY)$.

La covarianza es una medida de la dependencia *lineal* entre dos variables aleatorias. Si la covarianza es positiva indica que las dos variables aleatorias se mueven en la misma dirección, mientras que si es negativa indica que las dos variables se mueven en direcciones opuestas. Interpretar la *magnitud* de la covarianza puede resultar un poco difícil, como se verá en breve.

Como la covarianza es una medida de cómo están relacionadas dos variables, es natural preguntar por la relación entre la covarianza y el concepto de independencia. Esta relación está dada por la siguiente propiedad.

Propiedad COV.1: Si X y Y son independientes, entonces $\mathrm{Cov}(X,Y) = 0$.

Esta propiedad es consecuencia de la ecuación (B.27) y del hecho de que cuando X y Y son independientes, entondes $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$. Es importante recordar que el inverso de COV.1 *no* es verdad: cero covarianza entre X y Y no implica que X y Y sean independientes. En efecto, hay variables aleatorias X tales que si $Y = X^2$, $\mathrm{Cov}(X,Y) = 0$. [Toda variable aleatoria para la que $\mathrm{E}(X) = 0$ y $\mathrm{E}(X^3) = 0$ tiene esta propiedad.] Si $Y = X^2$, entonces es claro que X y Y no son independientes: una vez que se conoce X, se conoce Y. Resulta bastante raro que X y X^2 tengan covarianza cero, y esto revela una debilidad de la covarianza como medida general de la asociación entre dos variables aleatorias. La covarianza es útil en contextos en los que las relaciones son por lo menos aproximadamente lineales.

La segunda propiedad importante de la covarianza está relacionada con covarianzas entre funciones lineales.

Propiedad COV.2: Para todas las constantes a_1, b_1, a_2 y b_2,

$$\mathrm{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\mathrm{Cov}(X, Y). \tag{B.28}$$

Una consecuencia importante de cov.2 es que la covarianza entre dos variables aleatorias puede alterarse simplemente multiplicando una o las dos variables aleatorias por una constante. Esto es importante en economía debido a que las variables monetarias, las tasas de inflación, etcétera, pueden ser definidas en diversas unidades de medición sin que esto cambie su significado.

Por último, es útil saber que el valor absoluto de la covarianza entre dos variables aleatorias está acotado por el producto de sus desviaciones estándar; esto se conoce como *desigualdad de Cauchy-Schwart*.

Propiedad COV.3: $|\mathrm{Cov}(X,Y)| \leq \mathrm{de}(X)\,\mathrm{de}(Y)$.

Coeficiente de correlación

Suponga que se quiere conocer la relación entre cantidad de educación e ingreso anual en la población trabajadora. Para esto, educación se denota por X, e ingreso por Y, y se calcula su covarianza. Pero la respuesta que se obtenga dependerá de cómo se midan educación e ingreso. La propiedad COV.2 implica que la covarianza entre educación e ingresos depende de si éste se mide en dólares o en miles de dólares y de si la educación se mide en meses o en años. Es bastante claro que la manera en que se midan estas variables no tiene nada que ver con qué tan fuerte sea la relación entre ellas. Sin embargo, la covarianza entre ellas sí depende de las unidades de medición.

El hecho de que la covarianza dependa de las unidades de medición es una deficiencia que se supera mediante el **coeficiente de correlación** entre X y Y:

$$\mathrm{Corr}(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\mathrm{de}(X)\cdot\mathrm{de}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \qquad \textbf{B.29}$$

el coeficiente de correlación entre X y Y también suele denotarse ρ_{XY} (y se le suele llamar *correlación poblacional*).

Como σ_X y σ_Y son positivos, $\mathrm{Cov}(X,Y)$ y $\mathrm{Corr}(X,Y)$ siempre tienen el mismo signo y $\mathrm{Corr}(X,Y) = 0$ si y solo si $\mathrm{Cov}(X,Y) = 0$. Algunas de las propiedades de la covarianza se transfieren a la correlación. Si X y Y son independientes entonces $\mathrm{Corr}(X,Y) = 0$, pero que la correlación sea cero no implica independencia. (Como la covarianza, también el coeficiente de correlación es una medida de dependencia lineal.) Sin embargo, la magnitud del coeficiente de correlación es más fácil de interpretar que la magnitud de la covarianza debido a las propiedades siguientes.

Propiedad CORR.1: $-1 \leq \mathrm{Corr}(X,Y) \leq 1$.

Si $\mathrm{Corr}(X,Y) = 0$, o lo que es equivalente $\mathrm{Cov}(X,Y) = 0$, entonces no hay relación lineal entre X y Y, y se dice que X y Y son **variables aleatorias no correlacionadas**; de lo contrario, se dice que X y Y están *correlacionadas.* $\mathrm{Corr}(X,Y) = 1$ implica una relación lineal positiva perfecta, lo que significa que se puede escribir $Y = a + bX$ para alguna constante a y alguna constante $b > 0$. $\mathrm{Corr}(X,Y) = -1$ implica una relación lineal negativa perfecta, de manera que $Y = a + bX$ para alguna $b < 0$. Rara vez se presentan los casos extremos de 1 positivo o 1 negativo. Los valores de ρ_{XY} cercanos a 1 o a -1 indican una fuerte relación lineal.

Como se dijo antes, la correlación entre X y Y es invariante a las unidades de medición de X y de Y. Esto se enuncia de manera más general como sigue.

Propiedad CORR.2: Para todas las constantes a_1, b_1, a_2 y b_2, con $a_1a_2 > 0$,

$$\mathrm{Corr}(a_1X + b_1, a_2Y + b_2) = \mathrm{Corr}(X,Y).$$

Si $a_1a_2 < 0$, entonces

$$\mathrm{Corr}(a_1X + b_1, a_2Y + b_2) = -\mathrm{Corr}(X,Y).$$

Suponga, por ejemplo, que en la población trabajadora, la correlación entre ingreso y educación es .15. Esta medida no depende de si los ingresos se miden en dólares, miles de dólares o en cualquier otra unidad; tampoco depende de si le educación se mide en años, trimestres, meses, etcétera.

Varianza de sumas de variables aleatorias

Una vez que se han definido covarianza y correlación, se puede continuar con la lista de propiedades de la varianza.

Propiedad VAR.3: Para todas las constantes a y b,

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y).$$

Se sigue inmediatamente que, si X y Y no están correlacionadas —de manera que $\mathrm{Cov}(X,Y) = 0$— entonces

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \quad \textbf{B.30}$$

y

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \quad \textbf{B.31}$$

En el último caso observe que la varianza de las diferencias es igual a la *suma de las varianzas*, no a su diferencia.

Como ejemplo de (B.30), sea X los beneficios obtenidos en un restaurant en una noche de viernes y sea Y los beneficios obtenidos en la noche del sábado siguiente. Entonces, $Z = X + Y$ son los beneficios en las dos noches. Suponga que X y Y tienen, cada una, un valor esperado de \$300 y una desviación estándar de \$15 (de manera que su varianza es 225). Los beneficios esperados por las dos noches es $\mathrm{E}(Z) = \mathrm{E}(X) + \mathrm{E}(Y) = 2\cdot(300) = 600$ dólares. Si X y Y son independientes, y lo tanto no están correlacionadas, entonces la varianza de los beneficios totales es la suma de las varianzas: $\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 2\cdot(225) = 450$. Entonces, la desviación estándar de los beneficios totales es $\sqrt{450}$ es decir, aproximadamente \$21.21.

Las expresiones (B.30) y (B.31) pueden extenderse a más de dos variables aleatorias, para esta extensión se necesita una definición. Las variables aleatorias $\{X_1, \ldots, X_n\}$ son **variables aleatorias no correlacionadas entre sí**, si ninguna de las variables del conjunto está correlacionada con ninguna de las otras variables del conjunto. Es decir, $\mathrm{Cov}(X_i,X_j) = 0$, para toda $i \neq j$.

Propiedad VAR.4: Si $\{X_1, \ldots, X_n\}$ son variables aleatorias no correlacionadas entre sí y $\{a_i: i = 1, \ldots, n\}$ son constantes, entonces

$$\mathrm{Var}(a_1X_1 + \ldots + a_nX_n) = a_1^2\mathrm{Var}(X_1) + \ldots + a_n^2\mathrm{Var}(X_n).$$

Empleando la sumatoria, esto puede expresarse como

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\mathrm{Var}(X_i). \quad \textbf{B.32}$$

Un caso especial de la propiedad VAR.4 se presenta cuando se toma $a_i = 1$ para toda i. Entonces, para variables aleatorias no correlacionadas entre sí la varianza de la suma es igual a la suma de las varianzas:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i). \quad \textbf{B.33}$$

Como las variables aleatorias independientes no están correlacionadas (vea la propiedad COV.1), la varianza de una suma de variables aleatorias dependientes es igual a la suma de las varianzas.

Si las X_i son variable no correlacionadas entre sí, entonces la expresión para $\mathrm{Var}\left(\sum_{i=1}^{n} a_iX_i\right)$ es mucho más complicada; en el lado derecho de (B.32) es necesario agregar los términos $2a_i\, a_j\mathrm{Cov}(x_i, x_j)$ para toda $i > j$.

Empleando la ecuación (B.33) se puede obtener la varianza de una variable aleatoria binomial. Sea $X \sim$ Binomial(n,θ) si se escribe $X = Y_1 + \dots + Y_n$, donde las Y_i son variables aleatorias independientes Bernoulli(θ). Entonces, de acuerdo con (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \dots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

En el ejemplo de las reservaciones de la aerolínea, para $n = 120$ y $\theta = .85$, la varianza en la cantidad de pasajeros que utilicen su reservación es $120(.85)(.15) = 15.3$, de manera que la desviación estándar es 3.9.

Esperanza condicional

La covarianza y la correlación miden la relación lineal entre dos variables aleatorias tratándolas simétricamente. Pero en las ciencias sociales, es frecuente que se desee explicar una variable Y, en términos de otra, X. Además, si Y está relacionada con X en un forma que no sea lineal, también se desea saber esto. Sea Y la variable explicada y X la variable explicativa. Por ejemplo, Y puede ser el salario por hora y X los años de educación formal.

Ya antes se introdujo el concepto de una función de densidad de probabilidad condicional de Y dada X. Suponga que se quiere ver cómo cambia la distribución de los salarios con base en el nivel de educación. Sin embargo, por lo general se quiere encontrar una manera sencilla de resumir esta distribución. En este caso, un solo número no podrá ser suficiente, ya que la distribución de Y dado $X = x$ depende, por lo general, del valor de x. No obstante, la relación entre Y y X puede resumirse empleando la *esperanza condicional* de Y dado X, que también suele llamarse *media condicional.* La idea es esta: suponga que se sabe que X ha tomado un determinado valor, por ejemplo, x. Entonces, se puede calcular el valor esperado de Y, dado que se conoce este valor que ha tomado X. Este valor esperado se denota $\text{E}(Y|X = x)$, o también $\text{E}(Y|x)$ para abreviar. En general, cuando x cambia, también lo hace $\text{E}(Y|x)$.

Si Y es una variable aleatoria discreta que toma los valores $\{y_1, \dots, y_m\}$, entonces

$$\text{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

Si Y es continua, $\text{E}(Y|x)$ se define integrando $yf_{Y|X}(y|x)$ sobre todo los valores posibles de y. Como ocurre con las esperanzas no condicionales, la esperanza condicional es un promedio ponderado de los posibles valores de Y, pero ahora los pesos de ponderación reflejan el hecho de que X ha tomado un determinado valor. Por tanto, $\text{E}(Y|x)$ es una función de x, que dice cómo varía el valor esperado de Y de acuerdo con x.

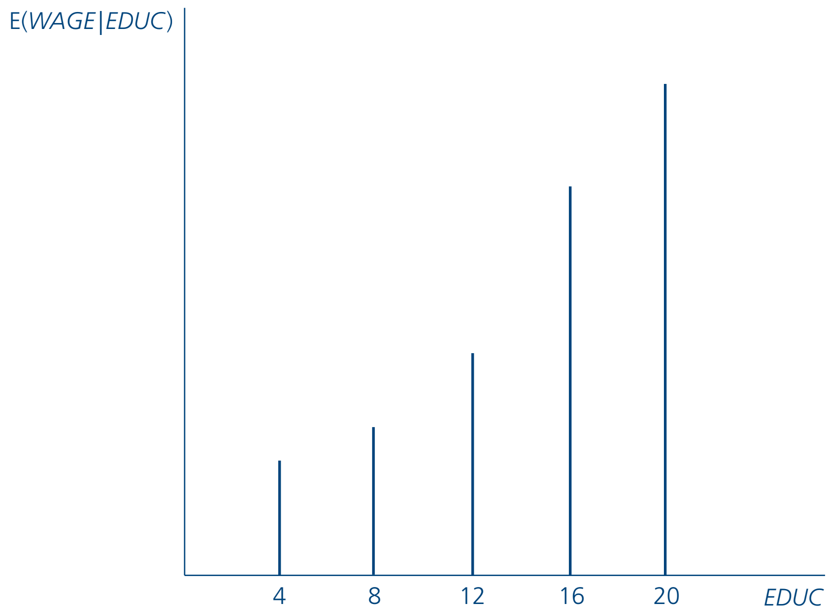
Por ejemplo, sea (X, Y) la población de todos los individuos que trabajan, donde X es años de educación y Y es salario por hora. Entonces, $\text{E}(Y|X = 12)$ es el salario promedio por hora de toda la población que tiene 12 años de educación (más o menos hasta bachillerato). $\text{E}(Y|X = 16)$ es el salario promedio por hora de todas las personas que cuentan con 16 años de educación. Los valores esperados correspondientes a varios niveles de educación proporcionan información importante sobre la relación entre salarios y educación. Véase la figura B.5.

En principio, pueden encontrarse los valores esperados de los salarios por hora correspondientes a cada nivel de educación y estas esperanzas pueden resumirse en una tabla. Pero dado que la educación puede variar mucho —e incluso puede medirse hasta en fracciones de años— esta es una forma muy complicada de medir la relación entre salario promedio y cantidad de educación. En econometría, lo que se acostumbra hacer es dar funciones sencillas que expresen esta relación. Por ejemplo, suponga que el valor esperado de *SALARIO* dada *EDUC* es una función lineal

$$\text{E}(WAGE|EDUC) = 1.05 + .45\ EDUC.$$

FIGURA B.5

Valor esperado del salario por hora dados diversos niveles de educación.



Si esta relación se satisface en la población de las personas que trabajan, el salario promedio de una persona que tenga 8 años de educación es $1.05 + .45(8) = 4.65$, o \$4.65, es decir, \$4.65. El salario promedio de una persona que tenga 16 años de educación será 8.25, es decir, \$8.25. El coeficiente de la variable *EDUC* implica que por cada año más de educación el salario esperado por hora aumenta .45, es decir 45¢.

Las esperanzas condicionales también pueden ser funciones no lineales. Por ejemplo, suponga que $\mathrm{E}(Y|x) = 10/x$, donde X es una variable aleatoria que es siempre mayor que cero. Esta función se grafica en la figura B.6 y puede representar la función de demanda, donde Y es cantidad demandada y X es precio. Si Y y X están relacionadas de esta manera, un análisis de asociación lineal, como el de correlación, será un análisis incompleto.

Propiedades de la esperanza condicional

En el análisis econométrico hay varias propiedades de las esperanzas condicionales que son muy útiles.

Propiedad CE.1: $\mathrm{E}[c(X)|X] = c(X)$, para toda función $c(X)$.

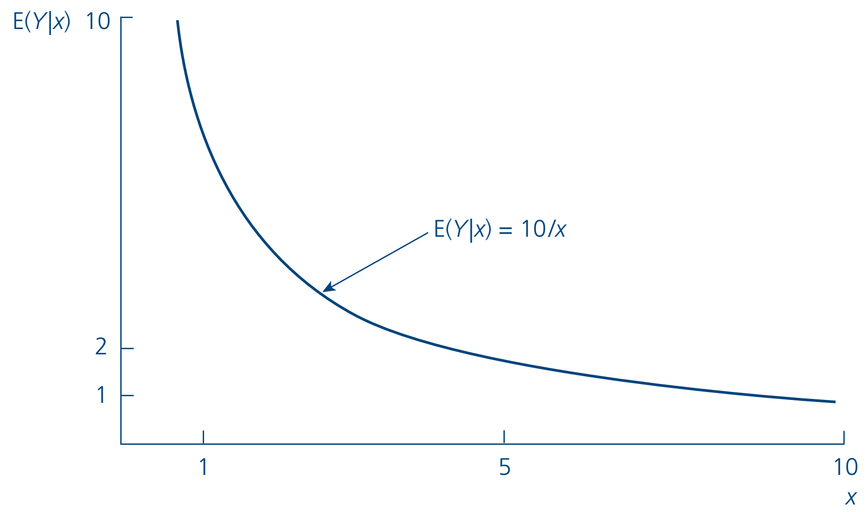
Esta primera propiedad significa que las funciones de X se comportan como constantes al calcular la esperanza condicional dado X. Por ejemplo, $\mathrm{E}(X^2|X) = X^2$. Intuitivamente esto significa que si se conoce X, también se conoce X^2.

Propiedad CE.2: Dadas las funciones $a(X)$ y $b(X)$,

$$\mathrm{E}[a(X)Y + b(X)|X] = a(X)\mathrm{E}(Y|X) + b(X).$$

FIGURA B.6

Gráfica de $\mathbf{E(Y|x) = 10/x}$.



Por ejemplo, la esperanza condicional de una función como $XY + 2X^2$: $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$ puede calcularse fácilmente.

En la propiedad siguiente se relacionan los conceptos de independencia y esperanza condicional.

Propiedad CE.3: Si X y Y son independientes, entonces $E(Y|X) = E(Y)$.

Esta propiedad dice que, si X y Y son independientes, entonces el valor esperado de Y dado X no depende de X, en cuyo caso, $E(Y|X)$ es siempre igual al valor esperado (incondicional) de Y. En el ejemplo del salario y la educación, si el salario es independiente de la educación, entonces el salario promedio de las personas con bachillerato y con universidad será el mismo. Ya que esto es falso, no se puede suponer que salario y educación sean independientes.

Un caso especial de la propiedad EC.3 es el siguiente: Si U y X son independientes y $E(U) = 0$, entonces $E(U|X) = 0$.

Hay también propiedades de la esperanza condicional que tienen que ver con el hecho de que $E(Y|X)$ sea una función de X, por ejemplo, $E(Y|X) = \mu(X)$. Como X es una variable aleatoria, $\mu(X)$ es también una variable aleatoria. Además, $\mu(X)$ tiene una distribución de probabilidad y por tanto un valor esperado. En general, el valor esperado de $\mu(X)$ puede ser muy difícil de calcular de manera directa. La **ley de las esperanzas iteradas** dice que el valor esperado de $\mu(X)$ es simplemente igual al valor esperado de Y. Esto se expresa como sigue.

Propiedad CE.4: $E[E(Y|X)] = E(Y)$.

A primera vista esta propiedad parece un poco difícil de entender. Significa que, si se obtienen primero $E(Y|X)$ como función de X y de esto se determina el valor esperado (por supuesto, respecto la distribución de X, entonces se acaba obteniendo $\mathrm{E}(Y)$. Esto no es muy obvio, pero se puede deducir usando la definición de valor esperado.

Un ejemplo de cómo usar la propiedad EC.4, sea $Y = SALARIO$ y $X = EDUC$, donde *SALARIO* se mide en horas y *EDUC*, en años. Suponga que el valor esperado de *SALARIO* dado *EDUC* es $\mathrm{E}(SALARIO|EDUC) = 4 + .60\ EDUC$. Además, $\mathrm{E}(EDUC) = 11.5$. Entonces, la ley de las esperanzas iteradas implica que $\mathrm{E}(SALARIO) = \mathrm{E}(4 + .60\ EDUC) = 4 + .60\ \mathrm{E}(EDUC) = 4 + .60(11.5) = 10.90$, es decir, \$10.90 por hora.

La propiedad siguiente da una versión más general de la ley de las esperanzas iteradas.

Propiedad CE.4′: $\mathrm{E}(Y|X) = \mathrm{E}[\mathrm{E}(Y|X,Z)|X]$.

En otras palabras, $\mathrm{E}(Y|X)$ se puede obtener en dos pasos. Primero, se encuentra $\mathrm{E}(Y|X,Z)$ para cualquier otra variable aleatoria Z. Después se encuentra el valor esperado de $\mathrm{E}(Y|X,Z)$, condicionado a X.

Propiedad CE.5: Si $\mathrm{E}(Y|X) = \mathrm{E}(Y)$, entonces $\mathrm{Cov}(X,Y) = 0$ [y entonces $\mathrm{Corr}(X,Y) = 0$]. De hecho, *toda* función de X es una función no correlacionada con Y.

Esta propiedad significa que, si el conocer X no modifica el valor esperado de Y, entonces X y Y no *deben* estar correlacionadas, lo que implica que si X y Y están correlacionadas, entonces $\mathrm{E}(Y|X)$ debe depender de X. El inverso de la propiedad EC.5 no es verdad: si X y Y no están correlacionadas, entonces $\mathrm{E}(Y|X)$ *puede* depender de X. Por ejemplo, suponga que $Y = X^2$. Entonces, $\mathrm{E}(Y|X) = X^2$, lo que es claramente una función de X. Sin embargo, como se dijo al hablar de la covarianza y de la correlación, es posible que X y X^2 no estén correlacionadas. La esperanza condicional capta las relaciones no lineales entre X y Y que el análisis de correlación no puede captar.

Las propiedades EC.4 y EC.5 tienen dos consecuencias importantes: si U y X son variables aleatorias tales que $\mathrm{E}(U|X) = 0$, entonces $\mathrm{E}(U) = 0$ y U y X *no* están correlacionadas.

Propiedad CE.6: Si $\mathrm{E}(Y^2) < \infty$ y $\mathrm{E}[g(X)^2] < \infty$ para alguna función g, entonces $\mathrm{E}\{[Y - \mu(X)]^2|X\} \leq \mathrm{E}\{[Y - g(X)]^2|X\}$ y $\mathrm{E}\{[Y - \mu(X)]^2\} \leq \mathrm{E}\{[Y - g(X)]^2\}$.

La propiedad EC.6 es muy útil en contextos de predicción o de pronóstico. La primera desigualdad indica que si se mide la inexactitud de una predicción como el cuadrado del error de predicción *esperado*, condicionado a X, entonces, para predecir X la media condicional es mejor que cualquier otra función de Y. La media condicional también minimiza el cuadrado del error de predicción esperado no condicional.

Varianza condicional

Dadas las variables aleatorias X y Y, la varianza de Y, condicionada a $X = x$, es simplemente la varianza asociada a la distribución condicional de Y, dado $X = x$: $\mathrm{E}\{[Y - \mathrm{E}(Y|x)]^2|x\}$. La fórmula

$$\mathrm{Var}(Y|X = x) = \mathrm{E}(Y^2|x) - [\mathrm{E}(Y|x)]^2$$

suele usarse para hacer cálculos. Sólo ocasionalmente habrá que calcular una varianza condicional. Pero para ciertos temas del análisis de regresión se tendrá que usar la varianza condicional y hacer supuestos acerca de ella.

Por ejemplo, sea $Y = AHORROS$ y $X = INGRESO$ (medidos ambos anualmente en la población de todas las familias). Suponga que $\mathrm{Var}(AHORROS|INGRESO) = 400 + .25\ INGRESO$. Esto indica que a medida que aumenta el ingreso, aumenta la varianza en los niveles de ahorro.

Es importante notar que la relación entre la varianza del *AHORROS* y el *INGRESO* es totalmente distinta a la relación entre *valor esperado* de los *AHORROS* e *INGRESO*.

A continuación se enuncia una útil propiedad acerca de la varianza condicional.

Propiedad CV.1: Si X y Y son independientes, entonces $\mathrm{Var}(Y|X) = \mathrm{Var}(Y)$.

Esta propiedad es bastante clara, ya que la distribución de Y dado X no depende de X y $\mathrm{Var}(Y|X)$ es sólo una característica de esta distribución.

B.5 La distribución normal y otras distribuciones semejantes

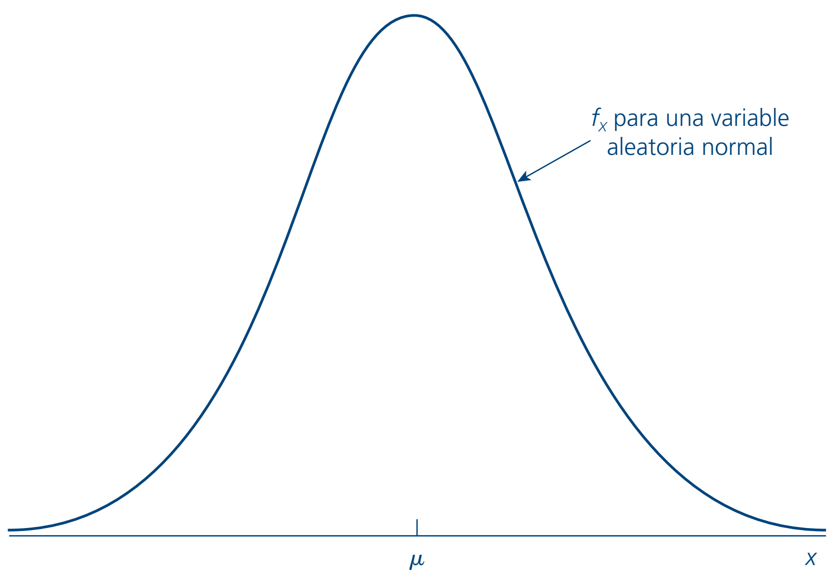
La distribución normal

La distribución normal y las distribuciones que de ella se derivan son las más empleadas en estadística y en econometría. El supuesto de que una variable aleatoria, definida para toda una población, tiene una distribución normal simplifica el cálculo de probabilidades. Además, la distribución normal y las distribuciones semejantes son muy usadas en la estadística y en la econometría para hacer inferencias —aun cuando la población subyacente no sea necesariamente normal. Por ahora, se deben posponer los detalles, pero con toda seguridad este tipo de distribuciones aparecerán con frecuencia en todo este libro.

Una variable aleatoria normal es una variable aleatoria continua que puede tomar cualquier valor. Su función de densidad de probabilidad tiene la conocida forma de curva de campana que se muestra en la figura B.7.

FIGURA B.7

Forma general de la función de densidad de probabilidad normal.



En términos matemáticos, la fdp de X se expresa como

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty, \qquad \textbf{B.34}$$

donde $\mu = \mathrm{E}(X)$ y $\sigma^2 = \mathrm{Var}(X)$. Se dice que X tiene una **distribución normal** cuyo valor esperado es μ y cuya varianza es σ^2, y se escribe $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Como la distribución normal es simétrica respecto a μ, μ es también la mediana de X. A la distribución normal también se le suele llamar *distribución gaussiana* en honor al famoso matemático C. F. Gauss.

Algunas variables aleatorias parecen seguir poco más o menos una distribución normal. Las estaturas y los pesos de los seres humanos, las puntuaciones de test y las tasas de desempleo en los condados (de Estados Unidos) tienen una fdp cuya forma es más o menos la de la figura B.7. Hay otras distribuciones, como la del ingreso, que no siguen una función de probabilidad normal. En la mayoría de los países, el ingreso no está distribuido simétricamente respecto a ningún valor; esta distribución es sesgada hacia la cola superior. En algunos casos, una variable puede transformarse para tener normalidad. Una transformación usual es el logaritmo natural, que sólo tiene sentido para variables aleatorias positivas. Si X es una variable aleatoria positiva, como ingreso, y $Y = \log(X)$ tiene una distribución normal, entonces se dice que X tiene una *distribución lognormal*. Se ha encontrado que en muchos países la distribución lognormal se ajusta bastante bien a la distribución del ingreso. Hay otras variables, como los precios de los bienes, que también pueden describirse como distribuidas lognormales.

La distribución normal estándar

Un caso especial de distribución normal es la distribución normal cuya media es cero y cuya varianza (y, por tanto, desviación estándar) es uno. Si una variable aleatoria Z tiene una distribución normal (0,1) entonces se dice que sigue la **distribución normal estándar**. La fdp de una variable normal estándar se denota $\phi(z)$; de acuerdo con (B.34), dado que $\mu = 0$ y $\sigma^2 = 1$, esta función está dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad -\infty < z < \infty. \qquad \textbf{B.35}$$

La función de la distribución acumulada normal estándar se denota $\Phi(z)$ y es el área bajo ϕ, a la izquierda de z; vea la figura B.8. Hay que recordar que $\Phi(z) = \mathrm{P}(Z \leq z)$; dado que Z es continua, $\Phi(z) = \mathrm{P}(Z < z)$ también es continua.

Para obtener los valores de $\Phi(z)$ no hay ninguna fórmula sencilla que pueda usarse [debido a que $\Phi(z)$ es la integral de la función dada en (B.35), y a que esta integral no tiene una forma cerrada]. No obstante, los valores de $\Phi(z)$ pueden tabularse fácilmente; en la tabla G.1 del apéndice G se dan estos valores para valores de z entre -3.1 y 3.1 Si $z \leq -3.1$, $\Phi(z)$ es menor que .001, y si $z \geq 3.1$, $\Phi(z)$ es mayor que .999. La mayoría del software para estadística y econometría tienen comandos sencillos para calcular los valores de la fda normal estándar, de manera que para obtener las probabilidades correspondientes a cualquier valor de z se puede prescindir totalmente de las tablas impresas

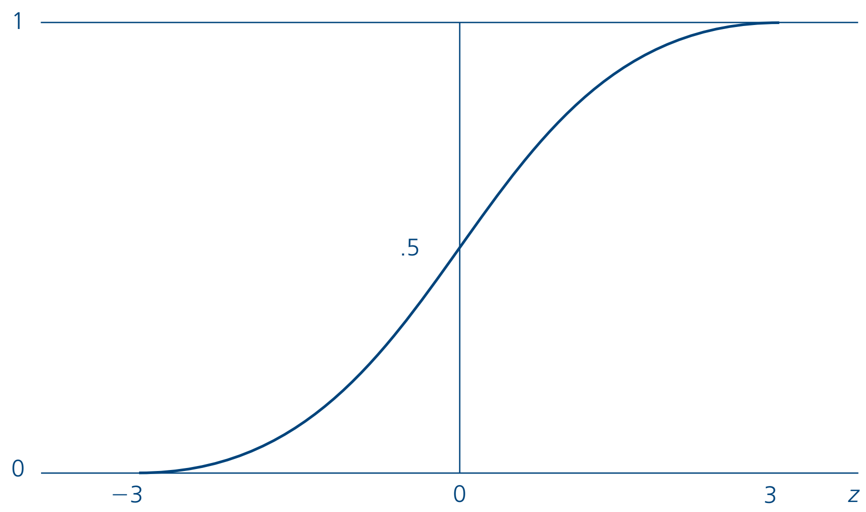
Empleando las propiedades básicas de la probabilidad —y, en particular, las propiedades (B.7) y (B.8) que se refieren a funciones de distribución acumulada— la fda normal estándar puede usarse para calcular la probabilidad de cualquier evento en el que intervenga una variable aleatoria normal estándar. Las fórmulas más importantes son

$$\mathrm{P}(Z > z) = 1 - \Phi(z), \qquad \textbf{B.36}$$

$$\mathrm{P}(Z < -z) = \mathrm{P}(Z > z), \qquad \textbf{B.37}$$

FIGURA B.8

Función de la distribución acumulada normal estándar.



y

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \quad \textbf{B.38}$$

Como Z es una variable aleatoria continua, estas tres fórmulas son válidas, ya sea que las desigualdades sean o no estrictas. Algunos ejemplos son $P(Z > .44) = 1 - .67 = .33$, $P(Z < -.92) = P(Z > .92) = 1 - .821 = .179$ y $P(-1 < Z \leq .5) = .692 - .159 = .533$.

Otra expresión útil es que, para toda $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \quad \textbf{B.39}$$

Por tanto, la probabilidad de que el valor absoluto de Z sea mayor que alguna constante positiva c es simplemente el doble de la probabilidad $P(Z > c)$; esto refleja la simetría de la distribución normal estándar.

En la mayoría de las aplicaciones, se parte de una variable aleatoria distribuida normalmente, $X \sim \text{Normal}(\mu, \sigma^2)$, donde μ es diferente de cero y $\sigma^2 \neq 1$. Todas las variables aleatorias normales pueden transformarse en variables aleatorias normales estándar usando la siguiente propiedad.

Propiedad Normal.1: Si $X \sim \text{Normal}(\mu, \sigma^2)$, entonces $(X - \mu)/\sigma \sim \text{Normal}(0, 1)$.

La propiedad Normal.1 indica cómo transformar cualquier variable aleatoria normal en una variable aleatoria normal estándar. Así, suponga que $X \sim \text{Normal}(3, 4)$ y que se quiere calcular

$P(X \le 1)$. Entre los pasos para hacer esto está la normalización de la variable X a una variable normal estándar:

$$P(X \le 1) = P(X - 3 \le 1 - 3) = P\left(\frac{X-3}{2} \le -1\right)$$

$$= P(Z \le -1) = \Phi(-1) = .159.$$

Ejemplo B.6

[Probabilidades para una variable aleatoria normal]

Primero, se calculará $P(2 < X \le 6)$ dado que $X \sim \text{Normal}(4, 9)$ (donde usar $<$ o $\le$ es irrelevante debido a que X es una variable aleatoria continua). Ahora,

$$P(2 < X \le 6) = P\left(\frac{2-4}{3} < \frac{X-4}{3} \le \frac{6-4}{3}\right) = P(-2/3 < Z \le 2/3)$$

$$= \Phi(.67) - \Phi(-.67) = .749 - .251 = .498.$$

Ahora, se calculará $P(|X| > 2)$:

$$\begin{aligned} P(|X| > 2) &= P(X > 2) + P(X < -2) \\ &= P[(X-4)/3 > (2-4)/3] + P[(X-4)/3 < (-2-4)/3] \\ &= 1 - \Phi(-2/3) + \Phi(-2) \\ &= 1 - .251 + .023 = .772. \end{aligned}$$

Propiedades adicionales de la distribución normal

Esta subsección concluye dando algunas otras propiedades de las distribuciones normales, que se usarán más adelante.

Propiedad Normal.2: Si $X \sim \text{Normal}(\mu, \sigma^2)$, entonces $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Por tanto, si $X \sim \text{Normal}(1, 9)$, entonces $Y = 2X + 3$ tiene una distribución normal con media $2E(X) + 3 = 5$ y varianza $2^2 \cdot 9 = 36$; $sd(Y) = 2sd(X) = 2 \cdot 3 = 6$.

Ya antes se discutió cómo, en general, correlación cero e independencia no son lo mismo. En el caso de variables aleatorias distribuidas normalmente, resulta que correlación cero es suficiente para independencia.

Propiedad Normal.3: Si X y Y son variables conjuntas distribuidas normalmente, entonces X y Y son independientes si y solo si, $\text{Cov}(X, Y) = 0$.

Propiedad Normal.4: Toda combinación lineal de variables aleatorias normales idénticamente distribuidas e independientes tiene una distribución normal.

Por ejemplo, sean X_i, para $i = 1$, 2 y 3, variables aleatorias independientes con distribución $\text{Normal}(\mu, \sigma^2)$. Se define $W = X_1 + 2X_2 - 3X_3$. Entonces, W está distribuida normalmente; sólo hace falta hallar su media y su varianza. Ahora,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

y,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

La propiedad Normal.4 también implica que el promedio de variables aleatorias normalmente distribuidas e independientes tiene una distribución normal. Si $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias independientes y cada una tiene la distribución Normal(μ, σ^2), entonces

$$\bar{Y} \sim \text{Normal}(\mu, \sigma^2/n). \tag{B.40}$$

Este resultado es crucial en la inferencia estadística acerca de la media de una población normal.

Hay otras propiedades de la distribución normal que vale la pena conocer, aunque no tengan un papel importante en este libro. Como una variable aleatoria normal es simétrica respecto a su media, tiene sesgo cero, es decir, $\text{E}[(X - \mu)^3] = 0$. Además, se puede demostrar que

$$\text{E}[(X - \mu)^4]/\sigma^4 = 3,$$

o $\text{E}(Z^4) = 3$, donde Z tiene distribución normal estándar. Como la distribución normal se encuentra con tanta frecuencia en probabilidad y estadística, la medida de la curtosis de cualquier variable aleatoria X (cuyo cuarto momento exista) suele definirse como $\text{E}[(X - \mu)^4]/\sigma^4 - 3$, es decir, relativo al valor de la distribución normal estándar. Si $\text{E}[(X - \mu)^4]/\sigma^4 > 3$, entonces la distribución de X tienen colas más gruesas que la distribución normal (algo que suele ocurrir, como por ejemplo en la distribución t, que se presentará en breve); si $\text{E}[(X - \mu)^4]/\sigma^4 < 3$, entonces la distribución tienen colas más delgadas que la normal (una situación poco frecuente).

La distribución *ji*-cuadrada

La distribución *ji*-cuadrada se obtiene directamente a partir de variables normales estándar, independientes. Sean Z_i, $i = 1, 2, \ldots, n$, variables aleatorias independientes, cada una con una distribución normal estándar. Se define una nueva variable como la suma de los cuadrados de las Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \tag{B.41}$$

Entonces, X tiene lo que se conoce como una **distribución *ji*-cuadrada** con n **grados de libertad** (gl). Esto se expresa $X \sim \chi_n^2$. En una distribución *ji*-cuadrada, los gl corresponden a la cantidad de términos en la suma de la ecuación (B.14). En este libro, el concepto de grados de libertad será muy importante en los análisis estadísticos y econométricos.

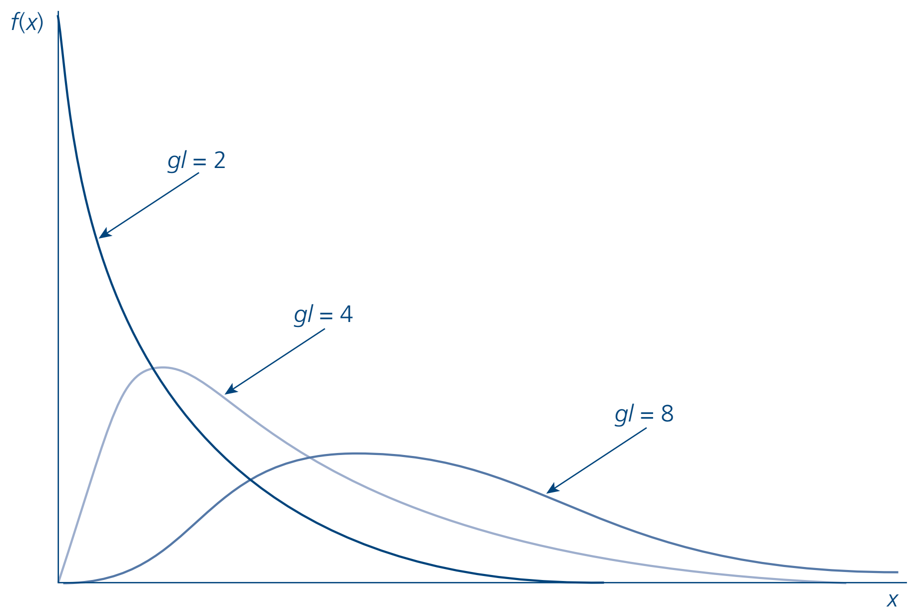
En la figura B.9 se presentan las fdp de distribuciones *ji*-cuadrada correspondientes a diversos grados de libertad; la fórmula de esta fdp no se necesitará, por lo que no se reproduce aquí. De acuerdo con la ecuación (B.41), es claro que las variables aleatorias *ji*-cuadrada son siempre no negativas, y que, a diferencia de la distribución normal, la distribución *ji*-cuadrada no es simétrica respecto a ningún punto. Se puede demostrar que es si $X \sim \chi_n^2$, entonces el valor esperado de X es n [el número de términos en (B.41)] y la varianza de X es $2n$.

La distribución *t*

La distribución t es el caballo de batalla de la estadística clásica y del análisis de regresión múltiple. Una distribución t se obtiene a partir de una variable aleatoria normal estándar y una variable aleatoria *ji*-cuadrada.

FIGURA B.9

Distribuciones *ji*-cuadrada para diversos grados de libertad.



Sea Z una variable aleatoria que tiene una distribución normal estándar y sea X una variable aleatoria que tiene una distribución *ji*-cuadrada con n grados de libertad. Suponga, además, que Z y X son independientes. Entonces, la variable aleatoria

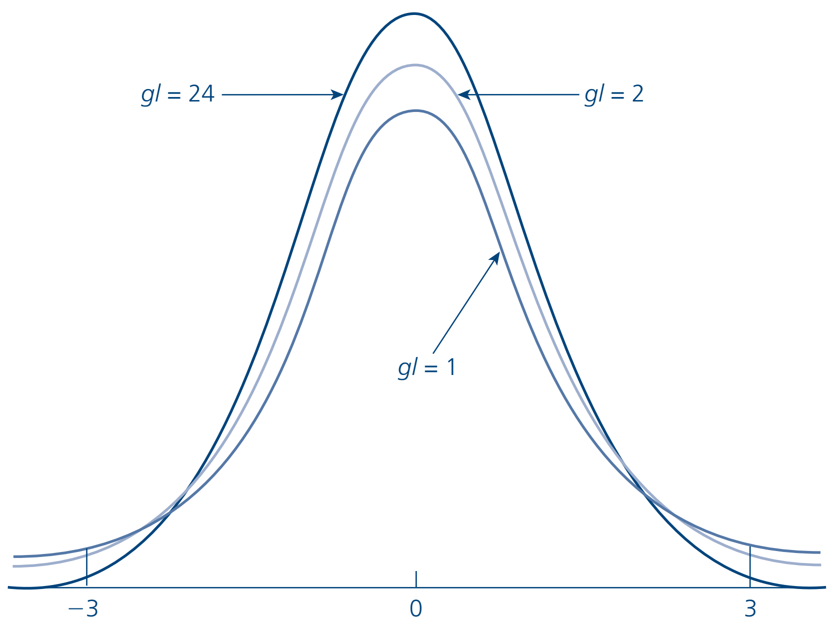
$$T = \frac{Z}{\sqrt{X/n}} \tag{B.42}$$

tiene una **distribución *t*** con n grados de libertad. Esto se denotará $T \sim t_n$. Las distribuciones t obtienen sus grados de libertad de la variable aleatoria *ji*-cuadrada en el denominador de la ecuación (B.42).

La fdp de la distribución t tienen una forma similar a la de la distribución normal estándar, sólo que es más dispersa, y por tanto tiene áreas mayores en las colas. El valor esperado de una variable aleatoria con distribución t es cero (estrictamente hablando, el valor esperado existe sólo para $n > 1$) y la varianza es is $n/(n - 2)$ para $n > 2$. (Para $n \leq 2$ no existe la varianza debido a que la distribución es demasiado dispersa.) En la figura B.10 se grafican distribuciones t correspondientes a diversos grados de libertad. A medida que los grados de libertad aumentan, las distribuciones t se aproximan a la distribución normal estándar.

FIGURA B.10

Distribuciones *t* con diversos grados de libertad.



La distribución *F*

Otra distribución importante para aquellos que se dedican a la estadística o a la econometría es la distribución *F*. Ésta se usará en especial para probar hipótesis en el contexto del análisis de regresión múltiple.

Para definir una variable aleatoria *F*, sean $X_1 \sim \chi^2_{k_1}$ y $X_2 \sim \chi^2_{k_2}$ y suponga que X_1 y X_2 son independientes. Entonces, la variable aleatoria

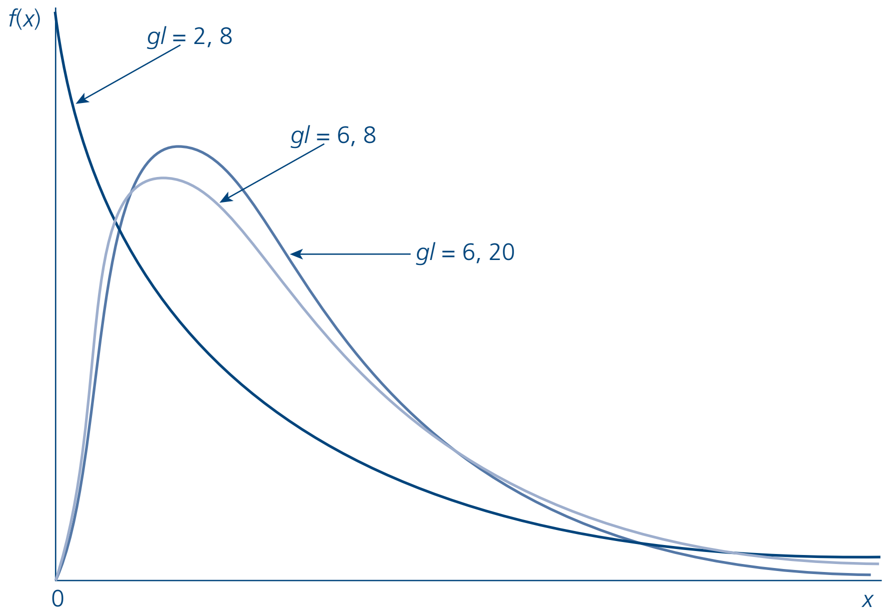
$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \tag{B.43}$$

tiene una **distribución *F*** con (k_1, k_2) grados de libertad. Esto se denota como $F \sim F_{k_1, k_2}$. En la figura B.11 se muestran las fdp de distribuciones *F* con diversos grados de libertad.

En F_{k_1, k_2} el orden de los grados de libertad es crítico. El entero k_1 son los *grados de libertad en el numerador*, debido a que corresponde a la variable *ji*-cuadrada del numerador. De igual manera, el entero k_2 son los *grados de libertad en el denominador*, debido a que corresponde a la variable *ji*-cuadrada del denominador. Aquí hay que tener cuidado, pues la ecuación (B.34) también puede expresarse como $(X_1k_2)/(X_2k_1)$, de manera que k_1 aparece en el denominador. Sólo hay que recordar que los *gl* del numerador es el entero asociado con la variable *ji*-cuadrada en el numerador de la ecuación (B.43) y de manera similar para *gl del denominador*.

FIGURA B.11

Distribuciones F_{k_1, k_2} para diversos grados de libertad, k_1 y k_2.



RESUMEN

En este apéndice se hizo un repaso de los conceptos de probabilidad que se emplean en la econometría. La mayoría de ellos resultará familiar al lector por los cursos introductorios de probabilidad y estadística. Algunos de los temas más avanzados, por ejemplo las propiedades de las esperanzas condicionales, no es necesario que se dominen ahora —habrá tiempo para eso cuando estos conceptos surjan en el contexto del análisis de regresión, en la parte 1.

En un curso introductorio de estadística, la atención se centra en los cálculos de medias, varianzas, covarianzas, etcétera de las diferentes distribuciones. En la parte 1 no se necesitarán estos cálculos: lo que se usará serán las *propiedades* de la esperanza, varianza, etcétera, vistas en este apéndice.

TÉRMINOS CLAVE

Coeficiente de correlación
Covarianza
Curtosis
Desviación estándar
Distribución binomial
Distribución *ji*-cuadrada
Distribución condicional
Distribución conjunta
Distribución *F*
Distribución normal
Distribución normal estándar
Distribución simétrica
Distribución *t*
Esperanza condicional
Experimento

Función de densidad de probabilidad (fdp)
Función de distribución acumulada (fda)
Grados de libertad
Ley de las esperanzas iteradas
Mediana
Sesgo
Valor esperado
Variable aleatoria
Variable aleatoria continua
Variable aleatoria de Bernoulli (o binaria)
Variable aleatoria discreta
Variable aleatoria estandarizada
Variables aleatorias independientes
Variables aleatorias no correlacionadas
Variables aleatorias no correlacionadas entre sí
Varianza

PROBLEMAS

B.1 Suponga que un estudiante de bachillerato se está preparando para presentar el examen SAT de admisión a la universidad. Explique por qué la calificación que obtendrá en este examen puede ser considerada como una variable aleatoria.

B.2 Sea X una variable aleatoria con distribuciones normal (5, 4). Determine las probabilidades de los eventos siguientes:

i) $P(X \leq 6)$.

ii) $P(X > 4)$.

iii) $P(|X - 5| > 1)$.

B.3 Se ha hablado mucho acerca de que hay algunos fondos mutualistas que superan el mercado año tras año (es decir, el rendimiento de las acciones de los fondos mutualistas es más alto que el de un portafolio como el S&P 500). En concreto, considere un periodo de 10 años y la población formada por los 4,170 fondos mutualistas publicados en *The Wall Street Journal* el 1 de enero de 1995. Decir que el desempeño en relación al mercado es aleatorio significa que cada año, cada fondo tiene 50 por ciento de posibilidades de superar el mercado y que su desempeño es independiente de un año a otro.

i) Si el desempeño en relación al mercado es realmente aleatorio, ¿cuál es la probabilidad de que un determinado fondo supere el mercado todos estos 10 años?

ii) Calcule la probabilidad de que en estos 10 años *por lo menos* uno de los 4,170 fondos supere el mercado?

iii) Empleando un software estadístico que calcule probabilidades binomiales, determine la probabilidad de que, en estos 10 años, por lo menos cinco fondos superen el mercado.

B.4 Dado un condado de Estados Unidos, tomado al azar, sea X la proporción de adultos mayores de 65 años que tienen un empleo, es decir, la tasa de empleo entre los adultos mayores. Entonces, X está restringida a tomar un valor entre cero y uno. Suponga que la función de distribución acumulada de X está dada por $F(x) = 3x^2 - 2x^3$ para $0 \leq x \leq 1$. Calcule la probabilidad de que la tasa de empleo entre los adultos mayores sea por lo menos .6 (60%).

B.5 Antes de elegir el jurado para el juicio por homicidio contra O. J. Simpson en 1995, en una encuesta se encontró que aproximadamente 20% de la población adulta creía que Simpson era inocente (dado que gran parte de las evidencias físicas del caso habían sido reveladas al público). Ignorando el hecho de que este 20% sea una estimación basada en una submuestra tomada de la población; considérese como ilustración, como el verdadero porcentaje de personas que, antes de la elección del jurado, pensaban que Simpson era inocente. Suponga que los 12 miembros del jurado fueron elegidos de la población de forma aleatoria e independiente (aunque esto resultó no ser cierto).

(i) Calcule la probabilidad de que entre los miembros del jurado haya habido por lo menos uno que antes de la elección del mismo haya creído en la inocencia de Simpson. [*Sugerencia:* defínase una variable aleatoria binomial (12, .20) X correspondiente a la cantidad de miembros del jurado que creían en la inocencia de Simpson.]

ii) Calcule la probabilidad de que entre los miembros del jurado haya habido al menos dos que hayan creído la inocencia de Simpson. [*Sugerencia:* $\mathrm{P}(X \geq 2) = 1 - \mathrm{P}(X \leq 1)$, y $\mathrm{P}(X \leq 1) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1)$.]

B.6 (Se requiere cálculo) Sea X los años de cárcel, en un determinado estado de Estados Unidos, para una persona acusada del robo de un automóvil. Suponga que la fdp de X sea

$$f(x) = (1/9)x^2, 0 < x < 3.$$

Use la integración para calcular los años de cárcel esperados.

B.7 Si un jugador de baloncesto tiene una probabilidad de 74% de anotar un tiro libre, entonces, en promedio, ¿cuántos tiros libres anotará en un juego en el que tiene ocho intentos de tiro libre?

B.8 Suponga que un estudiante universitario toma tres cursos: uno de dos créditos, otro de tres y otro más de cuatro. La calificación esperada en el curso de dos créditos es 3.5, mientras que la calificación esperada en los cursos de tres y cuatro créditos es 3.0. ¿Cuál es la calificación promedio esperada en el semestre? (Recuerde que cada calificación está ponderada por el número total de créditos.)

B.9 Sea X el salario anual de los profesores universitarios en Estados Unidos, medido en miles de dólares. Suponga que el sueldo promedio sea 52.3 y la desviación estándar 14.6. Halle la media y la desviación estándar si el sueldo se mide en dólares.

B.10 B.10 Suponga que en una universidad grande, el promedio de calificaciones obtenido, *GPA*, y la puntuación en el examen de admisión, *SAT*, se relacionan mediante la esperanza condicional $\mathrm{E}(GPA|SAT) = .70 + .002\ SAT$.

i) Determine el *GPA* esperado cuando $SAT = 800$. Calcule $\mathrm{E}(GPA|SAT = 1{,}400)$. Analice la diferencia.

ii) Si en esta universidad el *SAT* promedio es 1,100, ¿cuál será el *GPA* promedio? (*Sugerencia:* emplee la propiedad EC.4.)

iii) Si la puntuación obtenida en el *SAT* por un estudiante es 1,100, ¿significa esto que su calificación promedio, *GPA*, será la encontrada en el inciso ii)? Explique.

Apéndice C

Fundamentos de estadística matemática

C.1 Poblaciones, parámetros y muestreo aleatorio

La inferencia estadística involucra saber algo de una población, dada la disponibilidad de una muestra de esa población. Por **población** se entiende cualquier grupo bien definido de sujetos, que podrían ser individuos, empresas, ciudades o muchas otras posibilidades. "Saber" puede significar muchas cosas, las cuales se dividen en categorías de *estimación* y *prueba de hipótesis*.

Un par de ejemplos lo ayudarán a comprender estos términos. En la población de todos los adultos que trabajan en Estados Unidos, los economistas laborales están interesados en conocer el rendimiento de la educación, medido como el incremento porcentual promedio en los ingresos, dado otro año de educación. Sería poco práctico y costoso obtener la información sobre los ingresos y la educación de toda la población laboral en Estados Unidos, pero es posible obtener datos de un subconjunto de la población. Mediante los datos recabados, un economista laboral puede reportar que su mejor estimación del rendimiento por otro año de educación es de 7.5%. Esto es un ejemplo de *estimación puntual*. O quizá reporte un rango, como el "rendimiento de la educación se halla entre 5.6 y 9.4%." Éste es un ejemplo de *estimación de intervalo*.

Un economista urbano tal vez quiera saber si los programas vecinales de vigilancia contra la delincuencia están asociados con menores tasas de delitos. Después de comparar tales tasas en los vecindarios con y sin tales programas en una muestra de la población, el economista podría formular una de dos conclusiones: los programas de vigilancia vecinal afectan la delincuencia, o no lo hacen. Este ejemplo cae bajo el rubro de la prueba de hipótesis.

El primer paso en la inferencia estadística es identificar a la población de interés. Esto parece obvio, pero es importante ser muy específico. Una vez que se ha identificado a la población, se puede especificar un modelo para la relación poblacional de interés. Tales modelos involucran las distribuciones de probabilidad o las características de distribuciones de probabilidad, y éstas dependen de parámetros desconocidos. Los parámetros simplemente son constantes que determinan las direcciones y fortalezas de las relaciones entre las variables. En el ejemplo anterior de la economía laboral, el parámetro de interés es el rendimiento de la educación sobre la población.

Muestreo

Para analizar la inferencia estadística, se utilizará el escenario más simple posible. Sea Y una variable aleatoria que representa a una población con una función de densidad de probabilidad $f(y;\theta)$, que depende de un sólo parámetro θ. Se supone que se conoce la función de densidad de probabilidad (fdp) de Y, salvo por el valor de θ; diferentes valores de θ implican diferentes distribuciones poblacionales y, por tanto, lo que interesa aquí es el valor de θ. Es posible obtener ciertos tipos de muestras de la población, entonces podemos conocer algo acerca de θ. El esquema de muestreo más sencillo de manejar es el muestreo aleatorio.

Muestreo aleatorio. Si $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias independientes con una función de densidad de probabilidad común $f(y;\theta)$, entonces se dice que $\{Y_1, \ldots, Y_n\}$ es una **muestra aleatoria** de $f(y;\theta)$ [o muestra aleatoria de la población representada por $f(y;\theta)$].

Cuando $\{Y_1, \ldots, Y_n\}$ es una muestra aleatoria de la densidad $f(y;\theta)$, también se dice que Y_i son *variables aleatorias independientes, idénticamente distribuidas* (o *i.i.d.*) de $f(y;\theta)$. En algunos casos no se necesitará especificar por completo cuál es la distribución común.

La naturaleza aleatoria de $Y_1, Y_2, \ldots, Y_n$ en la definición de muestreo aleatorio refleja el hecho de que muchos resultados diferentes son posibles antes de que se realice realmente la muestra. Por ejemplo, si se obtiene el ingreso familiar para una muestra de $n = 100$ familias en Estados Unidos, los ingresos que se observan, por lo general, difieren para cada muestra diferente de 100 familias. Una vez que se obtenga la muestra se tiene un conjunto de números, por ejemplo, $\{y_1, y_2, \ldots, y_n\}$, que constituyen los datos con los que se trabajará. Si es o no correcto suponer que la muestra proviene de un esquema de muestreo aleatorio requiere conocer algo del proceso real de muestreo.

Las muestras aleatorias de una distribución de Bernoulli suelen utilizarse para ilustrar conceptos estadísticos, y también surgen en aplicaciones empíricas. Si $Y_1, Y_2, \ldots, Y_n$ son variables independientes aleatorias y cada una se distribuye como Bernoulli(θ), de manera que $P(Y_i = 1) = \theta$ y $P(Y_i = 0) = 1 - \theta$, entonces $\{Y_1, Y_2, \ldots, Y_n\}$ constituye una muestra aleatoria de la distribución de Bernoulli(θ). Como ilustración, considere el ejemplo de las reservaciones de la aerolínea que se realizó en el apéndice B. Cada Y_i denota si el cliente i muestra su reservación; $Y_i = 1$ si el pasajero i se presenta y $Y_i = 0$ de otra forma. Aquí, θ es la probabilidad de que una persona elegida al azar de la población de todas las personas que hicieron reservaciones aéreas se presente para su reservación.

En muchas otras aplicaciones se puede suponer que las muestras aleatorias se extraen de una distribución normal. Si $\{Y_1, \ldots, Y_n\}$ es una muestra aleatoria de la población Normal(μ,σ^2) entonces la población está caracterizada por dos parámetros, la media μ y la varianza σ^2. El interés principal suele enfocarse en μ, pero σ^2 es de interés por derecho propio debido a que las inferencias relativas a μ suelen requerir conocimientos acerca de σ^2.

C.2 Propiedades de muestras finitas de los estimadores

En esta sección, se estudia lo que se conoce como propiedades de los estimadores en muestras finitas. El término "muestra finita" proviene del hecho que las propiedades son válidas para una de cualquier tamaño, sin importar si es grande o pequeña. En ocasiones, éstas reciben el nombre de propiedades de muestras pequeñas. En la sección C.3, se cubrirán las "propiedades asintóticas", que tienen que ver con el comportamiento de los estimadores a medida que el tamaño de la muestra crece sin límite.

Estimadores y estimaciones

Para estudiar las propiedades de los estimadores se debe definir lo que se entiende por estimador. Dada una muestra aleatoria $\{Y_1, Y_2, \ldots, Y_n\}$ extraída de una distribución de la población que depende de un parámetro desconocido θ, un **estimador** de θ es una regla que asigna a cada resultado posible de la muestra un valor de θ. La regla se especifica antes de realizar cualquier muestreo y ésta es la misma, sin considerar realmente los datos obtenidos.

Como un ejemplo de un estimador, sea $\{Y_1, \ldots, Y_n\}$ una muestra aleatoria de una población con media μ. Un estimador natural de μ es el promedio de la muestra aleatoria:

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i. \tag{C.1}$$

$\bar{Y}$ recibe el nombre de **promedio muestral**, pero a diferencia del apéndice A, donde se definió el promedio muestral de un conjunto de números como un estadístico descriptivo, $\bar{Y}$ ahora se considera como un estimador. Dado cualquier resultado de las variables aleatorias $Y_1, \ldots, Y_n$, se utiliza la misma regla para estimar μ: simplemente se promedian. Para los datos reales $\{y_1, \ldots, y_n\}$, la **estimación** es sólo el promedio en la muestra: $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$.

Ejemplo C.1

[Tasas de desempleo en las ciudades]

Suponga que se obtiene la siguiente muestra de las tasas de desempleo en 10 ciudades de Estados Unidos:

Ciudad	Tasa de desempleo
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6
8	3.5
9	5.8
10	7.5

La estimación de la tasa de desempleo promedio en Estados Unidos es $\bar{y} = 6.0$. Cada muestra, por lo general, tiene como resultado una estimación diferente. Pero la *regla* para obtener la estimación es la misma, sin importar qué ciudades aparecen en la muestra, o cuántas.

En términos más generales, un estimador W de un parámetro θ se puede expresar como una fórmula matemática abstracta:

$$W = h(Y_1, Y_2, \ldots, Y_n), \tag{C.2}$$

para alguna función conocida h de las variables aleatorias $Y_1, Y_2, \ldots, Y_n$. Como en el caso especial del promedio muestral, W es una variable aleatoria pues depende de la muestra aleatoria: conforme se obtienen diferentes muestras aleatorias de la población, el valor de W puede cambiar. Cuando un conjunto de números en particular, por ejemplo $\{y_1, y_2, \ldots, y_n\}$, se inserta a la función h, se obtiene una *estimación* de θ, denotada como $w = h(y_1, \ldots, y_n)$. En ocasiones, W es llamado un estimador puntual y w una estimación puntual para distinguirlos de los estimadores y estimaciones de *intervalo*, que se estudiarán en la sección C.5.

Para evaluar los procedimientos de estimación, se estudian las diferentes propiedades de la distribución probabilística de la variable aleatoria W. La distribución de un estimador suele recibir el nombre de **distribución de muestreo**, debido a que ésta describe la probabilidad de varios resultados de W a través de diferentes muestras aleatorias. Debido a que existen reglas ilimitadas en la combinación de datos para estimar parámetros, se requieren algunos criterios sensibles para elegir entre los estimadores, o al menos para no considerar algunos de ellos. Por tanto, se abandona el ámbito de la estadística descriptiva, donde se calculan cuestiones como el promedio muestral para resumir simplemente un cuerpo de datos. En la estadística matemática se estudian las distribuciones muestrales de los estimadores.

Insesgadez

En principio, toda la distribución de muestreo de W se puede obtener dada la distribución probabilística de Y_i y la función h. Suele ser más fácil enfocarse en algunas características de la distribución de W cuando se evalúa como un estimador de θ. La primera propiedad importante de un estimador implica su valor esperado.

Estimador insesgado. Un estimador, W de θ, es un **estimador insesgado** si

$$\mathrm{E}(W) = \theta, \tag{C.3}$$

para todos los valores posibles de θ.

Si un estimador es insesgado, entonces su distribución de probabilidad tiene un valor esperado igual al parámetro que se supone estimará. La insesgadez *no* significa que la estimación que se obtuvo una muestra particular sea igual a θ, o incluso muy cercana a θ. Por el contrario, si se pudieran sacar *indefinidamente* muestras aleatorias en Y de la población, calcular una estimación cada vez y después promediarlas en todas las muestras aleatorias, se obtendría θ. Este experimento mental es abstracto, pues en la mayoría de las aplicaciones, sólo se tiene una muestra aleatoria con la cual trabajar.

Para un estimador que no es insesgado, se define su sesgo de la siguiente manera.

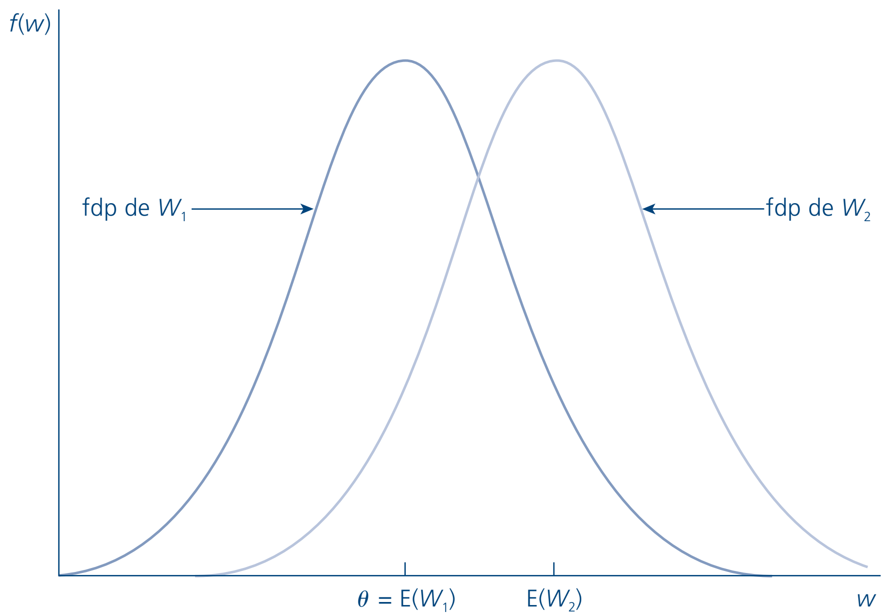
Sesgo de un estimador. Si W es un **estimador sesgado** de θ, su **sesgo** se define como

$$\mathrm{Sesgo}(W) \equiv \mathrm{E}(W) - \theta. \tag{C.4}$$

La figura C.1 muestra dos estimadores; el primero es insesgado y el segundo tiene un sesgo positivo.

FIGURA C.1

Un estimador insesgado, W_1, y un estimador con sesgo positivo, W_2.



La insesgadez de un estimador y la dimensión de cualquier sesgo posible dependen de la distribución de Y y de la función h. La distribución de Y suele estar fuera de control (aunque se suele elegir un *modelo* para esta distribución): puede estar determinado por la naturaleza o fuerzas sociales. Pero se tiene la opción de la regla h, y si se quiere un estimador, entonces se debe elegir h.

Por lo general, algunos estimadores pueden mostrar insesgadez. Por ahora se mostrará que el promedio muestral $\bar{Y}$ es un estimador insesgado de la media poblacional μ, sin importar la distribución de la población subyacente. Se utilizan las propiedades de los valores esperados (E.1 y E.2) que se cubrieron en la sección B.3:

$$\begin{aligned} \mathrm{E}(\bar{Y}) &= \mathrm{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\mathrm{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n} \mathrm{E}(Y_i)\right) \\ &= (1/n)\left(\sum_{i=1}^{n} \mu\right) = (1/n)(n\mu) = \mu. \end{aligned}$$

Para probar la hipótesis, se necesita estimar la varianza σ^2 de una población con media μ. Si $\{Y_1, \ldots, Y_n\}$ denota la muestra aleatoria de la población con $\mathrm{E}(Y) = \mu$ y $\mathrm{Var}(Y) = \sigma^2$, defina el estimador como

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{C.5}$$

lo cual suele recibir el nombre de **varianza muestral**. Se puede mostrar que S^2 es insesgada para σ^2: $\mathrm{E}(S^2) = \sigma^2$. La división entre $n - 1$, en lugar de n, se debe al hecho de que la media μ es estimada, más que conocida. Si se conociera μ un estimador insesgado de σ^2 sería $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$, pero μ rara vez se conoce en la práctica.

Aunque la insesgadez tiene un cierto atractivo como propiedad para un estimador, de hecho, su antónimo, "sesgado", tiene fuertes connotaciones negativas, no está exento de problemas. Una debilidad de la insesgadez es que algunos estimadores razonables, e incluso algunos muy buenos, no son insesgados. Un poco más adelante se verá un ejemplo.

Otra debilidad importante de la insesgadez es que existen estimadores insesgados que en realidad son muy pobres. Considere estimar la media μ de una población. En lugar de utilizar el promedio muestral $\bar{Y}$ para estimar μ, suponga que, después de recabar una muestra del tamaño n, se descartan todas las observaciones salvo la primera. Es decir, el estimador de μ es simplemente $W \equiv Y_1$. Este estimador es insesgado debido a que $\mathrm{E}(Y_1) = \mu$. Con algo de suerte, se piensa que ignorar todo excepto la primera observación, no es un método prudente para estimar: desecha la mayoría de la información en la muestra. Por ejemplo, con $n = 100$, se obtienen 100 resultados de la variable aleatoria Y, pero entonces se emplea sólo la primera de ellas para estimar $\mathrm{E}(Y)$.

La varianza de muestreo de los estimadores

El ejemplo al final de la subsección previa muestra que se necesitan criterios adicionales para evaluar los estimadores. La insesgadez o insesgamiento sólo asegura que la distribución muestral de un estimador tenga un valor medio igual al parámetro que se supone va a estimar. Esto está bien, pero también es necesario saber qué tan dispersa es la distribución de un estimador. Un estimador puede ser igual a θ, en promedio, pero también puede estar muy lejos con una probabilidad muy alta. En la figura C.2, W_1 y W_2 son estimadores insesgados de θ. Pero la distribución de W_1 está más centrada en torno a θ: la probabilidad de que W_1 sea mayor que cualquier distancia dada a partir de θ es menor que la probabilidad de que W_2 sea mayor que la misma distancia a partir de θ. Utilizar W_1 como estimador significa que es menos probable que se obtenga una muestra aleatoria que produzca una estimación muy alejada de θ.

Para resumir la situación que se muestra en la figura C2, se utiliza la varianza (o desviación estándar) de un estimador. Recuerde que esto da una sola medida de la dispersión en la distribución. La varianza de un estimador suele recibir el nombre de **varianza de muestreo**, pues es la varianza asociada con una distribución muestral. Recuerde que la varianza de muestreo no es una variable aleatoria; sino una constante, aunque podría ser desconocida.

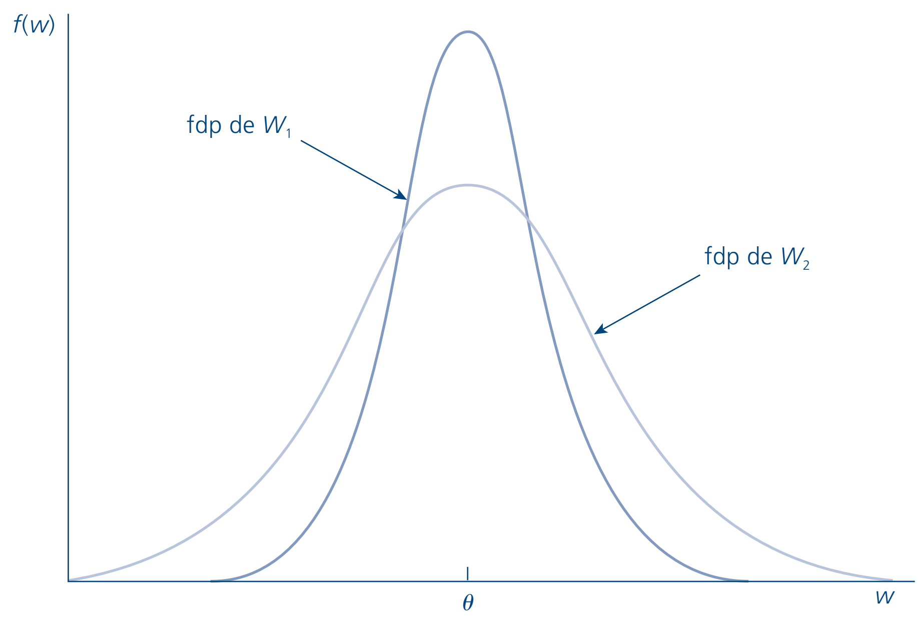
Se puede obtener ahora la varianza del promedio muestral para estimar la media μ de una población:

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n^2)\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n}\mathrm{Var}(Y_i)\right) \\ &= (1/n^2)\left(\sum_{i=1}^{n}\sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n.\end{aligned} \tag{C.6}$$

Observe cómo se emplean las propiedades de la varianza de las secciones B.3 y B.4 (VAR.2 y VAR.4), así como la independencia de Y_i. En suma: si $\{Y_i: i = 1, 2, \ldots, n\}$ es una muestra aleatoria de una población con una media μ y una varianza σ^2, entonces $\bar{Y}$ tiene la misma media que la población, pero su varianza muestral es igual a la varianza poblacional, σ^2, dividida entre el tamaño de la muestra.

FIGURA C.2

Distribuciones de dos estimadores insesgados de θ.



Una implicación importante de $\text{Var}(\bar{Y}) = \sigma^2/n$ es que puede acercarse mucho a cero al incrementar el tamaño muestral n. Esta es la característica clave de un estimador razonable, que se analizará en la sección C.3.

Como lo sugiere la figura C.2, entre los estimadores insesgados se prefiere el estimador con la varianza menor. Esto permite calcular ciertos estimadores de consideración. Para una muestra aleatoria de una población con media μ y varianza σ^2, se sabe que $\bar{Y}$ es insesgada y $\text{Var}(\bar{Y}) = \sigma^2/n$. ¿Qué hay del estimador Y_1, que es tan sólo la primera observación extraída? Debido a que Y_1 es una extracción aleatoria de la población, $\text{Var}(Y_1) = \sigma^2$. Por tanto, las diferencias entre $\text{Var}(Y_1)$ y $\text{Var}(\bar{Y})$ pueden ser grandes incluso para tamaños muestrales pequeños. Si $n = 10$, entonces $\text{Var}(Y_1)$ es 10 veces tan grande como $\text{Var}(\bar{Y}) = \sigma^2/10$. Esto da una forma formal de excluir Y_1 como estimador de μ.

Para enfatizar este punto, la tabla C.1 contiene el resultado de un pequeño estudio de simulación. Mediante el paquete estadístico Stata®, 20 muestras aleatorias de tamaño 10 se generaron de una distribución normal, con $\mu = 2$ y $\sigma^2 = 1$; aquí lo que interesa es estimar μ. Para cada una de las 20 muestras aleatorias se calculan dos estimaciones, y_1 y $\bar{y}$; estos valores están listados en la tabla C.1. Como se puede ver en la tabla, los valores de y_1 están mucho más dispersos que los de los rangos de $\bar{y}$: y_1 de -0.64 a 4.27, mientras que los rangos $\bar{y}$ sólo de 1.16 a 2.58. Además, en 16 de cada 20 casos, $\bar{y}$ es más cercana que y_1 a $\mu = 2$. El promedio de y_1 a través de las simulaciones es aproximadamente 1.89, mientras que el de $\bar{y}$ es 1.96. El hecho de que estos promedios sean cercanos a 2 ilustra el insesgamiento de ambos estimadores (y se podría hacer que estos promedios se acercaran más a 2 al hacer más de 20 réplicas). Pero comparar sólo los resultados promedio a través de las extracciones aleatorias oculta el hecho de que el promedio muestral $\bar{Y}$ es muy superior a Y_1 como estimador de μ.

TABLA C.1

Simulación de estimadores para una distribución Normal(μ,1) con μ = 2

Réplica	y_1	$\bar{y}$
1	−0.64	1.98
2	1.06	1.43
3	4.27	1.65
4	1.03	1.88
5	3.16	2.34
6	2.77	2.58
7	1.68	1.58
8	2.98	2.23
9	2.25	1.96
10	2.04	2.11
11	0.95	2.15
12	1.36	1.93
13	2.62	2.02
14	2.97	2.10
15	1.93	2.18
16	1.14	2.10
17	2.08	1.94
18	1.52	2.21
19	1.33	1.16
20	1.21	1.75

Eficiencia

Si se comparan las varianzas de $\bar{Y}$ y Y_1 de la sección anterior, se obtiene un ejemplo de un método general para comparar estimadores insesgados.

Eficiencia relativa. Si W_1 y W_2 son dos estimadores insesgados de θ, W_1 es eficiente en relación con W_2 cuando $\text{Var}(W_1) \leq \text{Var}(W_2)$ para toda θ, con desigualdad estricta al menos para un valor de θ.

Antes mostramos que, para estimar la media poblacional μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ para cualquier valor de σ^2 siempre que $n > 1$. Por tanto, $\bar{Y}$ es eficiente en relación con Y_1 para estimar μ. No siempre se puede elegir entre estimadores insesgados con base en el criterio de menor varianza: dados dos estimadores insesgados de θ, uno puede obtener una varianza menor de algunos valores de θ, mientras que el otro puede tener una varianza menor para otros valores de θ.

Si se restringe la atención a ciertas clases de estimadores, se puede mostrar que el promedio muestral tiene la varianza más pequeña. El problema C.2 pide mostrar que $\bar{Y}$ tiene la menor varianza entre todos los estimadores insesgados que son también funciones lineales de Y_1, Y_2, …, Y_n. Los supuestos consisten en que Y_i tiene una media y varianza común y que no están correlacionados a pares.

Si no se restringe la atención a estimadores insesgados, entonces no tiene sentido comparar las varianzas. Por ejemplo, cuando se estima la media poblacional μ, se puede usar un estimador trivial que es igual a cero, sin importar la muestra que se extraiga. Como es natural, la varianza de este estimador es cero (dado que es el mismo valor para toda muestra aleatoria). Pero el sesgo del estimador es $-\mu$, así que es un estimador muy pobre cuando $|\mu|$ es grande.

Una forma de comparar los estimadores que no necesariamente son insesgados, es calcular el **error cuadrático medio (ECM)** de los estimadores. Si W es un estimador de θ, entonces el ECM de W se define como $\text{ECM}(W) = \text{E}[(W - \theta)^2]$. El ECM mide, en promedio, qué tan alejado está el estimador de θ. Se puede mostrar que $\text{ECM}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$, así que $\text{ECM}(W)$ depende de la varianza y el sesgo (si lo hubiera). Esto permite comparar dos estimadores cuando uno o ambos son sesgados.

C.3 Propiedades asintóticas o de muestra grande de los estimadores

En la sección C.2 se encontró el estimador Y_1 para la media poblacional μ, y se vio que aunque es insesgado, es un estimador pobre debido a que su varianza puede ser mucho más grande que la media muestral. Una característica notable de Y_1 es que tiene la misma varianza para cualquier tamaño de muestra. Parece razonable exigir que el tamaño de la muestra aumente para que cualquier procedimiento de estimación mejore. Para estimar una media poblacional μ, $\bar{Y}$ mejora en el sentido de que la varianza disminuye a medida que n crece; Y_1 no mejora en este sentido.

Se pueden descartar ciertos estimadores inútiles al estudiar las propiedades *asintóticas* o de *muestra grande* de los estimadores. Además, se puede decir algo positivo acerca de los estimadores que no son insesgados y cuyas varianzas no son fáciles de encontrar.

El análisis asintótico implica aproximar las características de la distribución muestral de un estimador. Estas aproximaciones dependen del tamaño de la muestra. Por desgracia, necesariamente hay un límite en lo que se puede decir acerca de qué tan "grande" necesita ser un tamaño muestral para que el análisis asintótico sea apropiado; esto depende de la distribución de la población subyacente. Pero se sabe que las aproximaciones de muestras grandes funcionan bien para tamaños muestrales tan pequeños como $n = 20$.

Consistencia

La primera propiedad asintótica de los estimadores concierne a qué tan lejos es probable que esté el estimador del parámetro que se supone estimará, a medida que el tamaño de la muestra aumenta indefinidamente.

Consistencia. Sea W_n un estimador de θ basado en una muestra $Y_1, Y_2, \ldots, Y_n$ del tamaño n. Entonces, W_n es un **estimador consistente** de θ si para toda $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ cuando } n \to \infty. \quad \boxed{\text{C.7}}$$

Si W_n no es consistente para θ, entonces se dice que es **inconsistente**.

Cuando W_n es consistente, también se dice que θ es el **límite en probabilidad** de W_n, escrito como $\text{plim}(W_n) = \theta$.

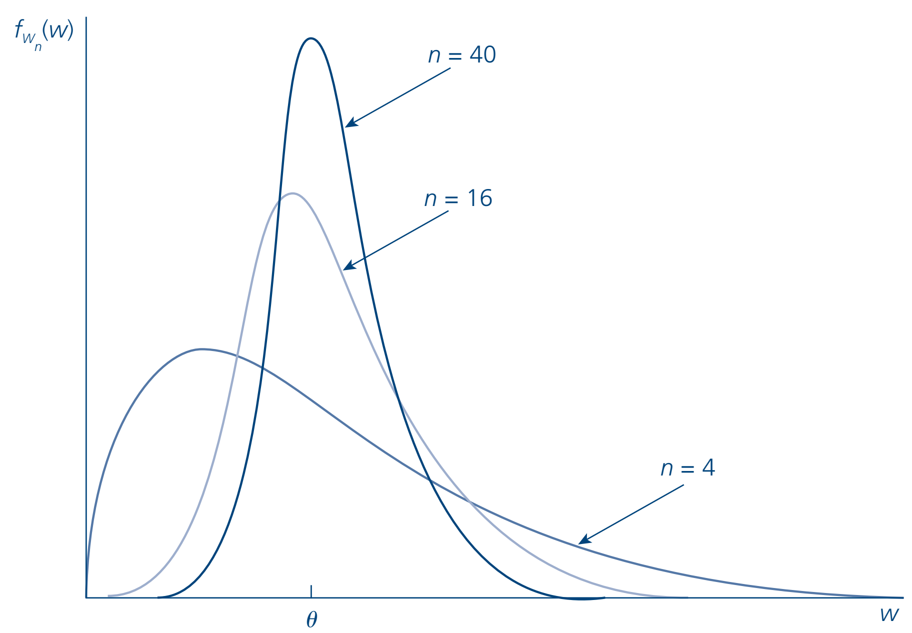
Al igual que el insesgamiento, que es una característica de un estimador de un tamaño de la muestra determinado, la consistencia implica el comportamiento de la distribución muestral de un estimador a medida que el tamaño muestral n aumenta. Para enfatizar esto, en esta definición, se ha indexado el estimador por el tamaño de la muestra y se continuará con esta convención a lo largo de esta sección.

La ecuación (C.7) parece técnica y puede ser muy difícil de establecer con base en principios fundamentales de probabilidad. Por el contrario, interpretar (C.7) es fácil. Significa que la distribución de W_n se concentra cada vez más en torno de θ, lo cual a grandes rasgos, significa que para tamaños de la muestra mayores, es cada vez menos probable que W_n se aleje mucho de θ. Esta tendencia se ilustra en la figura C.3.

Si un estimador no es consistente, entonces no es útil saber de θ, incluso con una cantidad ilimitada de datos. Por esta razón, la consistencia es un requisito mínimo de un estimador utilizado en estadística o econometría. Se encontrarán estimadores consistentes bajo ciertos supuestos e inconsistentes cuando tales supuestos fallen. Cuando los estimadores son inconsistentes, por

FIGURA C.3

Distribuciones de muestreo de un estimador consistente para tres tamaños de muestras.



lo general, se encuentran sus límites de probabilidad y será importante saber qué tan lejos se encuentran éstos de θ.

Como se observó antes, los estimadores insesgados no son necesariamente consistentes, pero aquellos con varianzas que se reducen a cero a medida que el tamaño muestral aumenta *son* consistentes. Esto se puede expresar formalmente: si W_n es un estimador insesgado de θ y $\text{Var}(W_n) \rightarrow 0$ cuando $n \rightarrow \infty$, entonces $\text{plim}(W_n) = \theta$. Los estimadores insesgados que utilizan la muestra de datos completa, por lo general, tienen una varianza que se reduce a cero a medida que el tamaño muestral aumenta y, por tanto, son consistentes.

Un buen ejemplo de un estimador consistente es el promedio de una muestra aleatoria extraída de la población con media μ y varianza σ^2. Ya se ha mostrado que el promedio muestral es insesgado para μ. En la ecuación (C.6) se derivó $\text{Var}(\bar{Y}_n)=\sigma^2/n$ para cualquier tamaño de muestra n. Por tanto, $\text{Var}(\bar{Y}_n) \rightarrow 0$ cuando $n \rightarrow \infty$, así, $\bar{Y}_n$ es un estimador consistente de μ (además de ser insesgado).

La conclusión de que $\bar{Y}_n$ es consistente para μ es válida incluso si $\text{Var}(\bar{Y}_n)$ no existiera. Este resultado clásico se conoce como la **ley de los grandes números (LGN)**.

Ley de los grandes números. Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias independientes, idénticamente distribuidas con media μ. Entonces,

$$\text{plim}(\bar{Y}_n) = \mu. \tag{C.8}$$

La ley de los grandes números significa que, si interesa estimar el promedio poblacional μ, es posible aproximarse arbitrariamente a μ si se elige una muestra suficientemente grande. Este resultado fundamental se puede combinar con las propiedades básicas de los plim para mostrar que estimadores muy complejos son consistentes.

Propiedad PLIM.1: Sea θ un parámetro y defina un nuevo parámetro, $\gamma = g(\theta)$, para alguna función *continua* $g(\theta)$. Suponga que $\text{plim}(W_n) = \theta$. Defina un estimador de γ por $G_n = g(W_n)$. Entonces,

$$\text{plim}(G_n) = \gamma. \tag{C.9}$$

Esto suele expresarse como

$$\text{plim}\ g(W_n) = g(\text{plim}\ W_n) \tag{C.10}$$

para una función continua $g(\theta)$.

El supuesto de que $g(\theta)$ es continua, es un requisito técnico que suele describirse de forma no técnica como "una función que puede graficarse sin levantar el lápiz del papel". Debido a que todas las funciones que se encuentran en este libro son continuas, no se ofrece una definición formal de una función continua. Ejemplos de funciones continuas son $g(\theta) = a + b\theta$ para constantes a y b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, y muchas variantes de éstas. No será necesario mencionar nuevamente el supuesto de continuidad.

Como ejemplo importante de un estimador consistente pero sesgado, considere la desviación estándar, σ, de una población con media μ y varianza σ^2. Se ha afirmado que la varianza muestral $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$ es insesgada para σ^2. Mediante la ley de los grandes números y un poco de álgebra, S_n^2 se puede mostrar que es consistente para σ^2. El estimador natural de

$\sigma = \sqrt{\sigma^2}$ es $S_n = \sqrt{S_n^2}$ (donde la raíz cuadrada es siempre la raíz cuadrada positiva). S_n, que recibe el nombre de **desviación estándar muestral**, *no* es un estimador insesgado debido a que el valor esperado de la raíz cuadrada *no* es la raíz cuadrada del valor esperado (vea la sección B.3). No obstante, por PLIM.1, plim $S_n = \sqrt{\text{plim } S_n^2} = \sqrt{\sigma^2} = \sigma$, así que S_n es un estimador consistente de σ.

He aquí otras propiedades útiles del límite en probabilidad:

Propiedad PLIM.2: Si $\text{plim}(T_n) = \alpha$ y $\text{plim}(U_n) = \beta$, entonces

i) $\text{plim}(T_n + U_n) = \alpha + \beta$;

ii) $\text{plim}(T_n U_n) = \alpha\beta$;

iii) $\text{plim}(T_n/U_n) = \alpha/\beta$, siempre que $\beta \neq 0$.

Estas tres propiedades sobre los límites en probabilidad permiten combinar estimadores consistentes en varias formas para obtener otros estimadores consistentes. Por ejemplo, sea $\{Y_1, \ldots, Y_n\}$ una muestra aleatoria de tamaño n sobre los ingresos anuales de la población de trabajadores con una educación de bachillerato y denótese la media poblacional por μ_Y. Sea $\{Z_1, \ldots, Z_n\}$ una muestra aleatoria del ingreso anual de la población de trabajadores con una educación universitaria y denote la media poblacional con μ_Z. Se quiere estimar la diferencia porcentual en el ingreso anual entre los dos grupos, la cual es $\gamma = 100 \cdot (\mu_Z - \mu_Y)/\mu_Y$. (Este es el porcentaje mediante el cual el ingreso promedio de los graduados universitarios difiere del ingreso promedio para los graduados de bachillerato.) Debido a que $\bar{Y}_n$ es consistente para μ_Y y $\bar{Z}_n$ es consistente para μ_Z, se desprende de PLIM.1 y el inciso iii) de PLIM.2 que

$$G_n \equiv 100 \cdot (\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

es un estimador consistente de γ. G_n es sólo la diferencia porcentual entre $\bar{Z}_n$ y $\bar{Y}_n$ en la muestra, así que es un estimador natural. G_n no es un estimador insesgado de γ, pero sigue siendo un buen estimador salvo, quizá, cuando n es pequeña.

Normalidad asintótica

La consistencia es una propiedad de los estimadores puntuales. Aunque ésta expresa que la distribución del estimador se está colapsando en torno al parámetro a medida que el tamaño de la muestra aumenta, no expresa esencialmente nada acerca de la *forma* de esa distribución para un tamaño muestral dado. Para construir estimadores de intervalo y pruebas de hipótesis, se necesita una forma de aproximar la distribución de los estimadores. La mayoría de los estimadores econométricos tiene distribuciones que se aproximan bien mediante una distribución normal para muestras grandes, lo que motiva la siguiente definición.

Normalidad asintótica. Sea $\{Z_n: n = 1, 2, \ldots\}$ una secuencia de variables aleatorias, tal que para todos los números z,

$$P(Z_n \leq z) \rightarrow \Phi(z) \text{ as } n \rightarrow \infty, \tag{C.11}$$

donde $\Phi(z)$ es la función de distribución acumulada normal estándar. Entonces, se dice que, Z_n tiene una *distribución normal estándar asintótica*. En este caso suele escribirse $Z_n \overset{a}{\sim} \text{Normal}(0,1)$. (La "*a*" sobre la tilde significa "asintóticamente" o "aproximadamente".)

La propiedad (C.11) significa que la función de Z_n se acerca cada vez más a la fda de la distribución normal estándar a medida que el tamaño de la muestra n aumenta. Cuando la **normalidad**

asintótica se mantiene para n grandes se tiene la aproximación $P(Z_n \leq z) \approx \Phi(z)$. Por tanto, las probabilidades concernientes a Z_n se pueden aproximar mediante probabilidades normales estándar.

El **teorema del límite central (TLC)** es uno de los resultados más poderosos en probabilidad y estadística. Indica que el promedio de una muestra aleatoria para *cualquier* población (con varianza finita), cuando se estandariza, tiene una distribución normal estándar asintótica.

Teorema del límite central. Sea $\{Y_1, Y_2, \ldots, Y_n\}$ una muestra aleatoria con una media μ y una varianza σ^2. Entonces,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \tag{C.12}$$

tiene una distribución normal estándar asintótica.

La variable Z_n en (C.12) es la versión estandarizada de $\bar{Y}_n$: se ha sustraído $E(\bar{Y}_n) = \mu$ y dividido entre $sd(\bar{Y}_n) = \sigma/\sqrt{n}$. Por tanto, sin importar la distribución poblacional de Y, Z_n tiene media cero y varianza uno, lo cual coincide con la media y varianza de la distribución normal estándar. Es de notar que toda la distribución de Z_n se acerca arbitrariamente a la distribución normal estándar a medida que n aumenta.

Se puede escribir la variable estandarizada en la ecuación (C.12) como $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, lo cual muestra que se debe multiplicar la diferencia entre la media muestral y la media poblacional por la raíz cuadrada del tamaño de la muestra para obtener una distribución limitante útil. Sin la multiplicación por $\sqrt{n}$, se tendría sólo $(\bar{Y}_n - \mu)/\sigma$, que converge en probabilidad a cero. En otras palabras, la distribución de $(\bar{Y}_n - \mu)/\sigma$ simplemente se colapsa a un solo punto conforme $n \to \infty$, lo cual, se sabe, no puede ser una buena aproximación a la distribución de $(\bar{Y}_n - \mu)/\sigma$ para tamaños muestrales razonables. Multiplicar por $\sqrt{n}$ asegura que la varianza de Z_n permanece constante. Prácticamente, suele tratarse a $\bar{Y}_n$ como distribuida normal y aproximadamente con media μ y varianza σ^2/n, y esto provee el procedimiento estadístico correcto, debido a que lleva a la variable estandarizada en la ecuación (C.12).

La mayoría de los estimadores encontrados en estadística y econometría se puede escribir como funciones de promedios muestrales, en cuyo caso se puede aplicar la ley de los grandes números y el teorema del límite central. Cuando dos estimadores consistentes tiene distribuciones normales asintóticas, se elige el estimador con la varianza asintótica menor.

Además, el promedio muestral estandarizado en (C.12), muchos otros estadísticos que dependen de promedios muestrales resultan ser asintóticamente normales. Uno importante es el que se obtiene al reemplazar σ con su estimador consistente S_n en la ecuación (C1.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \tag{C.13}$$

también tiene una distribución normal estándar aproximada para n grande. Las distribuciones exactas (muestra finita) de (C.12) y (C.13) definitivamente no son iguales, pero la diferencia suele ser tan pequeña como para ignorarse cuando n es grande.

A lo largo de esta sección, cada estimador tiene un subíndice n para enfatizar la naturaleza del análisis muestral grande o asintótico. Continuar esta convención desorganizaría la notación sin ofrecer información adicional, una vez que se comprendieran los principios básicos del análisis asintótico. Por tanto, se descartará el subíndice n, con la esperanza de que se recuerde que los estimadores dependen del tamaño muestral, y las propiedades como consistencia y normalidad asintótica se refieren al crecimiento sin límite del tamaño de la muestra.

C.4 Métodos generales para estimar parámetros

Hasta este punto se ha utilizado el promedio muestral para ilustrar las propiedades finitas y de muestra grande de los estimadores. Es natural preguntar: ¿existen métodos generales para estimar que produzcan estimadores con buenas propiedades, como insesgadez, consistencia y eficiencia?

La respuesta es sí. Un análisis detallado de los diversos métodos que existen para estimar está más allá del alcance de este libro; aquí sólo se ofrecerá un análisis informal. Un análisis detallado se puede encontrar en Larsen y Marx (1986, capítulo 5).

El método de momentos

Dado un parámetro θ que aparece en una distribución poblacional, suele haber diversas formas de obtener estimadores insesgados y consistentes de θ. No sería práctico, intentar todas las diferentes posibilidades y compararlas con base en los criterios de las secciones C.2 y C.3. Por fortuna, se ha mostrado que algunos métodos tienen buenas propiedades generales y, en la mayor parte, la lógica en la que se basan es atractiva, desde un punto de vista intuitivo.

En secciones anteriores, se ha estudiado el promedio muestral como un estimador insesgado del promedio poblacional y la varianza muestral como un estimador insesgado de la varianza poblacional. Estos estimadores son ejemplos de los estimadores del **método de momentos**. En general, la estimación del método de momentos es la siguiente. El parámetro θ está relacionado con algún valor esperado en la distribución de Y, que suele ser $E(Y)$ o $E(Y^2)$ (aunque en ocasiones se utilizan opciones más exóticas). Suponga, por ejemplo, que el parámetro de interés, θ, está relacionado con la media poblacional como $\theta = g(\mu)$ para alguna función g. Debido a que el promedio muestral $\bar{Y}$ es insesgado y es un estimador consistente de μ, es natural remplazar μ con $\bar{Y}$, lo cual genera un estimador $g(\bar{Y})$ de θ. El estimador $g(\bar{Y})$ es consistente para θ y si $g(\mu)$ es una función lineal de μ, entonces $g(\bar{Y})$ es insesgado también. Lo que se ha hecho es remplazar el momento poblacional, μ, con su contraparte muestral, $\bar{Y}$. De ahí el nombre de "método de momentos".

Se abarcan dos estimadores adicionales del método de momentos que serán útiles para la discusión del análisis de momentos. Recuerde que la covarianza entre dos variables aleatorias X y Y se define como $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. El método de momentos sugiere estimar σ_{XY} por $n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$. Este es un estimador consistente de σ_{XY}, pero resulta ser sesgado esencialmente por la misma razón de que la varianza muestral es sesgada si n, y no $n - 1$, se usa como el divisor. La **covarianza muestral** se define como

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}). \tag{C.14}$$

Se puede mostrar que este es un estimador insesgado de σ_{XY}. (Remplazar n con $n - 1$ no produce diferencia a medida que el tamaño muestral aumenta de forma indefinida, así que este estimador sigue siendo consistente).

Como se analizó en la sección B.4, la covarianza entre dos variables suele ser difícil de interpretar. Por lo general, lo que interesa más es la correlación. Debido a que la correlación poblacional es $\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$, el método de momentos sugiere que estimar ρ_{XY} como

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)^{1/2} \left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)^{1/2}}, \tag{C.15}$$

lo cual recibe el nombre de **coeficiente de correlación muestral** (o *correlación muestral*). Observe que se ha cancelado la división mediante $n - 1$ en la covarianza muestral y las desviaciones estándar muestrales. De hecho, se puede dividir cada una entre n, y se llegaría finalmente a la misma fórmula.

Se puede mostrar que el coeficiente de correlación muestral siempre está en el intervalo $[-1,1]$, como debe ser. Debido a que S_{XY}, S_X y S_Y son consistentes para el parámetro poblacional correspondiente, R_{XY} es un estimador consistente de la correlación poblacional, ρ_{XY}. No obstante, R_{XY} es un estimador sesgado por dos razones. Primero, S_X y S_Y son estimadores sesgados de σ_X y σ_Y, respectivamente. Segundo, R_{XY} es una razón de estimadores, así que no puede ser insesgado, aunque S_X y S_Y lo fueran. Para los fines de este libro, esto no es importante, a pesar del hecho de que un resultado clásico en la estadística matemática es que no existe un estimador insesgado de ρ_{XY}.

Máxima verosimilitud

Otro método general para estimar es el método de *máxima verosimilitud*, un tema que se cubre en muchos cursos introductorios de estadística. Un breve resumen sobre el caso más simple bastará aquí. Sea $\{Y_1, Y_2, \ldots, Y_n\}$ una muestra aleatoria de la distribución poblacional $f(y;\theta)$. Debido al supuesto de muestreo aleatorio, la distribución conjunta de $\{Y_1, Y_2, \ldots, Y_n\}$ es simplemente el producto de las densidades: $f(y_1;\theta)f(y_2;\theta) \cdots f(y_n;\theta)$. En el caso discreto, éste es $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$. Ahora, defínase la *función de verosimilitud* como

$$L(\theta;Y_1, \ldots, Y_n) = f(Y_1;\theta)f(Y_2;\theta) \cdots f(Y_n;\theta),$$

que es una variable aleatoria debido a que depende del resultado de la muestra aleatoria $\{Y_1, Y_2, \ldots, Y_n\}$. El **estimador de máxima verosimilitud** de θ, llámese W, es el valor de θ que maximiza la función de verosimilitud. (Ésta es la razón de que se escriba L en función de θ, seguida por la muestra aleatoria). Por supuesto, este valor depende de la muestra aleatoria. El principio de máxima verosimilitud dice que, de todos los valores posibles de θ, se debe elegir el valor que haga que la verosimilitud de los datos observados sea mayor. Intuitivamente, éste es un método razonable para estimar θ.

Por lo general, es más conveniente trabajar con la *función log-probabilidad*, que se obtiene mediante el logaritmo natural de la función de verosimilitud:

$$\log [L(\theta; Y_1, \ldots, Y_n)] = \sum_{i=1}^{n} \log [f(Y_i; \theta)], \tag{C.16}$$

donde se utiliza el hecho de que el logaritmo del producto es la suma de los logaritmos. Dado que (C.16) es la suma de variables aleatorias idénticamente distribuidas e independientes, analizar los estimadores que provienen de (C.16) es relativamente fácil.

La estimación de máxima verosimilitud (EMV) suele ser consistente y, en ocasiones, insesgada. Pero existen muchos otros estimadores. El atractivo tan difundido de EMV se debe a que, en general es el estimador más asintóticamente eficiente cuando el modelo poblacional $f(y;\theta)$ se especifica correctamente. Además, el EMV algunas veces es el **estimador insesgado de varianza mínima**, es decir, tiene la varianza menor entre todos los estimadores insesgados de θ. [Vea Larsen y Marx (1986, capítulo 5) para verificar estas afirmaciones.]

En el capítulo 17 se necesitará una máxima verosimilitud para estimar los parámetros de más modelos econométricos avanzados. En econometría casi siempre se tiene interés en la distribución de Y condicional en un conjunto de variables explicativas, por ejemplo, $X_1, X_2, \ldots, X_k$. Entonces, se remplaza la densidad en (C.16) con $f(Y_i|X_{i1}, \ldots, X_{ik}; \theta_1, \ldots, \theta_p)$, donde a esta densidad se le permite depender de p parámetros, $\theta_1, \ldots, \theta_p$. Por fortuna, para la aplicación exitosa de los métodos de máxima verosimilitud, no es necesaria una mayor profundización en las cuestiones computacionales o teorías estadísticas de muestra grande. Wooldridge (2002, capítulo 13) cubre la teoría de la estimación de máxima verosimilitud.

Mínimos cuadrados

Un tercer tipo de estimador, y que desempeña la mayor de las funciones en el libro, es el llamado **estimador de mínimos cuadrados**. Ya se ha visto un ejemplo de mínimos cuadrados: la media muestral, $\bar{Y}$, es un estimador de mínimos cuadrados de la media poblacional, μ. Ya se sabe que $\bar{Y}$ es un estimador del método de momentos. ¿Qué lo convierte en un estimador de mínimos cuadrados? Se puede mostrar que el valor de m que compone la suma de las desviaciones cuadradas

$$\sum_{i=1}^{n}(Y_i - m)^2$$

tan pequeña como sea posible es $m = \bar{Y}$. Demostrar esto no es difícil, por lo que se omite el álgebra.

Para algunas distribuciones importantes, incluidas la normal y la de Bernoulli, el promedio muestral $\bar{Y}$ es también el estimador de máxima verosimilitud de la media poblacional μ. Por tanto, los principios de los mínimos cuadrados, método de momentos y máxima verosimilitud suelen generar el *mismo* estimador. En otros casos, los estimadores son similares, mas no idénticos.

C.5 Estimación de intervalos e intervalos de confianza

La naturaleza de la estimación de intervalos

Una estimación puntual obtenida de una muestra particular no proporciona, por sí misma, suficiente información para probar teorías económicas o para análisis de políticas. Una estimación puntual puede ser el mejor cálculo del valor poblacional, pero, por su naturaleza, no ofrece información sobre qué tanta "probabilidad" hay de que la estimación esté cercana al parámetro de la población. Por ejemplo, suponga que un investigador reporta, con base en una muestra aleatoria de trabajadores, que los subsidios a la capacitación laboral incrementan el salario por hora en 6.4%. ¿Cómo se sabrá si esto es semejante o no al efecto en la población de trabajadores que pudieran capacitarse? Debido a que se ignora el valor poblacional, no se puede saber qué tan próxima es una estimación para una muestra en particular. No obstante, se pueden hacer planteamientos que comprendan probabilidades, y aquí es donde entra la estimación de intervalo.

Ya se conoce una forma de evaluar la incertidumbre en un estimador: determinar su desviación estándar muestral. Revelar la desviación estándar del estimador, junto con la estimación puntual, ofrece cierta información sobre la precisión de la estimación en cuestión. No obstante, aunque se ignore el problema de la dependencia de la desviación estándar sobre los parámetros de población desconocidos, dar cuenta de la desviación estándar junto con la estimación puntual no ofrece ningún planteamiento directo sobre dónde es probable que el valor poblacional se ubique con relación a la estimación. Esta limitación se supera al construir el **intervalo de confianza**.

Se ilustra el concepto de intervalo de confianza con un ejemplo. Suponga que la población tiene una distribución Normal(μ,1) y sea $\{Y_1, \ldots, Y_n\}$ una muestra aleatoria de esta población. (Se supone que la varianza poblacional es conocida y que es igual a la unidad, para fines ilustrativos; más adelante se mostrará qué hacer en el caso más realista en que la varianza se ignora.)

El promedio muestral, $\bar{Y}$, tiene una distribución normal con una media μ y una varianza $1/n$: $\bar{Y} \sim \text{Normal}(\mu,1/n)$. A partir de esto, se puede estandarizar $\bar{Y}$, y, debido a que la versión estándar de $\bar{Y}$ tiene una distribución normal estándar, se tiene

$$P\left(-1.96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = .95.$$

El evento entre paréntesis es idéntico al evento $\bar{Y} - 1.96/\sqrt{n} < \mu < \bar{Y} + 1.96/\sqrt{n}$, así que

$$P(\bar{Y} - 1.96/\sqrt{n} < \mu < \bar{Y} + 1.96/\sqrt{n}) = .95. \quad \text{C.17}$$

La ecuación (C.17) es interesante, debido a que indica que la probabilidad de que el intervalo aleatorio $[\bar{Y} - 1.96/\sqrt{n}, \bar{Y} + 1.96/\sqrt{n}]$ contenga la media poblacional μ es .95, o 95%. Esta información permite construir una *estimación de intervalo* de μ, lo cual se obtiene al insertar el resultado muestral del promedio, $\bar{y}$. Por tanto,

$$[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}] \quad \text{C.18}$$

es un ejemplo de una estimación de intervalo de μ. También recibe el nombre de un intervalo de confianza a 95%. Una notación abreviada para este intervalo es $\bar{y} \pm 1.96/\sqrt{n}$.

El intervalo de confianza en la ecuación (C.18) es fácil de calcular, una vez que se observan los datos muestrales $\{y_1, y_2, \dots, y_n\}$; $\bar{y}$ es el único factor que depende de los datos. Por ejemplo, suponga que $n = 16$ y el promedio de las 16 observaciones es 7.3. Entonces, el intervalo de confianza a 95% para μ es $7.3 \pm 1.96/\sqrt{16} = 7.3 \pm .49$, lo cual se puede escribir en forma de intervalo como [6.81,7.79]. Por construcción, $\bar{y} = 7.3$ está en el centro de este intervalo.

A diferencia de este cálculo, el significado de intervalo de confianza es más difícil de comprender. Cuando se dice que la ecuación (C.18) está a un intervalo de confianza a 95% para μ, significa que el intervalo *aleatorio*

$$[\bar{Y} - 1.96/\sqrt{n}, \bar{Y} + 1.96/\sqrt{n}] \quad \text{C.19}$$

contiene a μ con probabilidad de .95. En otras palabras, antes de que se extraiga la muestra aleatoria, existe una probabilidad de 95% de que (C.19) contenga a μ. La ecuación (C.19) es un ejemplo de un **estimador de intervalo**. Es un intervalo aleatorio, dado que los puntos finales cambian con diferentes muestras.

Un intervalo de confianza suele interpretarse de la siguiente manera: "La probabilidad de que μ esté en el intervalo (C.18) es de .95." Esto es incorrecto. Una vez que la muestra se observa y $\bar{y}$ se calcula, los límites del intervalo de confianza son simplemente números (6.81 y 7.79 en el ejemplo anterior). El parámetro poblacional, μ, aunque desconocido, es también algún otro número. Por tanto, μ está o no en el intervalo (C.18) (y nunca se sabrá con certidumbre cuál es el caso). La probabilidad no desempeña ninguna función una vez que el intervalo de confianza se ha calculado para los datos particulares a la mano. La interpretación probabilística proviene del hecho que para 95% de todas las muestras aleatorias, el intervalo de confianza construido contendrá a μ.

Para enfatizar el significado de un intervalo de confianza, la tabla C.2 contiene los cálculos para 20 muestras aleatorias (o réplicas) de la distribución Normal(2,1) con un tamaño muestral $n = 10$. Para cada una de las 20 muestras, se obtiene, $\bar{y}$ y (C.18) se calcula como $\bar{y} \pm 1.96/\sqrt{10} = \bar{y} \pm .62$ (cada uno de ellos redondeado a dos decimales). Como puede verse, el intervalo cambia con cada muestra aleatoria. Diecinueve de los 20 intervalos contienen al valor poblacional de μ. Sólo para la réplica número 19 μ no está en el intervalo de confianza. En otras palabras, 95% de las muestras genera un intervalo de confianza que contiene a μ. Esto no tendría que ser el caso con sólo 20 réplicas, pero así se ideó para esta simulación en particular.

TABLA C.2

Intervalos de confianza simulados de una distribución Normal(μ,1) con $\mu = 2$

Réplica	$\bar{y}$	Intervalo a 95%	¿Contiene a μ?
1	1.98	(1.36,2.60)	Sí
2	1.43	(0.81,2.05)	Sí
3	1.65	(1.03,2.27)	Sí
4	1.88	(1.26,2.50)	Sí
5	2.34	(1.72,2.96)	Sí
6	2.58	(1.96,3.20)	Sí
7	1.58	(.96,2.20)	Sí
8	2.23	(1.61,2.85)	Sí
9	1.96	(1.34,2.58)	Sí
10	2.11	(1.49,2.73)	Sí
11	2.15	(1.53,2.77)	Sí
12	1.93	(1.31,2.55)	Sí
13	2.02	(1.40,2.64)	Sí
14	2.10	(1.48,2.72)	Sí
15	2.18	(1.56,2.80)	Sí
16	2.10	(1.48,2.72)	Sí
17	1.94	(1.32,2.56)	Sí
18	2.21	(1.59,2.83)	Sí
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Sí

Intervalos de confianza para la media de una población normalmente distribuida

El intervalo de confianza derivado en la ecuación (C.18) ayuda a ilustrar cómo construir e interpretar intervalos de confianza. En la práctica, la ecuación (C.18) no es muy útil para la media de una población normal, pues se supone que el valor conocido de la varianza es uno. Es fácil extender (C.18) al caso en el que se sabe que la desviación estándar σ toma cualquier valor: el intervalo de confianza a 95% es

$$[\bar{y} - 1.96\sigma/\sqrt{n}, \bar{y} + 1.96\sigma/\sqrt{n}]. \tag{C.20}$$

Por tanto, siempre que se conoce σ, se puede decir que se puede construir fácilmente un intervalo de confianza para μ. Al considerar que σ, es desconocida, se debe utilizar una estimación. Sea

$$s = \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2} \tag{C.21}$$

la cual denota la desviación estándar muestral. Después, se obtiene un intervalo de confianza que depende por completo de los datos observados al remplazar σ en la ecuación (C.20) con su estimación, s. Por desgracia, esto no preserva el nivel de confianza a 95% debido a que s depende de la muestra particular. En otras palabras, el intervalo aleatorio $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ ya no contiene a μ con probabilidad de .95 debido a que la constante σ se remplazó con la variable aleatoria S.

¿Cómo se debe proceder? En lugar de utilizar la distribución normal estándar, se debe depender de la distribución t. La distribución t surge del hecho de que

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{C.22}$$

donde $\bar{Y}$ es el promedio muestral y S es la desviación estándar muestral de la muestra aleatoria $\{Y_1, \ldots, Y_n\}$. No se probará (C.22); una demostración detallada puede encontrarse en varios libros [por ejemplo, Larsen y Marx (1986, capítulo 7)].

Para construir un intervalo de confianza a 95%, sea c el 97.5-ésimo percentil en la distribución t_{n-1}. En otras palabras, c es el valor tal que 95% del área de la distribución t_{n-1} está entre $-c$ y c: $P(-c < t_{n-1} < c) = .95$. (El valor de c depende de los grados de libertad $n - 1$, aunque no se explicite). La elección de c se ilustra en la figura C.4. Una vez que c se ha elegido de manera apropiada, el intervalo aleatorio $[\bar{Y} - c\cdot S/\sqrt{n}, \bar{Y} + c\cdot S/\sqrt{n}]$ contiene a μ con probabilidad de .95. Para una muestra particular, el intervalo de confianza a 95% se calcula como

$$[\bar{y} - c\cdot s/\sqrt{n}, \bar{y} + c\cdot s/\sqrt{n}]. \tag{C.23}$$

Los valores de c para varios grados de libertad se pueden obtener de la tabla G.2 en el apéndice G. Por ejemplo, si $n = 20$, de manera que los *gl* son $n - 1 = 19$, entonces $c = 2.093$. Por tanto, el intervalo de confianza a 95 % es $[\bar{y} \pm 2.093(s/\sqrt{20})]$, donde $\bar{y}$ y s son los valores obtenidos de la muestra. Aunque $s = \sigma$ (lo cual es muy improbable), el intervalo de confianza en (C.23) es más amplio que el de (C.20) debido a que $c > 1.96$. Para grados de libertad pequeños, (C.23) es mucho más amplio.

En términos más generales, sea c_α el percentil $100(1 - \alpha)$ en la distribución t_{n-1}. Entonces, se obtiene un intervalo de confianza a $100(1 - \alpha)\%$ como

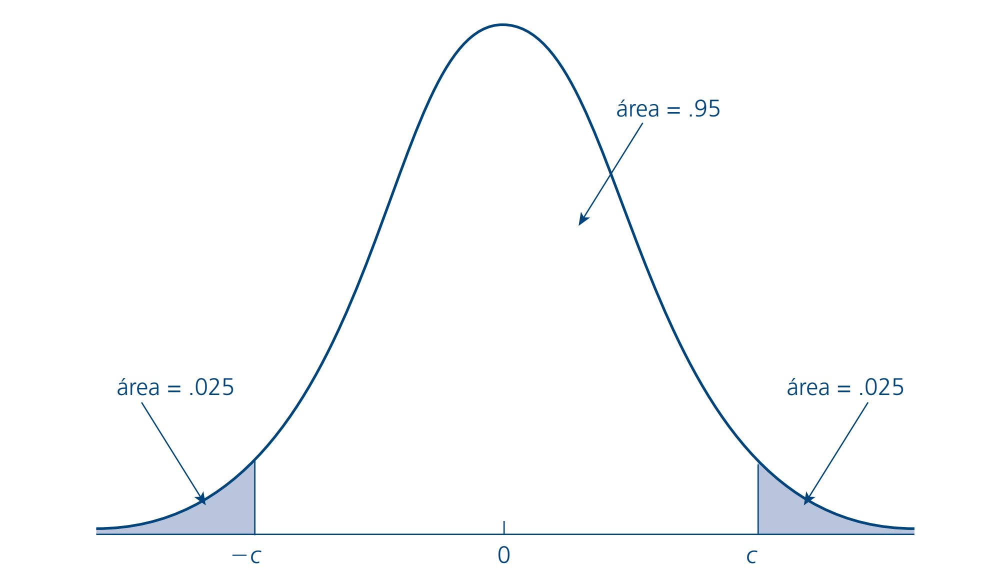
$$[\bar{y} - c_{\alpha/2}s/\sqrt{n}, \bar{y} + c_{\alpha/2}s/\sqrt{n}]. \tag{C.24}$$

Obtener $c_{\alpha/2}$ requiere que se elija α y se conozcan los grados de libertad $n - 1$; entonces, se puede utilizar la tabla G.2. En general, se considerarán los intervalos de confianza a 95%.

Existe una sencilla forma de recordar cómo construir un intervalo de confianza para la media de una distribución normal. Recuerde que $\mathrm{sd}(\bar{Y}) = \sigma/\sqrt{n}$. Por tanto, $s/\sqrt{n}$ es la estimación puntual de $\mathrm{sd}(\bar{Y})$. La variable aleatoria asociada, $S/\sqrt{n}$, algunas veces recibe el nombre de **error**

FIGURA C.4

El 97.5-ésimo percentil, *c*, en una distribución *t*.



estándar de $\bar{Y}$. Debido a la estimación puntual $s/\sqrt{n}$, que se presenta en las fórmulas, se define el error estándar de $\bar{y}$ como $\text{ee}(\bar{y}) = s/\sqrt{n}$. Entonces, (C.24) se puede escribir de forma abreviada como

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{ee}(\bar{y})]. \tag{C.25}$$

Esta ecuación muestra por qué la noción de error estándar de una estimación desempeña una importante función en econometría.

Ejemplo C.2

[Efecto de los subsidios para la capacitación laboral en la productividad de los trabajadores]

Holzer, Block, Cheatham y Knott (1993) estudiaron los efectos de los subsidios para la capacitación laboral en la productividad de los trabajadores al recabar información sobre "tasa de desperdicio" para una muestra de empresas manufactureras de Michigan que recibieron subsidios para la capacitación laboral en 1988. La tabla C.3 lista las tasas de desperdicio, medidas como número de artículos no utilizables y, por tanto, que tenían que desecharse por cada 100 producidos, entre 20 empresas. Cada una de estas empresas recibió un subsidio para la capacitación laboral en 1988; en 1987 no existían los subsidios. Se desea construir un intervalo de confianza para el cambio en la tasa de desperdicio de 1987 a 1988 para la población de todas las empresas manufactureras que pudieran haber recibido subsidios.

Se supone que el cambio en la tasa de desperdicio tiene una distribución normal. Dado que $n = 20$, un intervalo de confianza a 95% para el cambio en la media de las tasas de desperdicio μ es $[\bar{y} \pm 2.093 \cdot \text{ee}(\bar{y})]$, donde $\text{ee}(\bar{y}) = s/\sqrt{n}$. El valor de 2.093 es el 97.5-ésimo percentil en una distribución t_{19}. Para los valores de la muestra en particular, $\bar{y} = -1.15$ y $\text{ee}(\bar{y}) = .54$ (cada uno de ellos redondeado a dos decimales), así que el intervalo de confianza a 95% es $[-2.28, -.02]$. El valor cero se excluye de este intervalo, así que se concluye que, con una confianza a 95%, el cambio promedio en las tasas de desperdicio de la población no es cero.

TABLA C.3

Tasas de desperdicio para 20 empresas manufactureras de Michigan

Empresa	1987	1988	Cambio
1	10	3	–7
2	1	1	0
3	6	5	–1
4	.45	.5	.05
5	1.25	1.54	.29
6	1.3	1.5	.2
7	1.06	.8	–.26
8	3	2	–1
9	8.18	.67	–7.51
10	1.67	1.17	–.5
11	.98	.51	–.47
12	1	.5	–.5
13	.45	.61	.16
14	5.03	6.7	1.67
15	8	4	–4
16	9	7	–2
17	18	19	1
18	.28	.2	–.08
19	7	5	–2
20	3.97	3.83	–.14
Promedio	4.38	3.23	–1.15

En este punto, el ejemplo C.2 es más ilustrativo pues tiene, como un análisis económico, fallas potencialmente más serias. Pero lo más importante es que supone que cualquier reducción sistemática en las tasas de desperdicio se debe a los subsidios en la capacitación laboral. Pero, desde luego, en el transcurso de un año, muchas cosas pueden suceder que modifiquen la productividad del trabajador. A partir de este análisis, no hay forma de saber si la disminución en las tasas de desperdicio se puede atribuir a los subsidios para la capacitación laboral, o si, al menos en parte, es responsable alguna fuerza externa.

Una sencilla regla general para un intervalo de confianza a 95%

El intervalo de confianza en (C.25) se puede calcular para cualquier tamaño muestral y cualquier nivel de confianza. Como se vio en la sección B.5, la distribución t se aproxima a la distribución normal estándar a medida que aumentan los grados de libertad. En particular, para $\alpha = .05$, $c_{\alpha/2} \rightarrow 1.96$ conforme $n \rightarrow \infty$, aunque $c_{\alpha/2}$ siempre es mayor que 1.96 para cada n. Una *regla general* para un nivel de confianza aproximado a 95% es

$$[\bar{y} \pm 2 \cdot \mathrm{ee}(\bar{y})]. \tag{C.26}$$

En otras palabras, se obtiene $\bar{y}$ y su error estándar, y después se calcula $\bar{y}$ menos el doble de su error estándar para obtener el intervalo de confianza. Esto es demasiado amplio para una n, muy grande, y es demasiado reducido para una n pequeña. Como puede verse en el ejemplo C.2, incluso para una n tan pequeña como 20, (C.26) está en el ámbito del intervalo de confianza para la media de una distribución normal. Esto significa que es posible obtener algo muy aproximado a 95% del intervalo de confianza sin tener que referirse a las tablas t.

Intervalos de confianza asintóticos para poblaciones no normales

En algunas aplicaciones, la población es claramente no normal. Un caso sobresaliente es la distribución de Bernoulli, donde la variable aleatoria asume sólo valores de cero y uno. En otros casos, la población no normal no tiene distribución estándar. Esto no importa, siempre y cuando el tamaño de la muestra sea lo bastante grande para el teorema del límite central para dar una buena aproximación de la distribución de la media muestral $\bar{Y}$. Para una n, grande, un intervalo de confianza *aproximado* a 95% es

$$[\bar{y} \pm 1.96 \cdot \mathrm{ee}(\bar{y})], \tag{C.27}$$

donde el valor 1.96 es el 97.5-ésimo percentil en la distribución normal estándar. En términos mecánicos, calcular un intervalo de confianza aproximado no difiere del caso normal. Una pequeña diferencia consiste en que el número que multiplica el error estándar proviene de la distribución normal estándar, y no de la distribución t debido a que se están usando asintóticas. Debido a que la distribución t se aproxima a la normal estándar a medida que los *gl* aumentan, la ecuación (C.25) también se legitima perfectamente como una aproximación al intervalo de 95%; algunos prefieren ésta a (C.27), pues la primera es exacta para poblaciones normales.

Ejemplo C.3

[Discriminación racial en las contrataciones]

El Urban Institute realizó un estudio en 1988 en Washington, D.C. para examinar el grado de discriminación racial en las contrataciones. Se entrevistó a cinco pares de personas para varios puestos. En cada par, una persona era blanca y la otra negra, las cuales ofrecieron currículos que indicaba que eran virtualmente iguales en términos de experiencia, educación y otros factores que determinaba la calificación para el trabajo. La idea era hacer que los individuos fueran tan similares como fuera posible con excepción de la raza. Las personas de cada par pasaban por una entrevista para solicitar el mismo trabajo y los investigadores anotaron qué solicitante recibió cada oferta de trabajo. Este es un ejemplo de un *análisis de pares* *igualados,* donde cada prueba consiste en datos de dos personas (o dos empresas, ciudades, etcétera) que se consideran similares en muchos aspectos pero diferentes en una característica importante.

Sea θ_B la probabilidad de que la persona negra obtenga el trabajo y θ_W la probabilidad de que la persona blanca lo obtenga. Lo que aquí interesa es la diferencia, $\theta_B - \theta_W$. Sea B_j una variable de Bernoulli igual a uno si la persona negra obtiene una oferta de trabajo de su empleador i, y cero de otra manera. Asimismo, $W_i = 1$ si la persona blanca obtiene una oferta de trabajo del empleador i, y cero en caso contrario. Al reunir los resultados de los cinco pares de personas, hubo un total de $n = 241$ pruebas (pares de entrevistas con empleadores). Los estimadores insesgados de θ_B y θ_W son $\bar{B}$ y $\bar{W}$, las fracciones de entre-vistas en las cuales personas blancas y negras recibieron ofertas de trabajo, respectivamente.

Para calcular un intervalo de confianza para la media de una población, defina una nueva variable $Y_i = B_i - W_i$. Ahora, Y_i puede tomar tres valores: -1 si la persona negra no obtiene el trabajo, pero la persona blanca sí, 0 si ninguna de las personas obtiene el trabajo o si las dos lo obtienen, y 1 si la persona negra obtiene el trabajo y la persona blanca no. Entonces, $\mu \equiv \mathrm{E}(Y_i) = \mathrm{E}(B_i) - \mathrm{E}(W_i) = \theta_B - \theta_W$.

La distribución de Y_i sin lugar a dudas es no normal; es discreta y sólo toma tres valores. Sin embargo, un intervalo de confianza aproximado para $\theta_B - \theta_W$ se puede obtener mediante métodos muestrales grandes.

Mediante los 241 puntos de datos observados $\bar{b} = .224$ y $\bar{w} = .357$, así que $\bar{y} = .224 - .357 = -.133$. Por tanto, 22.4% de los solicitantes negros recibieron oferta de trabajo, mientras que 35.7% de los solicitantes blancos recibieron oferta de trabajo. Esto es a *primera vista* una evidencia de la discriminación en contra de las personas negras, pero es posible saber mucho más al calcular un intervalo de confianza para μ. Para calcular un intervalo de confianza a 95%, se necesita la desviación estándar muestral. Ésta resulta ser $s = .482$ [mediante la ecuación (C.21)]. Mediante (C.27) se obtiene un IC a 95% para $\mu = \theta_B - \theta_W$ cuando $-.133 \pm 1.96(.482/\sqrt{241}) = -.133 \pm .031 = [-.164, -.102]$. El IC aproximado a 99% es $-.133 \pm 2.58(.482/\sqrt{241}) = [-.213, -.053]$. Naturalmente, esto contiene a un rango más amplio de valores que el IC a 95%. Pero incluso el IC a 99% no contiene al valor cero. Por tanto, se tiene un alto grado de confianza en que la diferencia poblacional $\theta_B - \theta_W$ no sea cero.

Antes de regresar a la prueba de hipótesis, es útil revisar las diferentes cantidades muestrales y poblacionales que miden las diferencias entre las distribuciones poblacionales y las distribuciones de muestreo de los estimadores. Estas cantidades suelen aparecer en los análisis estadísticos y sus extensiones son importantes para el análisis de regresión en el texto principal. La cantidad σ es la desviación estándar de la población (desconocida); es la medida de la diferencia en la distribución de Y. Cuando se divide σ entre $\sqrt{n}$, se obtiene la **desviación estándar muestral** de $\bar{Y}$ (el promedio muestral). Si bien σ es una característica fija de la población, $\mathrm{sd}(\bar{Y}) = \sigma/\sqrt{n}$ se reduce a cero conforme $n \to \infty$: el estimador de μ se vuelve cada vez más preciso a medida que el tamaño de la muestra aumenta.

La estimación de σ para una muestra en particular, s, recibe el nombre de desviación estándar muestral debido a que se obtiene de la muestra. (También recibe el nombre de variable aleatoria subyacente, S, que cambia a través de muestras diferentes, la desviación estándar muestral.) Al igual que $\bar{y}$ como estimación de μ, s es el “mejor supuesto” para σ dada la muestra en cuestión. La cantidad $s/\sqrt{n}$ es lo que se llama error estándar de $\bar{y}$, y es la mejor estimación de $\sigma/\sqrt{n}$. Los intervalos de confianza para el parámetro poblacional μ dependen directamente de $\mathrm{ee}(\bar{y}) = s/\sqrt{n}$. Debido a que este error estándar se reduce a cero a medida que el tamaño de la muestra aumenta, un tamaño de muestra más grande, por lo general, significa un intervalo de confianza menor. Por tanto, se puede ver claramente que un beneficio de tener más datos es que generan intervalos de confianza más estrechos. La noción de error estándar de una estimación, que en la vasta mayoría de los casos se reduce a cero a una tasa de $1/\sqrt{n}$, desempeña una función fundamental en la prueba de hipótesis (como se verá en la siguiente sección) y para intervalos de confianza y prueba en el contexto de regresión múltiple (como se analizó en el capítulo 4).

C.6 Prueba de hipótesis

Hasta ahora se ha analizado cómo evaluar los estimadores puntuales, y se ha visto, en el caso de una media poblacional, cómo construir e interpretar intervalos de confianza. Pero en ocasiones la cuestión que interesa es tener una respuesta definitiva de sí o no. He aquí algunos ejemplos: 1) ¿Un programa de capacitación laboral aumenta efectivamente la productividad promedio del trabajador? (vea el ejemplo C.2); 2) ¿Las personas negras sufren discriminación en el proceso de contratación? (vea el ejemplo C.3); 3) ¿Las estrictas leyes estatales contra la conducción en estado de ebriedad reducen el número de arrestos por conducir en esas condiciones? Diseñar métodos para contestar tales preguntas, mediante una muestra de datos, recibe el nombre de prueba de hipótesis.

Fundamentos de la prueba de hipótesis

Para ilustrar las cuestiones que implica la prueba de hipótesis, considere un ejemplo electoral. Suponga que se postulan dos candidatos en una elección, el candidato A y el B. El candidato A reportó haber recibido 42% del voto popular, mientras que el candidato B recibió 58%. Se supone que estos son porcentajes verdaderos entre la población votante y se tratarán como tales.

El candidato A está convencido de que más personas votaron por él, así que le gustaría investigar si la elección fue fraudulenta. Con cierto conocimiento estadístico, el candidato A contrata a una agencia consultora para hacer una muestra al azar de 100 votantes para saber si cada persona votó o no por él. Suponga que, en la muestra recabada, 53 personas votaron por el candidato A. Esta estimación muestral de 53% excede claramente el valor reportado de la población de 42%. ¿El candidato A debe concluir que en realidad la elección fue fraudulenta?

Si bien todo parece indicar que hubo una omisión de votos en el conteo para el candidato A, es posible que, en una muestra de 100, se observen 53 personas que en realidad votaron por el candidato A. La pregunta es: ¿Qué tan *sólida* es la evidencia muestral frente al porcentaje oficialmente reportado de 42%?

Una forma de proceder es establecer una **prueba de hipótesis**. Sea θ la verdadera proporción de la población que votó por el candidato A. La hipótesis de que los resultados reportados son precisos se puede expresar como

$$H_0: \theta = .42. \tag{C.28}$$

Éste es un ejemplo de **hipótesis nula**. Siempre se denotará aquí la hipótesis nula como H_0. En la prueba de hipótesis, la hipótesis nula desempeña una función similar a la de un defensor de oficio en muchos sistemas judiciales; tal como el acusado se presume inocente hasta demostrarse culpable, la hipótesis nula se presume verdadera hasta que los datos sugieran de manera contundente lo contrario. En este ejemplo, el candidato A debe presentar evidencia concluyente en contra de (C.28) con el fin de ganar el derecho a un recuento.

La **hipótesis alternativa** en el ejemplo de la elección es que la verdadera proporción que votó por el candidato A en la elección es mayor que .42:

$$H_1: \theta > .42. \tag{C.29}$$

Con el fin de concluir que H_0 es falsa y H_1 es verdadera, se debe tener la evidencia "más allá de la duda razonable" en contra de H_0. ¿Cuántos votos de 100 serían necesarios antes de parecer que la evidencia es contundente en contra de H_0? La mayoría estaría de acuerdo en que 43 votos de una muestra de 100 no son suficientes para anular los resultados de la elección original; tal resultado está dentro de la variación muestral esperada. Por otra parte, no es necesario observar 100 votos para el candidato A para que se dude de H_0. Si 53 de 100 basta para rechazar H_0 es mucho menos claro. La respuesta depende de cómo se cuantifique la expresión "más allá de la duda razonable".

Antes de regresar a la cuestión de cuantificar la incertidumbre en la prueba de hipótesis, es necesario prevenir cualquier posible confusión. Quizá se haya observado que la hipótesis en las ecuaciones (C.28) y (C.29) no agota todas las posibilidades: podría ser que θ sea menor que .42. Para la aplicación en cuestión, esa posibilidad no es de particular interés; no tiene nada que ver con anular los resultados de la elección. Por tanto, sólo se puede afirmar al principio que se están ignorando las alternativas θ con $\theta < .42$. Sin embargo, algunos autores prefieren establecer hipótesis nulas y alternativas, con fines de exhaustividad, en cuyo caso la hipótesis nula aquí sería H_0: $\theta \leq .42$. Expresado de esta forma, la hipótesis nula es una hipótesis nula *compuesta*, debido a que permite más de un valor bajo H_0. [Por el contrario, la ecuación (C.28) es un ejemplo de una hipótesis nula *simple*]. Para este tipo de ejemplos, no importa si se establece la hipótesis nula como en (C.28) o como una nula compuesta: el valor más difícil de rechazar si $\theta \leq .42$ es $\theta = .42$. (Es decir, si se rechaza el valor $\theta = .42$ frente a $\theta > .42$, entonces lógicamente se deberá rechazar cualquier valor menor que .42.) Por tanto, este procedimiento de prueba basado en (C.28) lleva a la misma prueba como si H_0: $\theta \leq .42$. En este libro siempre se establecerá una hipótesis nula como una hipótesis nula simple.

En la prueba de hipótesis se pueden cometer dos tipos de errores. Primero, se puede rechazar la hipótesis nula cuando de hecho es verdadera. Esto recibe el nombre de **error tipo I**. En el ejemplo de la elección ocurre un error tipo I si se rechaza H_0 cuando la verdadera proporción de personas que votan por el candidato A es en realidad .42. El segundo tipo de error es no rechazar H_0 cuando en realidad es falsa. Esto recibe el nombre de **error tipo II**. En el ejemplo de la elección, ocurre un error tipo II si $\theta > .42$ pero no se rechaza H_0.

Después de que haber tomado la decisión de rechazar o no la hipótesis nula, se habrá decidido correctamente o se habrá cometido un error. Nunca se sabrá con certeza si se cometió un error. No obstante, se puede calcular la *probabilidad* de cometer un error tipo I o tipo II. Las reglas de prueba de hipótesis se construyen para hacer que la probabilidad de cometer un error tipo I sea muy pequeña. En general, se define el **nivel de significancia** (o simplemente el *nivel*) de una prueba como la probabilidad de un error tipo I; suele denotarse por α. En símbolos, se tiene

$$\alpha = \mathrm{P}(\text{Reject } \mathrm{H}_0|\mathrm{H}_0). \qquad \textbf{C.30}$$

El lado derecho se lee como: "La probabilidad de rechazar H_0 dado que H_0 es verdadera."

La prueba clásica de la hipótesis requiere que inicialmente se especifique un nivel de significancia para una prueba. Cuando se especifica un valor para α, esencialmente se está cuantificando la tolerancia para un error tipo I. Los valores comunes de α son .10, .05 y .01. Si $\alpha = .05$, entonces el investigador desea rechazar falsamente H_0 5% de las veces, con el fin de detectar desviaciones de H_0.

Una vez que se ha elegido el nivel de significancia, entonces quizá se desee minimizar la probabilidad del error tipo II. Otra posibilidad es que se quiera maximizar la **potencia de una prueba** contra cualquier alternativa relevante. La potencia de una prueba es sólo uno menos la probabilidad del error tipo II. Matemáticamente,

$$\pi(\theta) = \mathrm{P}(\text{Reject } \mathrm{H}_0|\theta) = 1 - \mathrm{P}(\text{Type II}|\theta),$$

donde θ denota el valor real del parámetro. Naturalmente, se desearía poder igualar a uno si la hipótesis nula es falsa. Pero eso es imposible de lograr mientras se mantiene pequeño el nivel de significancia. En cambio, se opta porque las pruebas maximicen la potencia para un nivel de significancia dado.

Pruebas de hipótesis para la media de una población normal

Con el fin de probar una hipótesis nula contra una alternativa, es necesario elegir un estadístico de prueba (o estadístico, para abreviar) y un valor crítico. La elección del estadístico y el valor crítico está basada en la conveniencia y el deseo de maximizar la potencia dado un nivel de significancia para la prueba. En esta subsección se revisará cómo comprobar las hipótesis para la media de una población normal.

Un **estadístico de prueba**, denotado como T, es alguna función de la muestra aleatoria. Cuando se calcula el estadístico para un resultado en particular, se obtiene un resultado para el estadístico de prueba, que se denotará por t.

Dado un estadístico de prueba, se puede definir una regla de rechazo que determine cuando H_0 se rechace en favor de H_1. En esta prueba todas las reglas de rechazo están basadas en comparar el valor de un estadístico de prueba, t, respecto a un **valor crítico**, c. Los valores de t que resulten del rechazo de la hipótesis nula se conocen de manera conjunta como la **región de rechazo**. Para determinar el valor crítico, se debe decidir primero el nivel de significancia de la prueba. Después, dada α, el valor crítico asociado con α se determina mediante la distribución de T, *suponiendo* que H_0 sea verdadera. Se escribirá este valor crítico como c y se eliminará el hecho de que depende de α.

Probar la hipótesis acerca de la media μ de una población Normal(μ,σ^2) es sencillo. La hipótesis nula se expresa como

$$H_0\text{: } \mu = \mu_0, \tag{C.31}$$

donde μ_0 es un valor que se especifica. En la mayoría de las aplicaciones, $\mu_0 = 0$, pero el caso general no es más difícil.

La regla de rechazo que se eligió depende de la naturaleza de la hipótesis alternativa. Las tres alternativas de interés son

$$H_1\text{: } \mu > \mu_0, \tag{C.32}$$

$$H_1\text{: } \mu < \mu_0, \tag{C.33}$$

y

$$H_1\text{: } \mu \neq \mu_0. \tag{C.34}$$

La ecuación (C.32) da una **alternativa de una cola**, como (C.33). Cuando la hipótesis alternativa es (C.32), la nula es efectivamente H_0: $\mu \leq \mu_0$, dado que se rechaza H_0 sólo cuando $\mu > \mu_0$. Esto es apropiado cuando se tiene interés en el valor de μ sólo cuando μ es al menos tan grande como μ_0. La ecuación (C.34) es una **alternativa de dos colas**. Esta es apropiada cuando se tiene interés en cualquier desviación a partir de la hipótesis nula.

Considere primero la alternativa en (C.32). Intuitivamente, se debe rechazar H_0 en favor de H_1 cuando el valor del promedio muestral, $\bar{y}$, es lo "bastante" mayor que μ_0. Pero, ¿cómo se debe determinar cuando $\bar{y}$ es lo bastante mayor para que H_0 se rechace al nivel de significancia elegido? Esto requiere conocer la probabilidad de rechazar la hipótesis nula, cuando ésta es verdadera. En lugar de trabajar directamente con $\bar{y}$, se utiliza su versión estandarizada, donde σ se remplaza con la desviación estándar muestral, s:

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{ee}(\bar{y}), \tag{C.35}$$

donde $\text{ee}(\bar{y}) = s/\sqrt{n}$ es el error estándar de $\bar{y}$. Dada la muestra de datos, es fácil obtener t. Se trabaja con t debido a que, bajo la hipótesis nula, la variable aleatoria

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

tiene una distribución t_{n-1}. Ahora, suponga que se ha establecido un nivel de significancia a 5%. Entonces se elije el valor crítico c, de manera que $P(T > c|H_0) = .05$; es decir, la probabilidad de un error tipo I es de 5%. Una vez que se ha encontrado c, la regla de rechazo es

$$t > c, \qquad \textbf{C.36}$$

donde c es el percentil $100(1 - \alpha)$ en una distribución t_{n-1}; como porcentaje, el nivel de significancia es $100 \cdot \alpha\%$. Este es un ejemplo de una **prueba de una cola** debido a que la región de rechazo está en una cola de la distribución t. Para un nivel de significancia de 5%, c es el percentil 95-énesimo en la distribución t_{n-1}; esto se ilustra en la figura C.5. Un nivel de significancia diferente ocasiona un valor crítico diferente.

El estadístico en la ecuación (C.35) suele recibir el nombre de **estadístico *t*** para prueba de H_0: $\mu = \mu_0$. El estadístico t mide la distancia de $\bar{y}$ a μ_0 *con relación* al error estándar de $\bar{y}$, $\text{ee}(\bar{y})$.

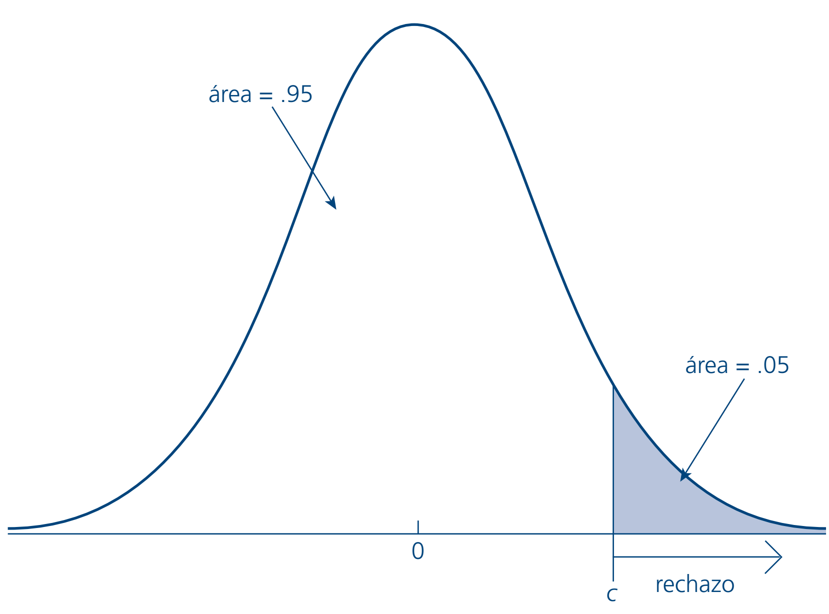
Ejemplo C.4

[Efecto de las zonas empresariales en las inversiones de negocios]

En la población de ciudades que otorgan zonas empresariales en un estado en particular [vea Papke (1994) para Indiana], sea Y el cambio porcentual en la inversión a partir del año anterior al año en que una ciudad se

FIGURA C.5

Región de rechazo para una prueba con un nivel de significancia de 5% frente a la alternativa de una cola $\mu > \mu_0$.



convirtió en una zona empresarial. Suponga que Y tiene una distribución Normal(μ,σ^2) . La hipótesis nula de que las zonas empresariales no tienen efecto en la inversión de negocios es H_0: $\mu = 0$; la alternativa de que tiene un efecto positivo es H_1: $\mu > 0$. (Se supone que no tienen un efecto negativo). Suponga que se desea probar H_0 al nivel de 5%. El estadístico de prueba en este caso es

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{\text{ee}(\bar{y})}. \tag{C.37}$$

Suponga que se tiene una muestra de 36 ciudades a las que les han otorgado zonas empresariales. Entonces, el valor crítico es $c = 1.69$ (vea la tabla G.2) y se rechaza H_0 en favor de H_1 si $t > 1.69$. Suponga que la muestra produce $\bar{y} = 8.2$ y $s = 23.9$. Entonces, $t \approx 2.06$, y por tanto H_0 se rechaza al nivel de 5%. Por consiguiente, se concluye que, al nivel de significancia de 5%, las zonas empresariales tienen un efecto en la inversión promedio. El valor crítico de 1% es 2.44, así que H_0 no se rechaza al nivel de 1%. La misma advertencia se mantiene aquí como en el ejemplo C.2: no se controlan otros factores que pudieran afectar la inversión en las ciudades a través del tiempo, así que no se puede afirmar que el efecto sea causal.

La regla de rechazo es similar a la alternativa de una cola (C.33). Una prueba con un nivel de significancia de $100\cdot\alpha\%$ rechaza H_0 frente a (C.33) siempre que

$$t < -c; \tag{C.38}$$

en otras palabras, se está buscando valores negativos del estadístico t lo cual implica $\bar{y} < \mu_0$ que están lo bastante lejos de cero para rechazar H_0.

Para alternativas de dos colas, se debe tener cuidado en elegir el valor crítico de manera que el nivel de significancia de la prueba siga siendo α. Si H_1 está dada por H_1: $\mu \neq \mu_0$, entonces se rechaza H_0 si $\bar{y}$ está lejos de μ_0 en *valor absoluto*: una $\bar{y}$ mucho mayor o mucho menor que μ_0 ofrece evidencia contra H_0 en favor de H_1. Una prueba al nivel $100\cdot\alpha\%$ se obtiene de la regla de rechazo

$$|t| > c, \tag{C.39}$$

donde $|t|$ es el valor absoluto del estadístico t en (C.35). Esto da una **prueba de dos colas**. Se debe tener cuidado en elegir el valor crítico: c es el percentil $100(1 - \alpha/2)$ en la distribución t_{n-1}. Por ejemplo, si $\alpha = .05$, entonces el valor crítico es el percentil 97.5-ésimo en la distribución t_{n-1}. Esto asegura que H_0 se rechace sólo 5% de las veces cuando es verdadera (vea la figura C.6). Por ejemplo, si $n = 22$, entonces el valor crítico es $c = 2.08$, el 97.5-ésimo percentil en una distribución t_{21} vea la tabla G.2). El valor absoluto del estadístico t debe exceder a 2.08 con el fin de rechazar H_0 frente a H_1 al nivel 5%.

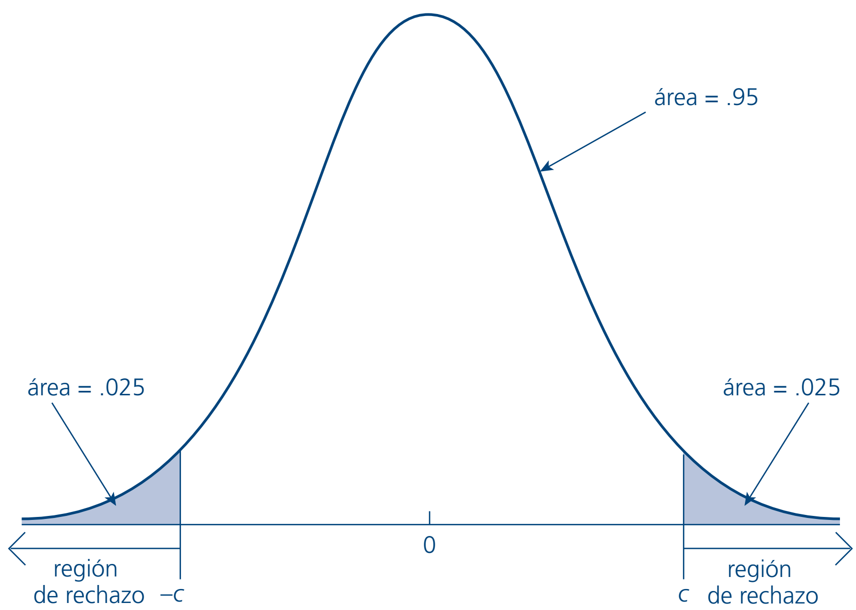
Es importante conocer el lenguaje apropiado en la prueba de hipótesis. En ocasiones la frase apropiada “no se rechazó H_0 en favor de H_1 al nivel de significancia de 5%” se remplaza con “se acepta H_0 al nivel de significancia de 5%.” Esta última redacción es incorrecta. Con el mismo conjunto de datos, existen por lo general, numerosas hipótesis que no se pueden rechazar. En el ejemplo de las elecciones, sería ilógico decir que se “aceptan” H_0: $\theta = .42$ y H_0: $\theta = .43$ dado que sólo una es verdadera. Pero es totalmente posible que ninguna de las hipótesis pueda rechazarse. Por esta razón, siempre se dice que “no se rechazó H_0” en lugar de “se aceptó H_0.”

Pruebas asintóticas para poblaciones no normales

Si el tamaño muestral es lo bastante grande para invocar el teorema del límite central (vea la sección C.3) la mecánica de la prueba de hipótesis para las medias poblacionales es la *misma*,

FIGURA C.6

Región de rechazo para una prueba al nivel de significancia de 5% frente a la alternativa de dos colas H_1: $\mu \neq \mu_0$.



sin importar si la distribución poblacional es normal o no. La justificación teórica proviene del hecho que, bajo la hipótesis nula

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Por tanto, con n, grande, se puede comparar el estadístico t en (C.35) con los valores críticos de una distribución normal estándar. Debido a que la distribución t_{n-1} converge a la distribución normal estándar a medida que n aumenta, t y los valores críticos normales estándar estarán muy cerca para una n extremadamente grande. Debido a que la teoría asintótica está basada en una n que crece de manera ilimitada, no puede indicar si son mejores los valores críticos normales estándares o los t. Para los valores moderados de n, por ejemplo, entre 30 y 60, es tradicional emplear la distribución t debido a que se sabe que es lo correcto para las poblaciones normales. Para $n > 120$, la elección entre t y las distribuciones normales estándar suele ser irrelevante debido a que los valores críticos prácticamente son los mismos.

Debido a que los valores críticos elegidos utilizan la distribución t o la normal estándar son sólo aproximadamente válidos para poblaciones no normales, los niveles de significancia elegidos son sólo aproximados; por tanto, para poblaciones no normales, los niveles de significancia son realmente niveles de significancia asintóticos. De este modelo, si se elige un nivel de significancia de 5%, pero la población es no normal, entonces el nivel de significancia real será mayor o menor que 5% (y no se puede saber cuál es el caso). Cuando el tamaño muestral es grande, el nivel de significancia real será muy cercano a 5%. En términos prácticos, la distinción no es importante, por tanto, se elimina el calificativo "asintótico".

Ejemplo C.5

[Discriminación racial en las contrataciones]

En el estudio del Urban Institute sobre discriminación en la contratación (vea el ejemplo C.3), el interés principal estribaba en probar H_0: $\mu = 0$ frente a H_1: $\mu < 0$, donde $\mu = \theta_B - \theta_W$ es la diferencia de probabilidades de que las personas negras recibieran ofertas de trabajo con relación a los blancos. Recuerde que μ es la media poblacional de la variable $Y = B - W$, donde B y W son los indicadores binarios. Mediante las $n = 241$ comparaciones pareadas, se obtuvo $\bar{y} = -.133$ y $\text{ee}(\bar{y}) = .482/\sqrt{241} \approx .031$. El estadístico t para probar H_0: $\mu = 0$ es $t = -.133/.031 \approx -4.29$. Se recordará del apéndice B que la distribución normal estándar es, para fines prácticos, indistinguible de la distribución t con 240 grados de libertad. El valor de -4.29 está tan lejano en el extremo izquierdo de la distribución que se rechaza H_0 a cualquier nivel de significancia razonable. De hecho, el valor crítico de .005 (medio punto porcentual para una prueba de una cola) es de aproximadamente -2.58. Un valor t de -4.29 es una evidencia *muy* sólida contra H_0 y en favor de H_1. Por tanto, se concluye que existe discriminación en la contratación.

Cálculo y uso de los valores-*p*

El requerimiento tradicional de usar un nivel de significancia por adelantado significa que investigadores diferentes, con ayuda de los mismos datos y el mismo procedimiento para comprobar la misma hipótesis, pueden llegar a conclusiones diferentes. Reportar el nivel de significancia al cual se lleva a cabo la prueba resuelve hasta cierto punto este problema, pero no lo elimina por completo.

Para ofrecer más información, se puede plantear la siguiente pregunta: ¿cuál es el nivel de significancia *máximo* al cual se puede realizar la prueba y aun así no rechazar la hipótesis nula? El valor se conoce como el **valor-*p*** de una prueba (en ocasiones llamado el valor de la prob). En comparación con la elección de un nivel de significancia por adelantado y obtener un valor crítico, comparar un valor-p es un tanto más difícil. Pero con la ayuda de los cálculos computacionales rápidos y económicos, los valores-p ahora son muy fáciles de obtener.

Como ejemplo, considere el problema de probar H_0: $\mu = 0$ en una población Normal(μ,σ^2) El estadístico t en este caso es $T = \sqrt{n}\cdot\bar{Y}/S$, y se supone que n es suficientemente grande para tratar T como si contara con una distribución normal estándar según H_0. Suponga que el valor observado de T para la muestra es $t = 1.52$. (Observe cómo se ha saltado el paso de elegir un nivel de significancia). Ahora que se ha visto el valor t, se encuentra el nivel de significancia más grande al cual no se lograría rechazar H_0. Este es el nivel de significancia asociado con usar t como valor crítico. Debido a que la prueba del estadístico T tiene una distribución normal estándar según H_0, se tiene

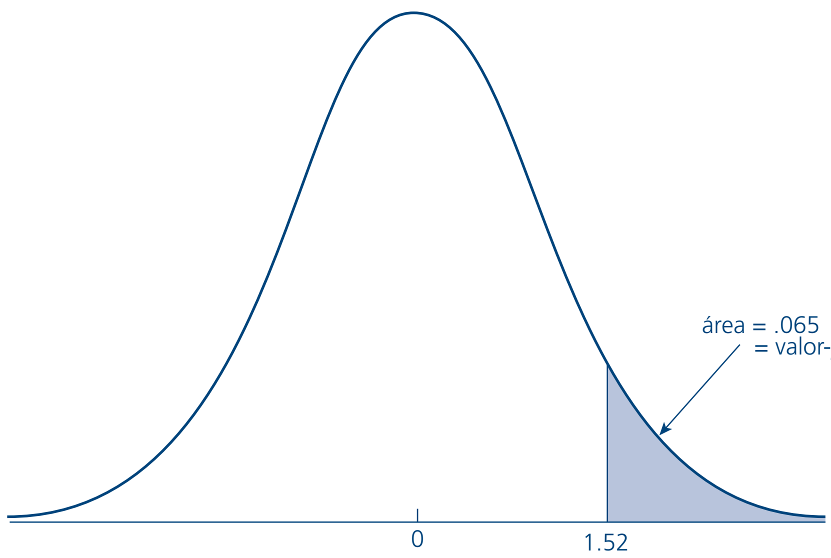
$$\text{valor-}p = P(T > 1.52|H_0) = 1 - \Phi(1.52) = .065, \tag{C.40}$$

donde $\Phi(\cdot)$ denota la fda normal estándar. En otras palabras el valor-p en este ejemplo es simplemente el área a la derecha de 1.52, el valor observado del estadístico de prueba, en una distribución normal estándar. Vea la figura C.7 como ejemplo.

Debido a que el valor-p = .065, el nivel de significancia mayor al cual se puede realizar esta prueba y no rechazar es 6.5%. Si se realiza la prueba al nivel por debajo de 6.5% (como a 5%), no se rechaza H_0. Si se realiza la prueba a un nivel mayor que 6.5% (como a 10%), se rechaza H_0. Con el valor-p a la mano, se puede realizar la prueba a cualquier nivel.

FIGURA C.7

El valor-p cuando $t = 1.52$ para una alternativa de una cola $\mu > \mu_0$.



El valor-p en este ejemplo tiene otra interpretación útil: es la probabilidad que observe un valor de T tan grande como 1.52 cuando la hipótesis nula es verdadera. Si la hipótesis nula en realidad es verdadera, se observaría un valor de T tan grande como 1.52 debido a la probabilidad de sólo 6.5% de las veces. Si esto es lo bastante pequeño para rechazar H_0 depende de la tolerancia que se tenga para un error tipo I. El valor-p tiene una interpretación similar en todos los demás casos, como se verá.

En general, los valores-p pequeños son evidencia *contra* H_0, dado que indican que el resultado de los datos ocurre con una probabilidad pequeña si H_0 es verdadera. En el ejemplo anterior, si t hubiera sido un valor mayor, por ejemplo, $t = 2.85$, entonces el valor-p habría sido $1 - \Phi(2.85) \approx .002$. Esto significa que, si la hipótesis nula fuera verdadera, se observaría un valor de T hasta de 2.85 con una probabilidad de .002. ¿Cómo interpretar esto? Ya sea que se obtenga una muestra muy inusual o que la hipótesis nula sea falsa. A menos que se tenga una tolerancia *muy* pequeña para un error de tipo I, se rechazaría la hipótesis nula. Por otra parte, un valor-p grande es una evidencia débil contra H_0. Si se obtuvo $t = .47$ en el ejemplo anterior, entonces el valor-$p = 1 - \Phi(.47) = .32$. Observar a un valor de T mayor que .47 sucede con probabilidad de .32 incluso cuando H_0 es verdadera; esto es lo bastante grande de manera que la duda acerca de H_0 es insuficiente, a menos que se tenga una tolerancia muy alta para un error tipo I.

Para probar hipótesis acerca de la media de una población mediante la distribución t, se necesitan tablas detalladas para calcular los valores-p. La tabla G.2 sólo permite poner límites a los valores-p. Por fortuna, numerosos paquetes estadísticos y econométricos ahora calculan los valores-p de manera rutinaria, y también ofrecen cálculos de las fda para t y otras distribuciones usadas para calcular valores-p.

Ejemplo C.6

[Efecto de los subsidios para la capacitación laboral en la productividad de los trabajadores]

Considere nuevamente los datos de Holzer *et al*. (1993) en el ejemplo C.2. Desde una perspectiva de políticas, existen dos cuestiones de interés. Primero, ¿cuál es la mejor estimación del cambio promedio en las tasas de desperdicios industriales, μ? Ya se obtuvo esto de la muestra de 20 empresas listadas en la tabla C.3: el promedio muestral en el cambio en las tasas de desperdicio es de -1.15. Con relación a la tasa inicial de desperdicio en 1987, esto representa una disminución en la tasa de desperdicio de aproximadamente 26.3% ($-1.15/4.38 \approx -.263$), lo cual es un efecto importante.

También se desea saber si la muestra ofrece una evidencia contundente para un efecto en la población de empresas manufactureras que pudieran haber recibido los subsidios. La hipótesis nula es H_0: $\mu = 0$, y se prueba esto contra H_1: $\mu < 0$, donde μ es el cambio promedio en las tasas de desperdicio. Según la nula, los subsidios de capacitación laboral no tienen efecto en las tasas promedio de desperdicio. La alternativa establece que existe un efecto. La alternativa $\mu > 0$ no es de interés; la hipótesis nula es efectivamente H_0: $\mu \geq 0$.

Dado que $\bar{y} = -1.15$ y ee$(\bar{y}) = .54$, $t = -1.15/.54 = -2.13$. Esto se halla por debajo del valor crítico de 5% de -1.73 (de una distribución t_{19}) pero por encima del valor crítico de 1%, -2.54. El valor-p en este caso se calcula como

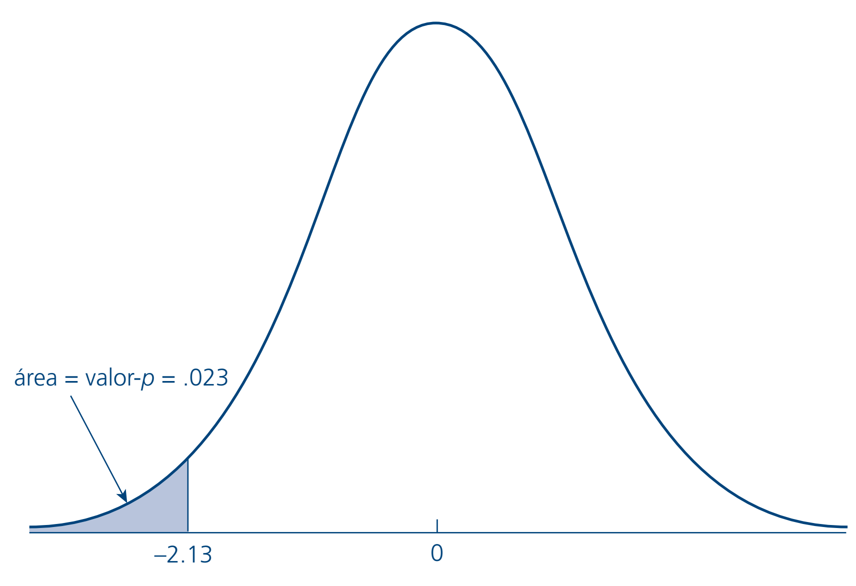
$$\text{valor-}p = \text{P}(T_{19} < -2.13), \tag{C.41}$$

donde T_{19} representa una variable aleatoria que se distribuye como una t con 19 grados de libertad. La desigualdad se invierte de (C.40) debido a que la alternativa tiene la forma en (C.33). La probabilidad en (C.41) es el área a la izquierda de -2.13 en una distribución t_{19} (vea la figura C.8).

Mediante la tabla G.2, lo más que se puede decir es que el valor-p está entre .025 y .01, pero esto es más cercano a .025 (dado que el 97.5-ésimo percentil es de aproximadamente 2.09). Mediante un software

FIGURA C.8

El valor-p cuando $t = -2.13$ con 19 grados de libertad para una alternativa de una cola $\mu < 0$.



estadístico, como Stata, se puede calcular el valor-p exacto. Resulta ser de aproximadamente .023, que es una evidencia razonable contra H_0. Ésta es ciertamente suficiente evidencia para rechazar la hipótesis nula de que los subsidios para la capacitación no tienen efecto al nivel de significancia de 2.5% (y por tanto, al nivel de 5%).

Calcular un valor-p para una prueba de dos colas es similar, pero se debe dar cuenta de la naturaleza de dos colas de la regla de rechazo. Para la prueba t en torno a la media poblacional, el valor-p se calcula como

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \tag{C.42}$$

donde t es el valor del estadístico de prueba y T_{n-1} es una variable aleatoria t. (Para una n grande, remplazar T_{n-1} con una variable aleatoria estándar normal). Por tanto, se calcula el valor absoluto del estadístico t, se encuentra el área a la derecha de este valor en una distribución t_{n-1} y se multiplica el área por dos.

Para poblaciones no normales, el valor-p exacto puede ser difícil de obtener. Sin embargo, se pueden encontrar valores-p *asintóticos* al utilizar estos mismos cálculos. Estos valores-p son válidos para tamaños muestrales grandes. Para una n mayor que, por ejemplo, 120, se puede emplear también la distribución normal estándar. La tabla G.1 es lo bastante detallada para obtener valores-p precisos, pero también se puede usar un programa estadístico o de econometría.

Ejemplo C.7

[Discriminación racial en las contrataciones]

Mediante los datos pareados del Urban Institute ($n = 241$), se obtuvo $t = -4.29$. Si Z es una variable aleatoria normal estándar, $P(Z < -4.29)$ es, para fines prácticos, cero. En otras palabras, el valor-p asintótico para este ejemplo es esencialmente cero. Esta es una evidencia contundente contra H_0.

Resumen de cómo usar los valores-*p*:

i) Elija un estadístico T de prueba y decida la naturaleza de la alternativa. Esto determina si la regla de rechazo es $t > c$, $t < -c$, o $|t| > c$.

ii) Use el valor observado del estadístico t como el valor crítico y calcule el nivel de significancia correspondiente de la prueba. Este es el valor-p. Si la regla de rechazo es de la forma $t > c$, entonces el valor-$p = P(T > t)$. Si la regla de rechazo es $t < -c$, entonces el valor-$p = P(T < t)$; si la regla de rechazo es $|t| > c$, entonces el valor-$p = P(|T| > |t|)$.

iii) Si se eligió un nivel de significancia α entonces se rechaza H_0 al nivel $100{\cdot}\alpha\%$ si el valor-$p < \alpha$. Si el valor-$p \geq \alpha$, entonces no se rechaza H_0 al nivel $100{\cdot}\alpha\%$. Por tanto, este es un valor-p pequeño que lleva al rechazo.

La relación entre intervalos de confianza y pruebas de hipótesis

Debido a que construir intervalos de confianza y pruebas de hipótesis implican enunciados de probabilidad, es natural pensar que existe una conexión entre ellos. Resulta que así es. Después de que se construye el intervalo de confianza, se pueden realizar varias pruebas de hipótesis.

Los intervalos de confianza que se han analizado son todos de naturaleza de dos colas. (En este libro, no habrá necesidad de construir intervalos de confianza de una cola). Por tanto, los intervalos de confianza se pueden usar para probar contra alternativas de *dos colas*. En el caso de una media poblacional, la nula está dada por (C.31) y la alternativa por (C.34). Suponga que se ha construido un intervalo de confianza a 95% para μ. Entonces, si el valor hipotético de μ según H_0, μ_0, no está en el intervalo de confianza, entonces H_0: $\mu = \mu_0$ se rechaza contra H_1: $\mu \neq \mu_0$ al nivel de 5%. Si μ_0 yace en este intervalo, entonces no se rechaza H_0 al nivel de 5%. Observe cómo cualquier valor de μ_0 se puede probar una vez que se construye el intervalo de confianza y, dado que un intervalo de confianza contiene a más de un valor, existen varias hipótesis nulas que no se rechazarán.

Ejemplo C.8

[Subsidios de capacitación y productividad de los trabajadores]

En el ejemplo de Holzer *et al*. se construyó un intervalo de confianza a 95% para un cambio medio en la tasa de desperdicio μ cuando $[-2.28, -.02]$. Dado que el cero se excluyó de este intervalo, se rechazó H_0: $\mu = 0$ contra H_1: $\mu \neq 0$ al nivel de 5%. Este intervalo de confianza a 95% también significa que no se rechazó H_0: $\mu = -2$ al nivel de 5%. De hecho, existe una sucesión de hipótesis nulas que no se rechazan, dado el intervalo de confianza.

Significancia práctica frente a significancia estadística

En los ejemplos que se han dado hasta ahora, se han producido tres tipos de evidencia concernientes a los parámetros poblacionales: las estimaciones puntuales, los intervalos de confianza y las pruebas de hipótesis. Estas herramientas para aprender acerca de los parámetros poblacionales son igualmente importantes. Existe una tendencia comprensible de los estudiantes a enfocarse en los intervalos de confianza y las pruebas de hipótesis, pues son cosas a las que se puede atribuir niveles de confianza o significancia. Pero en cualquier estudio se deben interpretar las *magnitudes* de las estimaciones puntuales.

El signo y magnitud de $\bar{y}$ determina su **significancia práctica** y nos permite analizar la dirección de un efecto de intervención o política, y si el efecto estimado es "grande" o "pequeño". Por otra parte, la **significancia estadística** de $\bar{y}$ depende de la magnitud de su estadístico t. Para probar H_0: $\mu = 0$, el estadístico t es simplemente $t = \bar{y}/\mathrm{ee}(\bar{y})$. En otras palabras, la significancia estadística depende de la razón entre $\bar{y}$ y su error estándar. En consecuencia, un estadístico t puede ser grande debido a que $\bar{y}$ es grande o $\mathrm{ee}(\bar{y})$ es pequeño. En aplicaciones, es importante discutir tanto la significancia práctica como la estadística, y ser conscientes de que una estimación puede ser significativa estadísticamente sin ser especialmente grande en un sentido práctico. Que una estimación sea importante en la práctica depende del contexto así como del juicio propio, así que no existe un conjunto de reglas para determinar la significancia práctica.

Ejemplo C.9

[Efecto del ancho de la autopista en el tiempo de traslado]

Sea Y el cambio en el tiempo de traslado, medido en minutos, para los viajantes que viven en un área metropolitana antes de que una autopista se ampliará a después de su ampliación. Suponga que $Y \sim \mathrm{Normal}(\mu,\sigma^2)$. La hipótesis nula de que la ampliación no reduciría el tiempo promedio de traslado es H_0: $\mu = 0$; la alternativa de que redujo el tiempo promedio de traslado es H_1: $\mu < 0$. Suponga una muestra aleatoria de viajeros del tamaño $n = 900$ se obtiene para determinar la efectividad del proyecto de la autopista. El cambio promedio en el tiempo de traslado se calcula como de $\bar{y} = -3.6$ y la desviación estándar es $s = 32.7$; por tanto, $\text{ee}(\bar{y}) = 32.7/\sqrt{900} = 1.09$. El estadístico t es $t = -3.6/1.09 \approx -3.30$, lo cual es muy significativo en términos estadísticos; el valor-p es de aproximadamente .0005. Así, se concluye que la ampliación de la autopista tendría un efecto estadísticamente significativo en el tiempo de traslado.

Si el resultado de la prueba de hipótesis es la que se reportó en el estudio, se podría caer en un error. Reportar sólo la significancia estadística oculta el hecho de que la reducción estimada en el tiempo promedio de traslado, 3.6 minutos, es mínimo. Para ser francos, se debe reportar la estimación puntual de -3.6, junto con la prueba de significancia.

Encontrar estimaciones puntuales que sean estadísticamente significativas sin ser prácticamente trascendentes puede ocurrir cuando se trabaja con muestras grandes. Para analizar por qué sucede esto, es útil tener la siguiente información.

Consistencia de la prueba. Una **prueba consistente** rechaza H_0 con probabilidad más cerca de uno a medida que el tamaño de la muestra aumenta siempre que H_1 sea verdadera.

Otra forma de decir que una prueba es consistente es que, a medida que el tamaño de la muestra tiende a infinito, la potencia de la prueba se acerca cada vez más a la unidad siempre que H_1 sea verdadera. Todas las pruebas que se cubren en este libro tienen esta propiedad. En caso de la prueba de hipótesis acerca de una media poblacional, la consistencia de la prueba se desprende debido a que la varianza de $\bar{Y}$ converge a cero a medida que el tamaño de la muestra aumenta. El estadístico t para probar H_0: $\mu = 0$ es $T = \bar{Y}/(S/\sqrt{n})$. Dado que $\text{plim}(\bar{Y}) = \mu$ y $\text{plim}(S) = \sigma$, se desprende que si, por ejemplo, $\mu > 0$, entonces T aumenta cada vez más (con una probabilidad alta) cuando $n \to \infty$. En otras palabras, no importa qué tan cercana esté μ de cero, casi se puede estar seguro de rechazar H_0: $\mu = 0$ dado un tamaño de muestra suficientemente grande. Esto no dice nada acerca de si μ es grande en un sentido práctico.

C.7 Comentarios sobre la notación

En el repaso de probabilidad y estadística aquí y en el apéndice B, se ha tenido cuidado de utilizar convenciones estándar para denotar variables aleatorias, estimadores y estadísticos de prueba. Por ejemplo, se ha usado W para indicar un estimador (variable aleatoria) y w para denotar una estimación particular (resultado de la variable aleatoria W). Distinguir entre un estimador y una estimación es importante para comprender varios conceptos en las estimaciones y pruebas de hipótesis. No obstante, hacer esta distinción rápidamente se convierte en una carga en el análisis econométrico debido a que los modelos son más complicados: muchas variables aleatorias y parámetros estarán implicados, y apegarse a las convenciones usuales de la probabilidad y estadística requiere muchos símbolos adicionales.

En el texto principal, se utiliza una convención más simple que se emplea ampliamente en econometría. Si θ es un parámetro poblacional, la notación $\hat{\theta}$ ("theta con gorro") se usará para denotar tanto un estimador como una estimación de θ. La notación es útil en cuanto a que ofrece una forma sencilla de unir un estimador al parámetro poblacional que se supone se va a estimar. Por tanto, si el parámetro poblacional es β, entonces $\hat{\beta}$ denota un estimador o estimación de β; si el parámetro es σ^2, $\hat{\sigma}^2$ es un estimador o estimación de σ^2; y así sucesivamente. En ocasiones se analizarán dos estimadores del mismo parámetro, en cuyo caso se necesitará una notación diferente, como $\tilde{\theta}$ ("theta con gorro").

Aunque desechar las convenciones de probabilidad y estadística para indicar estimadores, variables aleatorias y estadísticos de prueba le impone más responsabilidad, no es muy difícil una vez que se ha comprendido la diferencia entre un estimador y una estimación. Si se están analizando las propiedades *estadísticas* de $\hat{\theta}$ como derivar si es insesgada o no, entonces se estará considerando necesariamente a $\hat{\theta}$ como un estimador. Por otra parte, si se escribe algo como $\hat{\theta} = 1.73$, entonces claramente se está denotando un estimador puntual de una muestra dada de datos. La confusión que puede surgir utilizando $\hat{\theta}$ para denotar ambos debería ser mínima una vez que se ha comprendido bien la probabilidad y la estadística.

RESUMEN

Se han analizado temas de estadística matemática que se basan en el análisis econométrico. La noción de un estimador, que es simplemente una regla para combinar datos para estimar un parámetro poblacional, es fundamental. Se han cubierto varias propiedades de los estimadores. Las propiedades más importantes de la muestra pequeña son el insesgamiento y la eficiencia, esta última depende de comparar varianzas cuando los estimadores son insesgados. Las propiedades de las muestras grandes tienen que ver con la secuencia de estimadores obtenida a medida que el tamaño de la muestra crece, y también son importantes en la econometría. Cualquier estimador útil es consistente. El teorema del límite central implica que, en las muestras grandes, la distribución de muestreo de la mayoría de los estimadores es aproximadamente normal.

La distribución muestral de un estimador se puede utilizar para construir intervalos de confianza. Se ve esto en la estimación de la media de una distribución normal y en el cálculo de intervalos de confianza aproximados en casos que no son normales. La clásica prueba de hipótesis, que requiere especificar una hipótesis nula, una hipótesis alternativa y un nivel de significancia, se realiza al comparar un estadístico de prueba con un valor crítico. Alternativamente, se puede calcular un valor-*p* que permita realizar una prueba a cualquier nivel de significancia.

TÉRMINOS CLAVE

Alternativa de dos colas
Alternativa de una cola
Coeficiente de correlación muestral
Covarianza muestral
Desviación estándar de muestreo
Desviación estándar muestral
Distribución de muestreo
Error cuadrático medio (ECM)
Error estándar
Error tipo I
Error tipo II
Estadístico de prueba
Estadístico *t*
Estimación
Estimador
Estimador consistente
Estimador de intervalo
Estimador de máxima verosimilitud
Estimador de mínimos cuadrados
Estimador insesgado
Estimador insesgado de varianza mínima
Estimador sesgado
Hipótesis alternativa
Hipótesis nula
Inconsistente
Intervalo de confianza
Ley de los grandes números (LGN)
Límite en probabilidad
Método de momentos
Muestra aleatoria
Nivel de significancia
Normalidad asintótica
Población
Potencia de una prueba
Promedio muestral
Prueba de dos colas
Prueba de hipótesis
Prueba de una cola
Región de rechazo
Sesgo
Significancia estadística
Significancia práctica
Teorema del límite central
Valor crítico
Valor-*p*
Varianza de muestreo
Varianza muestral

PROBLEMAS

C.1 Sea Y_1, Y_2, Y_3 y Y_4 variables aleatorias independientes, distribuidas idénticamente, de una población con media μ y varianza σ^2. Sea $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ el promedio de estas cuatro variables aleatorias.

i) ¿Cuál es el valor esperado y la varianza de $\bar{Y}$ en términos de μ y σ^2?

ii) Ahora, considere un estimador diferente de μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

Este es un ejemplo de promedio *ponderado* de Y_i. Muestre que W también es un estimador insesgado de μ. Determine la varianza de W.

iii) Con base en las respuestas en los incisos i) y ii), ¿qué estimador de μ se prefiere, $\bar{Y}$ o W?

C.2 Esta es una versión más general del problema C.1. Sean $Y_1, Y_2, \ldots, Y_n$ sean n variables aleatorias no correlacionadas con media común μ y varianza común σ^2. Sea $\bar{Y}$ el promedio muestral.

i) Defina la clase de *estimadores lineales* de μ mediante

$$W_a = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n,$$

donde a_i son constantes. ¿Qué restricción sobre a_i es necesaria para que W_a sea un estimador insesgado de μ?

ii) Determine la $\text{Var}(W_a)$.

iii) Para cualesquiera números $a_1, a_2, \ldots, a_n$, aplica la siguiente desigualdad: $(a_1 + a_2 + \ldots + a_n)^2/n \leq a_1^2 + a_2^2 + \ldots + a_n^2$. Use ésta, junto con las partes i) y ii) para mostrar que $\text{Var}(W_a) \geq \text{Var}(\bar{Y})$ siempre que W_a es insesgada, así que $\bar{Y}$ es el *mejor estimador insesgado lineal*. [*Sugerencia:* ¿en qué se convierte la desigualdad cuando a_i satisface la restricción del inciso i)?].

C.3 Sea $\bar{Y}$ el promedio muestral de una muestra aleatoria con media μ y varianza σ^2. Considere dos estimadores alternativos de μ: $W_1 = [(n-1)/n]\bar{Y}$ y $W_2 = \bar{Y}/2$.

i) Muestre que W_1 y W_2 son estimadores insesgados de μ y encuentre el sesgo. ¿Qué sucede a los sesgos cuando $n \to \infty$? Comente cualquier diferencia importante en el sesgo para los dos estimadores a medida que la muestra aumenta.

ii) Determine los límites en probabilidad de W_1 y W_2. {*Sugerencia:* Use las propiedades PLIM.1 y PLIM.2; para W_1, observe que $\text{plim}\,[(n-1)/n] = 1$.} ¿Qué estimador es consistente?

iii) Determine $\text{Var}(W_1)$ y $\text{Var}(W_2)$.

iv) Argumente que W_1 es un mejor estimador que $\bar{Y}$ si μ está "cercana" a cero. (Considere tanto el sesgo como la varianza).

C.4 Para las variables aleatorias positivas X y Y, suponga que el valor esperado de Y dada X es $\text{E}(Y|X) = \theta X$. El parámetro desconocido θ muestra cómo cambia el valor esperado de Y con X.

i) Defina la variable aleatoria $Z = Y/X$. Muestre que $\text{E}(Z) = \theta$. [*Sugerencia:* use la propiedad CE.2 junto con la ley de las esperanzas iteradas, la propiedad CE.4. En particular, primero muestre que $\text{E}(Z|X) = \theta$ y después utilice CE.4].

ii) Use el inciso i) para demostrar que el estimador $W_1 = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ es insesgado para θ, donde $\{(X_i,Y_i): i = 1, 2, \ldots, n\}$ es una muestra aleatoria.

iii) Explique por qué el estimador $W_2 = \bar{Y}/\bar{X}$, donde las barras superiores denotan promedios muestrales, no es lo mismo que W_1. Sin embargo, muéstrese que W_2 también es insesgado para θ.

iv) La siguiente tabla contiene a los datos sobre la producción de maíz para varios condados de Iowa. La USDA predice el número de hectáreas de maíz en cada país con base en fotos satelitales. Los investigadores cuentan el número de "pixeles" de maíz en la foto satelital (comparado con, por ejemplo, el número de pixeles de frijol de soya o de terrenos no cultivados) y los utilizan para predecir el número real de hectáreas. Para desarrollar una ecuación de predicción que empleen los condados en general, la USDA aplicó una encuesta a los granjeros en condados seleccionados para obtener las producciones de maíz en hectáreas. Sea Y_i = la producción de maíz en el condado i y X_i = el número de pixeles de maíz en la fotografía satelital para el condado i. Existen $n = 17$ observaciones para ocho condados. Utilice esta muestra para calcular las estimaciones de θ diseñadas en los incisos ii) y iii). ¿Las estimaciones son similares?

Gráfica	Producción de maíz	Pixeles de maíz
1	165.76	374
2	96.32	209
3	76.08	253
4	185.35	432
5	116.43	367
6	162.08	361
7	152.04	288
8	161.75	369
9	92.88	206
10	149.94	316
11	64.75	145
12	127.07	355
13	133.55	295
14	77.70	223
15	206.39	459
16	108.33	290
17	118.17	307

C.5 Sea Y una variable aleatoria de Bernoulli(θ) con $0 < \theta < 1$. Suponga que interesa estimar la *razón probabilística*, $\gamma = \theta/(1 - \theta)$, que es la probabilidad de éxito sobre la probabilidad de fracaso. Dada una muestra aleatoria $\{Y_1, \ldots, Y_n\}$, se sabe que un estimador insesgado y consistente de θ es $\bar{Y}$, la proporción de éxito en n pruebas. Un estimador natural de γ es $G = \bar{Y}/(1 - \bar{Y})$, la proporción de éxitos sobre la proporción de fracasos en la muestra.

i) ¿Por qué G no es un estimador insesgado de γ?

ii) Use PLIM.2(iii) para mostrar que G es un estimador consistente de γ.

C.6 A usted lo contrata el gobernador para estudiar si un impuesto sobre el licor ha reducido el consumo promedio de licor en su estado. Usted puede obtener, para una muestra de individuos seleccionados al azar, la diferencia en el consumo de licor (en onzas) para los años anteriores y posteriores al impuesto. Para la persona i que forma parte de la muestra aleatoria de la población, Y_i denota el cambio en el consumo de licor. Trate a éstas como una muestra aleatoria de una distribución Normal(μ,σ^2) .

i) La hipótesis nula es que no hubo cambio en el consumo de licor. Exprese esto formalmente en términos de μ.

ii) La alternativa es que hubo un declive en el consumo del licor; enuncie la alternativa en términos de μ.

iii) Ahora, suponga que el tamaño de su muestra es $n = 900$ y que obtiene las estimaciones $\bar{y} = -32.8$ y $s = 466.4$. Calcule el estadístico t para probar H_0 contra H_1; obtenga el valor-p para la prueba. (Debido al tamaño grande de la muestra, sólo use la distribución normal estándar tabulada en la tabla G.1). ¿Rechaza H_0 al nivel de 5%? ¿Al nivel de 1%?

iv) ¿Diría usted que la disminución estimada en el consumo es grande en magnitud? Comente sobre la significancia estadística contra la significancia práctica de esta estimación.

v) ¿Qué se ha supuesto implícitamente en su análisis acerca de otras determinantes del consumo de licor durante el periodo de dos años con el fin de inferir la causalidad del cambio fiscal en el consumo de licor?

C.7 La nueva administración de una pastelería afirma que ahora los trabajadores son más productivos que con la antigua administración, debido a que los salarios hayan “generalmente aumentado”. Sea W_i^b el salario del trabajador i bajo la antigua administración y sea W_i^a el salario del trabajador i después del cambio. La diferencia es $D_i \equiv W_i^a - W_i^b$. Suponga que las D_i son muestras aleatorias de una distribución Normal(μ,σ^2).

i) Mediante los siguientes datos sobre 15 trabajadores, construya un intervalo de confianza a 95% para μ.

ii) Formalmente exprese la hipótesis nula de que no ha habido cambio en el promedio de los salarios. En particular, ¿qué es $E(D_i)$ según H_0? Si lo contratan para examinar la validez de la afirmación de la nueva dirección, ¿cuál es la hipótesis alternativa relevante en términos de $\mu = E(D_i)$?

iii) Pruebe la hipótesis nula del inciso ii) frente a la alternativa expresada a los niveles de 5% y 1%.

iv) Obtenga el valor-p para la prueba en el inciso iii).

Trabajador	Salario anterior	Salario posterior
1	8.30	9.25
2	9.40	9.00

(*continúa*)

Trabajador	Salario anterior	Salario posterior
3	9.00	9.25
4	10.50	10.00
5	11.40	12.00
6	8.75	9.50
7	10.00	10.25
8	9.50	9.50
9	10.80	11.50
10	12.55	13.10
11	12.00	11.50
12	8.65	9.00
13	7.75	7.75
14	11.25	11.50
15	12.65	13.00

C.8 El *New York Times* (2/5/90) reportó las anotaciones de canasta de tres puntos de los mejores diez tiradores en esta categoría de la NBA. La siguiente tabla resume estos datos:

Jugador	IA-AL
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Nota: *IA* = intentos de anotación y AL anotaciones logradas.

Para un jugador dado, el resultado de un tiro particular se puede representar como una variable Bernoulli (cero-uno): si Y_i es el resultado del tiro i, entonces $Y_i = 1$ si se logró anotar y $Y_i = 0$ si se falló el tiro. Sea θ la probabilidad de hacer cualquier intento de lanzamiento de tres puntos. El estimador natural de θ es $\bar{Y} = FGM/FGA$.

i) Estime θ para Mark Price.

ii) Encuentre la desviación estándar del estimador $\bar{Y}$ en términos de θ y el número de intentos de anotación, n.

iii) La distribución asintótica de $(\bar{Y} - \theta)/\text{ee}(\bar{Y})$ es normal estándar, donde $\text{ee}(\bar{Y}) = \sqrt{\bar{Y}(1 - \bar{Y})/n}$. Use este dato para probar H_0: $\theta = .5$ frente a H_1: $\theta < .5$ para Mark Price. Use un nivel de significancia a 1%.

C.9 Suponga que un dictador militar en un país determinado realiza un plebiscito (un voto de confianza de sí o no) y afirma que tiene el apoyo de 65% de los votantes. Un grupo de derechos humanos sospecha que hay algún tipo de irregularidad y lo contrata a usted para probar la validez de la afirmación del dictador. Usted tiene un presupuesto que le permite obtener una muestra aleatoria de 200 votantes de ese país.

i) Sea X el número de votos afirmativos obtenidos de una muestra aleatoria de 200 de entre toda la población votante. ¿Cuál es el valor esperado de X si, de hecho, 65% de todos los votantes apoyaron al dictador?

ii) ¿Cuál es la desviación estándar de X, si se supone nuevamente que la fracción verdadera que votó que sí en el plebiscito es .65?

iii) Ahora, usted reúne una muestra de 200, y encuentra que 115 personas en realidad votaron que sí. Use el TLC para aproximar la probabilidad de que encuentre 115 o menos votos de sí de una muestra aleatoria de 200 si, en realidad, 65% de la población entera votó que sí.

iv) ¿Cómo explicaría la relevancia del resultado en el inciso iii) a alguien que no tiene conocimientos de estadística?

C.10 Antes de que una huelga concluyera prematuramente la temporada de la liga mayor de béisbol en 1994, Tony Gwynn de los Padres de San Diego había realizado 165 *hits* en 419 turnos al bate, para un promedio de bateo de .394. Hubo una discusión sobre si Gwynn hubiera podido lograr .400 ese año. Este aspecto puede expresarse en términos de la probabilidad de que Gwynn consiguiera un *hit* en un bateo determinado, llámelo θ. Sea Y_i el indicador de Bernoulli(θ) igual a la unidad si Gwynn hace un *hit* durante su i-ésimo bateo y cero si no. Así, $Y_1, Y_2, \ldots, Y_n$ es una muestra aleatoria de una distribución de Bernoulli(θ) en la que θ es la probabilidad de éxito, y $n = 419$.

La mejor estimación puntual de θ es el promedio de bateo de Gwynn, el cual es sólo la proporción de éxitos: $\bar{y} = .394$. Mediante el hecho de que $\text{ee}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, constrúyase un intervalo de confianza aproximado a 95% para θ, utilizando la distribución normal estándar. ¿Se diría que ésta es una evidencia contundente en contra de que Gwynn fuera un bateador potencial de .400? Explique.

Apéndice D

Resumen de álgebra matricial

Este apéndice resume los conceptos de álgebra matricial, incluida el álgebra probabilística, necesarias para el estudio de modelos de regresión lineal múltiple mediante las matrices del apéndice E. Nada de este material se utiliza en el texto principal.

D.1 Definiciones básicas

Definición D.1 (Matriz). Una **matriz** es un arreglo rectangular numérico. En otras palabras y en términos más precisos, una matriz $m \times n$ tiene m filas y n columnas. El entero positivo m se llama *dimensión de la fila o renglón*, y n es la *dimensión de la columna*.

Se usan letras mayúsculas en bold para denotar matrices. En términos generales se puede escribir una matriz $m \times n$ como

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix},$$

donde a_{ij} representa el elemento en el fila i-ésima y en la columna j-ésima. Por ejemplo, a_{25} representa el número en la fila segunda y la quinta columna de $\mathbf{A}$. Un ejemplo específico de una matriz 2×3 es

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \qquad \boxed{\mathbf{D.1}}$$

donde $a_{13} = 7$. La abreviatura $\mathbf{A} = [a_{ij}]$ suele usarse para definir las operaciones de matrices.

Definición D.2 (Matriz cuadrada). Una **matriz cuadrada** tiene el mismo número de filas que de columnas. La dimensión de una matriz cuadrada es su número de filas y columnas.

Definición D.3 (Vectores)

i) Una matriz $1 \times m$ se denomina **vector fila** (de dimensión m) y se puede escribir como $\mathbf{x} \equiv (x_1, x_2, \ldots, x_m)$.

ii) Una matriz $n \times 1$ se llama **vector columna** y se puede escribir como

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Definición D.4 (Matriz diagonal). Una matriz cuadrada **A** es una **matriz diagonal** si todos sus elementos fuera de la diagonal son cero, es decir, $a_{ij} = 0$ para toda $i \neq j$. La matriz diagonal se puede escribir como:

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ 0 & a_{22} & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & a_{nn} \end{bmatrix}.$$

Definición D.5 (Matriz identidad y matriz cero)

i) La **matriz identidad** $n \times n$, denotada como **I**, o algunas veces $\mathbf{I}_n$ para enfatizar su dimensión, es la matriz diagonal con la unidad (uno) en cada posición diagonal y cero en cualquier otro lado:

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

ii) La **matriz cero** $m \times n$ denotada como **0**, es la matriz $m \times n$ con cero para todas las entradas. No necesariamente tiene que ser una matriz cuadrada.

D.2 Operaciones matriciales

Suma matricial

Dos matrices **A** y **B**, cada una con dimensión $m \times n$, se pueden sumar elemento por elemento: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. De manera más precisa,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \ldots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \ldots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \ldots & a_{mn} + b_{mn} \end{bmatrix}.$$

Por ejemplo,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Las matrices de dimensiones diferentes no pueden sumarse.

Multiplicación escalar

Dado cualquier número real γ (llamado escalar), la **multiplicación escalar** se define como $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, o

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \ldots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \ldots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \ldots & \gamma a_{mn} \end{bmatrix}.$$

Por ejemplo, si $\gamma = 2$ y $\mathbf{A}$ es la matriz en la ecuación (D.1), entonces

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

Multiplicación matricial

Para multiplicar la matriz $\mathbf{A}$ por la matriz $\mathbf{B}$ y formar el producto $\mathbf{AB}$, la dimensión *columna* de $\mathbf{A}$ debe ser igual a la dimensión *fila* de $\mathbf{B}$. Por tanto, sea $\mathbf{A}$ una matriz $m \times n$ y sea $\mathbf{B}$ una matriz $n \times p$. Entonces, la **multiplicación matricial** se define como

$$\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik}b_{kj}\right].$$

En otras palabras, el elemento $(i,j)^{\text{ésimo}}$ de una nueva matriz $\mathbf{AB}$ se obtiene al multiplicar cada elemento en la i-ésima fila de $\mathbf{A}$ por el elemento correspondiente en la j-ésima columna de $\mathbf{B}$ y después sumar esos productos n. Un esquema puede ayudar a aclarar este proceso:

$$i\text{-ésima fila} \rightarrow \overset{\mathbf{A}}{\begin{bmatrix} \\ a_{i1}\, a_{i2}\, a_{i3} \ldots a_{in} \\ \\ \end{bmatrix}} \overset{\mathbf{B}}{\underset{\substack{\uparrow \\ \text{columna} \\ j\text{-ésima}}}{\begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix}}} = \overset{\mathbf{AB}}{\underset{\substack{\uparrow \\ \text{elemento} \\ (i,j)\text{-ésimo}}}{\begin{bmatrix} \\ \sum_{k=1}^{n} a_{ik}b_{kj} \\ \\ \end{bmatrix}}},$$

donde, por la definición del operador de suma del apéndice A,

$$\sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}.$$

Por ejemplo,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

También se puede multiplicar una matriz y un vector. Si $\mathbf{A}$ es una matriz $n \times m$ matrix y $\mathbf{y}$ es un vector $m \times 1$, entonces $\mathbf{Ay}$ es un vector $n \times 1$. Si $\mathbf{x}$ es un vector $1 \times n$, entonces $\mathbf{xA}$ es un vector $1 \times m$.

La suma matricial, la multiplicación escalar y la multiplicación matricial pueden combinarse de varias formas, y estas operaciones satisfacen varias reglas que son familiares debido a las operaciones numéricas básicas. En la siguiente lista de propiedades, $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ son matrices con dimensiones apropiadas para aplicar cada operación, y α y β son números reales. La mayoría de esas propiedades son fáciles de ilustrar a partir de las definiciones.

Propiedades de la multiplicación matricial. 1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; 2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; 3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; 4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; 5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; 6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; 7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; 8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; 9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; 10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; 11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; 12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; 13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$ y 14) $\mathbf{AB} \neq \mathbf{BA}$, incluso cuando ambos productos estén definidos.

La última propiedad merece un comentario adicional. Si $\mathbf{A}$ es $n \times m$ y $\mathbf{B}$ es $m \times p$, entonces $\mathbf{AB}$ está definida, pero $\mathbf{BA}$ está definida sólo si $n = p$ (la dimensión fila de $\mathbf{A}$ es igual a la dimensión columna de $\mathbf{B}$). Si $\mathbf{A}$ es $m \times n$ y $\mathbf{B}$ es $n \times m$, entonces $\mathbf{AB}$ y $\mathbf{BA}$ están definidas, pero generalmente no son iguales; de hecho, tienen diferentes dimensiones, a menos que $\mathbf{A}$ y $\mathbf{B}$ no sean matrices cuadradas. Aun cuando si $\mathbf{A}$ y $\mathbf{B}$ sean cuadradas $\mathbf{AB} \neq \mathbf{BA}$, salvo en circunstancias especiales.

Transposición

Definición D.6 (Transposición). Sea $\mathbf{A} = [a_{ij}]$ una matriz $m \times n$. La **transpuesta** de $\mathbf{A}$, denotada como $\mathbf{A}'$ (llamada $\mathbf{A}$ *prima*), es la matriz $n \times m$ obtenida al intercambiar las filas y las columnas de $\mathbf{A}$. Se puede escribir esto como $\mathbf{A}' \equiv [a_{ji}]$.

Por ejemplo,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \qquad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

Propiedades de la transposición. 1) $(\mathbf{A}')' = \mathbf{A}$; 2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ para todo escalar α; 3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; 4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, donde $\mathbf{A}$ es $m \times n$ y $\mathbf{B}$ es $n \times k$; 5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$, donde $\mathbf{x}$ es un vector $n \times 1$; y 6) Si $\mathbf{A}$ es una matriz $n \times k$ con *filas* dadas por los vectores $1 \times k$, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, así que se puede escribir

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix},$$

entonces $\mathbf{A}' = (\mathbf{a}_1' \; \mathbf{a}_2' \ldots \mathbf{a}_n')$.

Definición D.7 (Matriz simétrica). Una matriz cuadrada $\mathbf{A}$ es una **matriz simétrica** si, y sólo si, $\mathbf{A}' = \mathbf{A}$.

Si $\mathbf{X}$ toda matriz $n \times k$ entonces $\mathbf{X}'\mathbf{X}$ está siempre definida y es una matriz simétrica, como puede verse al aplicar la primera y cuarta propiedades transpuestas (vea el problema D.3).

Multiplicación parcial particionada

Sea $\mathbf{A}$ una matriz $n \times k$ con filas dadas por los vectores $1 \times k$ $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, y sea $\mathbf{B}$ una matriz $n \times m$ con filas dadas por los vectores $1 \times m$ $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Entonces,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{b}_i,$$

donde para cada i, $\mathbf{a}_i'\mathbf{b}_i$ es una matriz $k \times m$. Por tanto, $\mathbf{A}'\mathbf{B}$ puede escribirse como la suma de n matrices, cada una de las cuales es $k \times m$. Como caso especial, se tiene

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{a}_i,$$

donde $\mathbf{a}_i'\mathbf{a}_i$ es una matriz $k \times k$ para toda i.

Traza

La traza de una matriz es una operación muy simple definida sólo para matrices *cuadradas*.

Definición D.8 (Traza). Para toda matriz $\mathbf{A}$ de $n \times n$, la **traza de una matriz A**, denotada como $\mathrm{tr}(\mathbf{A})$, es la suma de sus elementos diagonales. En términos matemáticos

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Propiedades de la traza. 1) $\mathrm{tr}(\mathbf{I}_n) = n$; 2) $\mathrm{tr}(\mathbf{A}') = \mathrm{tr}(\mathbf{A})$; 3) $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$; 4) $\mathrm{tr}(\alpha\mathbf{A}) = \alpha\mathrm{tr}(\mathbf{A})$, para toda escalar α; y 5) $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, donde $\mathbf{A}$ es $m \times n$ y $\mathbf{B}$ es $n \times m$.

Inversa

La noción de matriz inversa es muy importante para las matrices cuadradas.

Definición D.9 (Inversa). Una matriz $\mathbf{A}$ de $n \times n$ tiene una **inversa**, denotada como $\mathbf{A}^{-1}$, si $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ y $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. En este caso se dice que $\mathbf{A}$ es *invertible* o *no singular*. De lo contrario, se dice que es *no invertible* o *singular*.

Propiedades de la inversa. 1) Si existe una inversa, es única; 2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, si $\alpha \neq 0$ y $\mathbf{A}$ es invertible; 3) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, si tanto $\mathbf{A}$ como $\mathbf{B}$ son invertibles $n \times n$ y 4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

Aquí no interesa la mecánica de calcular la inversa de una matriz. Cualquier libro de álgebra matricial contiene ejemplos detallados de tales cálculos.

D.3 Independencia lineal y rango de una matriz

Para un conjunto de vectores que tienen la misma dimensión, es importante saber si un vector se puede expresar como una combinación lineal de los vectores restantes.

Definición D.10 (Independencia lineal). Sea $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ un conjunto de $n \times 1$ vectores. Éstos son **vectores linealmente independientes** si, y sólo si,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_r\mathbf{x}_r = \mathbf{0} \tag{D.2}$$

implica que $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$. Si (D.2) aplica para un conjunto de escalares no todas iguales a cero, entonces $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ es *linealmente dependiente*.

La afirmación de que $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ es linealmente dependiente es equivalente a decir que al menos un vector en este grupo se puede escribir como una combinación lineal de los demás.

Definición D.11 (Rango)

i) Sea $\mathbf{A}$ una matriz $n \times m$. El **rango de una matriz A**, denotado como rango($\mathbf{A}$), es el máximo número de columnas linealmente independientes de $\mathbf{A}$.

ii) Si $\mathbf{A}$ es $n \times m$ y rango($\mathbf{A}$) $= m$, entonces $\mathbf{A}$ tiene un *rango pleno por columnas*.

Si $\mathbf{A}$ es $n \times m$, su rango puede ser a lo sumo m. Una matriz tiene un rango pleno por columnas, si sus columnas forman un conjunto linealmente independiente. Por ejemplo, la matriz 3×2.

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

se puede tener a lo sumo un rango dos. De hecho, su rango es sólo uno porque la segunda columna es tres veces la primera.

Propiedades del rango. 1) rango($\mathbf{A}'$) $=$ rango($\mathbf{A}$); 2) Si $\mathbf{A}$ es $n \times k$, entonces rango($\mathbf{A}$) $\leq$ min(n,k) y 3) Si $\mathbf{A}$ es $k \times k$ y rango($\mathbf{A}$) $= k$, entonces $\mathbf{A}$ es invertible.

D.4 Formas cuadráticas y matrices definidas positivas

Definición D.12 (Forma cuadrática). Sea $\mathbf{A}$ una matriz $n \times n$ *simétrica*. La **forma cuadrática** asociada con la matriz $\mathbf{A}$ es la función de valor real definida para todo vector $n \times 1$ de $\mathbf{x}$:

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>1} a_{ij}x_ix_j.$$

Definición D.13 (Definida positiva y semidefinida positiva)

i) Se dice que una matriz **A** simétrica es **definida positiva** (**d.p.**) si

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ para todo vector } n \times 1 \text{ de } \mathbf{x} \text{ excepto } \mathbf{x} = \mathbf{0}.$$

ii) Una matriz simétrica **A** es **semidefinida positiva (p.s.d.)** si

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ para todo vector } n \times 1.$$

Si una matriz es definida positiva o semidefinida positiva, automáticamente se asume que es simétrica.

Propiedades de las matrices definidas positivas y semidefinidas positivas. 1) Una matriz definida positiva tiene elementos diagonales que son estrictamente positivos, mientas que una matriz semidefinida positiva tiene elementos diagonales no negativos; 2) si **A** es d.p., entonces $\mathbf{A}^{-1}$ existe y es su d.p.; 3) Si **X** es $n \times k$, entonces $\mathbf{X}'\mathbf{X}$ y $\mathbf{X}\mathbf{X}'$ son s.d.p.; y 4) Si **X** es $n \times k$ y rango($\mathbf{X}$) $= k$, entonces $\mathbf{X}'\mathbf{X}$ es d.p. (y por tanto, no singular).

D.5 Matrices idempotentes

Definición D.14 (matrices idempotentes). Sea **A** una matriz simétrica $n \times n$. Entonces, se dice que **A** es una **matriz idempotente** si y sólo si, $\mathbf{AA} = \mathbf{A}$.

Por ejemplo,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

es una matriz idempotente, como lo comprueba la multiplicación directa.

Propiedades de las matrices idempotentes. Sea **A** una matriz $n \times n$ idempotente. 1) rango($\mathbf{A}$) $=$ tr($\mathbf{A}$) y 2) **A** es semidefinida positiva.

Se pueden construir matrices idempotentes de manera muy general. Sea **X** una matriz $n \times k$ de rango($\mathbf{X}$) $= k$. Se define

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Entonces **P** y **M** son matrices simétricas, idempotentes, con rango($\mathbf{P}$) $= k$ y rango($\mathbf{M}$) $= n - k$. Los rangos se obtienen más fácilmente mediante la propiedad 1: tr($\mathbf{P}$) $=$ tr[$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$] (de la propiedad 5 de la traza) $=$ tr($\mathbf{I}_k$) $= k$ (por la propiedad 1 de la traza). Se desprende que tr($\mathbf{M}$) $=$ tr($\mathbf{I}_n$) $-$ tr($\mathbf{P}$) $= n - k$.

D.6 Diferenciación de formas lineales y cuadráticas

Para un vector **a** de $n \times 1$ considere la función lineal definida por

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x},$$

para todos los vectores $n \times 1$ de **x**. La derivada de f respecto a **x** es el vector de $1 \times n$ de las derivadas parciales, que simplemente es

$$\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'.$$

Para una matriz **A** simétrica de $n \times n$ se define la forma cuadrática

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

Entonces,

$$\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A},$$

que es un vector $1 \times n$.

D.7 Momentos y distribuciones de vectores aleatorios

Con el fin de derivar el valor esperado y la varianza de los estimadores MCO mediante matrices, es necesario definir el valor esperado y la varianza de un **vector aleatorio**. Como su nombre sugiere, un vector aleatorio simplemente es un vector de variables aleatorias. También es necesario definir la distribución normal multivariada. Estos conceptos son simplemente extensiones de los que se abordaron en el apéndice B.

Valor esperado

Definición D.15 (Valor esperado)

i) Si **y** es un vector $n \times 1$ aleatorio, el **valor esperado** de **y**, denotado como E(**y**), es el vector de valores esperados: $\mathrm{E}(\mathbf{y}) = [\mathrm{E}(y_1), \mathrm{E}(y_2), \ldots, \mathrm{E}(y_n)]'$.

ii) Si **Z** es una matriz aleatoria de $n \times m$, E(**Z**) es la matriz $n \times m$ de valores esperados: $\mathrm{E}(\mathbf{Z}) = [\mathrm{E}(z_{ij})]$.

Propiedades del valor esperado. 1) Si **A** es una matriz $m \times n$ y **b** es un vector $n \times 1$ ambos son no aleatorios, entonces $\mathrm{E}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}\mathrm{E}(\mathbf{y}) + \mathbf{b}$; y 2) Si **A** es $p \times n$ y **B** es $m \times k$, donde ambos son no aleatorios, entonces $\mathrm{E}(\mathbf{AZB}) = \mathbf{A}\mathrm{E}(\mathbf{Z})\mathbf{B}$.

Matriz varianza-covarianza

Definición D.16 (Matriz varianza-covarianza). Si **y** es un vector aleatorio $n \times 1$, su **matriz varianza-covarianza**, denotada como Var(**y**), se define como

$$\text{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix},$$

donde $\sigma_j^2 = \text{Var}(y_j)$ y $\sigma_{ij} = \text{Cov}(y_i, y_j)$. En otras palabras, la matriz varianza-covarianza tiene las varianzas de cada elemento de $\mathbf{y}$ bajo su diagonal, con términos de covarianza fuera de las diagonales. Debido a que $\text{Cov}(y_i, y_j) = \text{Cov}(y_j, y_i)$, se desprende que la matriz varianza-covarianza es simétrica.

Propiedades de la varianza. 1) Si $\mathbf{a}$ es un vector no aleatorio de $n \times 1$ entonces $\text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'[\text{Var}(\mathbf{y})]\mathbf{a} \geq 0$; 2) Si $\text{Var}(\mathbf{a}'\mathbf{y}) > 0$ para toda $\mathbf{a} \neq \mathbf{0}$, $\text{Var}(\mathbf{y})$ es definida positiva; 3) $\text{Var}(\mathbf{y}) = \text{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$, donde $\boldsymbol{\mu} = \text{E}(\mathbf{y})$; 4) Si los elementos de $\mathbf{y}$ no se correlacionan, $\text{Var}(\mathbf{y})$ es una matriz diagonal. Si, además, $\text{Var}(y_j) = \sigma^2$ para $j = 1, 2, \ldots, n$, entonces $\text{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$; y 5) Si $\mathbf{A}$ es una matriz de $m \times n$ no aleatoria y $\mathbf{b}$ es un vector no aleatorio de $n \times 1$, entonces $\text{Var}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}[\text{Var}(\mathbf{y})]\mathbf{A}'$.

Distribución normal multivariada

La distribución normal de una variable aleatoria se analizó con cierto detalle en el apéndice B. Se debe ampliar la distribución normal a vectores no aleatorios. No se proporcionará una expresión para la función de distribución de probabilidad, puesto que no es necesaria. Es importante saber que un vector aleatorio normal multivariado se caracteriza completamente por su media y su matriz varianza-covarianza. Por tanto, si $\mathbf{y}$ es un vector aleatorio normal multivariado de $n \times 1$ con media $\boldsymbol{\mu}$ y matriz varianza-covarianza $\boldsymbol{\Sigma}$, se escribe $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ahora se enunciarán las propiedades de la **distribución normal multivariada**.

Propiedades de la distribución normal multivariada. 1) Si $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces cada elemento de $\mathbf{y}$ se distribuye normalmente; 2) Si $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces y_i y y_j, dos elementos cualquiera de $\mathbf{y}$, son independientes si, y sólo si, no se correlacionan, es decir, $\sigma_{ij} = 0$; 3) Si $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces $\mathbf{Ay} + \mathbf{b} \sim \text{Normal}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, donde $\mathbf{A}$ y $\mathbf{b}$ son no aleatorias; 4) Si $\mathbf{y} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$, entonces, para las matrices aleatorias $\mathbf{A}$ y $\mathbf{B}$, $\mathbf{Ay}$ y $\mathbf{By}$ son independientes si, y sólo si, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. En particular, si $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, entonces $\mathbf{AB}' = \mathbf{0}$ es necesaria y suficiente para la independencia de $\mathbf{Ay}$ y $\mathbf{By}$; 5) Si $\mathbf{y} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{A}$ es una matriz no aleatoria $k \times n$, y $\mathbf{B}$ es una matriz $n \times n$ simétrica e idempotente, entonces $\mathbf{Ay}$ y $\mathbf{y}'\mathbf{By}$ son independientes si, y sólo si, $\mathbf{AB} = \mathbf{0}$; y 6) Si $\mathbf{y} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ y $\mathbf{A}$ y $\mathbf{B}$ son matrices idempotentes simétricas no aleatorias, entonces $\mathbf{y}'\mathbf{Ay}$ y $\mathbf{y}'\mathbf{By}$ son independientes si, y sólo si, $\mathbf{AB} = \mathbf{0}$.

Distribución *ji*-cuadrada

En el apéndice B, se definió una **variable aleatoria *ji*-cuadrada** como la suma de las variables *cuadradas* aleatorias independientes y normales estandarizadas. En notación vectorial, si $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, entonces $\mathbf{u}'\mathbf{u} \sim \chi_n^2$.

Propiedades de la distribución *ji*-cuadrada. 1) Si $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$ y $\mathbf{A}$ es una matriz $n \times n$ simétrica e idenmpotente con $\text{rango}(\mathbf{A}) = q$, entonces $\mathbf{u}'\mathbf{Au} \sim \chi_q^2$; 2) Si $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$ y $\mathbf{A}$ y $\mathbf{B}$

son matrices $n \times n$ simétricas, idempotentes tal que $\mathbf{AB} = \mathbf{0}$, entonces $\mathbf{u}'\mathbf{Au}$ y $\mathbf{u}'\mathbf{Bu}$ son variables independientes aleatorias *ji*-cuadradas; y 3) Si $\mathbf{z} \sim \text{Normal}(\mathbf{0},\mathbf{C})$ donde $\mathbf{C}$ es una matriz $m \times m$ no singular, entonces, $\mathbf{z}'\mathbf{C}^{-1}\mathbf{z} \sim \chi^2_m$.

Distribución *t*

También se definió la **distribución *t*** en el apéndice **B**. Ahora se agregará una propiedad importante.

Propiedad de la distribución *t*. Si $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, $\mathbf{c}$ es un vector de $n \times 1$ no aleatorio, $\mathbf{A}$ es una matriz $n \times n$ no aleatoria simétrica e idempotente con rango q, y $\mathbf{Ac} = \mathbf{0}$, entonces $\{\mathbf{c}'\mathbf{u}/(\mathbf{c}'\mathbf{c})^{1/2}\}/(\mathbf{u}'\mathbf{Au})^{1/2} \sim t_q$.

Distribución *F*

Recuerde que la **variable aleatoria *F*** se obtiene al determinar dos variables aleatorias *ji*-cuadradas *independientes* y se calcula la razón de cada una estandarizada por sus grados de libertad.

Propiedad de la distribución *F*. Si $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ y $\mathbf{A}$ y $\mathbf{B}$ son matrices no aleatorias simétricas e idempotentes con rango $n \times n$, $\text{rango}(\mathbf{A}) = k_1$, $\text{rango}(\mathbf{B}) = k_2$ y $\mathbf{AB} = \mathbf{0}$, entonces $(\mathbf{u}'\mathbf{Au}/k_1)/(\mathbf{u}'\mathbf{Bu}/k_2) \sim F_{k_1,k_2}$.

RESUMEN

Este apéndice contiene una forma condensada de la información general necesaria para estudiar el modelo lineal clásico mediante matrices. Aunque el material que aquí se presenta es independiente, tiene el objetivo de ser un repaso para los lectores familiarizados con el álgebra matricial y la estadística multivariada, y se usará ampliamente en el apéndice E.

TÉRMINOS CLAVE

Definida positiva
Distribución normal multivariada
Distribución *t*
Forma cuadrática
Inversa
Matriz
Matriz cero
Matriz cuadrada
Matriz diagonal
Matriz idempotente
Matriz identidad
Matriz simétrica
Matriz varianza-covarianza
Multiplicación escalar
Multiplicación matricial
Rango de una matriz
Semidefinida positiva
Transpuesta
Traza de una matriz
Valor esperado
Variable aleatoria *F*
Variable aleatoria *ji*-cuadrada
Vector aleatorio
Vector columna
Vector fila
Vectores linealmente independientes

PROBLEMAS

D.1 i) Determine el producto **AB** mediante

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

ii) ¿Existe **BA**?

D.2 Si **A** y **B** son matrices diagonales $n \times n$, muestre que $\mathbf{AB} = \mathbf{BA}$.

D.3 Sea **X** cualquier matriz $n \times k$. Muestre que $\mathbf{X}'\mathbf{X}$ es una matriz simétrica.

D.4 i) Use las propiedades de traza para argumentar que $\mathrm{tr}(\mathbf{A}'\mathbf{A}) = \mathrm{tr}(\mathbf{AA}')$ para toda matriz $n \times m$ de **A**.

ii) Para $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, verifique que $\mathrm{tr}(\mathbf{A}'\mathbf{A}) = \mathrm{tr}(\mathbf{AA}')$.

D.5 i) Use la definición de la inversa para demostrar lo siguiente: si **A** y **B** son matrices no singulares $n \times n$, entonces $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

ii) Si **A**, **B** y **C** son matrices $n \times n$ no singulares, encuentre $(\mathbf{ABC})^{-1}$ en términos de $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$ y $\mathbf{C}^{-1}$.

D.6 i) Muestre que si **A** es una matriz $n \times n$ simétrica y definida positiva, entonces **A** debe tener elementos diagonales estrictamente positivos.

ii) Escriba una matriz simétrica 2×2 con elementos diagonales estrictamente positivos, que *no* sea definida positiva.

D.7 Sea **A** una matriz $n \times n$ simétrica y definida positiva. Muestre que si **P** es una matriz $n \times n$ no singular, entonces $\mathbf{P}'\mathbf{AP}$ es definida positiva.

D.8 Demuestre la propiedad 5 de las varianzas de vectores mediante la propiedad 3.

Apéndice E

El modelo de regresión lineal en forma matricial

Este apéndice deriva diferentes resultados para la estimación de los mínimos cuadrados ordinarios del modelo de regresión lineal múltiple mediante notación y álgebra matricial (vea un resumen en el apéndice D). El material aquí presentado es mucho más avanzado que el del texto.

E.1 El modelo de estimación de los mínimos cuadrados ordinarios

En este apéndice se usa el subíndice t para indexar observaciones y n para denotar el tamaño muestral. Es útil escribir el modelo de regresión lineal múltiple con k parámetros así:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,\ t = 1, 2, \ldots, n, \tag{E.1}$$

donde y_t es la variable dependiente para la observación t y x_{tj}, $j = 1, 2, \ldots, k$, son las variables independientes. Como de costumbre, β_0 es el intercepto y $\beta_1, \ldots, \beta_k$ denotan los parámetros de las pendientes.

Para cada t, defina un vector $1 \times (k + 1)$, como $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$, y sea $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ el vector $(k + 1) \times 1$ para todos los parámetros. Entonces, (E.1) se puede escribir como

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t,\ t = 1, 2, \ldots, n. \tag{E.2}$$

[Algunos autores prefieren definir $\mathbf{x}_t$ como un vector de columna, en cuyo caso $\mathbf{x}_t$ se remplaza por $\mathbf{x}_t'$ en (E.2). Matemáticamente, tiene más sentido definirlo como vector de fila o renglón.] Se puede escribir (E.2) en notación matricial completa al definir apropiadamente los vectores de datos y matrices. Sea $\mathbf{y}$ el vector $n \times 1$ de observaciones de y: el elemento t-ésimo de $\mathbf{y}$ es y_t. Sea $\mathbf{X}$ el vector $n \times (k + 1)$ de observaciones sobre las variables explicativas. En otras palabras, la fila t-ésima de $\mathbf{X}$ consiste en el vector $\mathbf{x}_t$. Escrito en detalle,

$$\underset{n \times (k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}.$$

Finalmente, sea **u** el vector $n \times 1$ de errores o disturbios inobservables. Entonces, se puede escribir (E.2) para todas las n observaciones en **notación matricial**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{E.3}$$

Recuerde que, debido a que $\mathbf{X}$ es $n \times (k + 1)$ y $\boldsymbol{\beta}$ es $(k + 1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ es $n \times 1$.

La estimación de $\boldsymbol{\beta}$ se produce al minimizar la suma de los residuales cuadrados, como en la sección 3.2. Defina la suma de la función de los residuales cuadrados para todo vector de parámetros $\mathbf{b}$ de $(k + 1) \times 1$ como

$$\text{SRC}(\mathbf{b}) \equiv \sum_{t=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2.$$

El vector $(k + 1) \times 1$ vector de estimadores de mínimos cuadrados ordinarios, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$, minimiza SRC($\mathbf{b}$) para todos los vectores posibles de $(k + 1) \times 1$. Este es un problema de cálculo multivariado. Para que $\hat{\boldsymbol{\beta}}$ minimice la suma de los cuadrados de los residuales, debe resolver la **condición de primer orden**

$$\partial\text{SRC}(\hat{\boldsymbol{\beta}})/\partial\mathbf{b} \equiv 0. \tag{E.4}$$

Puesto que la derivada de $(y_t - \mathbf{x}_t\mathbf{b})^2$ con respecto a $\mathbf{b}$ es el vector $1 \times (k + 1)$, $-2(y_t - \mathbf{x}_t\mathbf{b})\mathbf{x}_t$, (E.4) es equivalente a

$$\sum_{t=1}^{n} \mathbf{x}_t'(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}) \equiv \mathbf{0}. \tag{E.5}$$

(Se ha dividido entre -2 y se toma la transpuesta.) Esta condición de primer orden se puede escribir como

$$\sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0$$
$$\sum_{t=1}^{n} x_{t1}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0$$
$$\vdots$$
$$\sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0,$$

que es idéntica a las condiciones de primer orden en la ecuación (3.13). Se puede escribir esto en forma de matricias para hacerla más fácil de usar. Mediante la fórmula para la multiplicación particionada en el apéndice D, se ve que (E.5) es equivalente a

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{E.6}$$

o

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{E.7}$$

Se puede mostrar que (E.7) siempre tiene al menos una solución. Las soluciones múltiples no son de ayuda, puesto que se busca un conjunto único de estimadores MCO dado el conjunto de datos. Si se asume que la matriz simétrica $\mathbf{X}'\mathbf{X}$ de $(k + 1) \times (k + 1)$ no sea singular, se pueden premultiplicar ambos lados de (E.7) por $(\mathbf{X}'\mathbf{X})^{-1}$ para resolver el estimador $\hat{\boldsymbol{\beta}}$ de MCO:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{E.8}$$

Esta es la fórmula crítica para el análisis matricial del modelo de regresión lineal múltiple. El supuesto de que $\mathbf{X}'\mathbf{X}$ es invertible y equivalente al supuesto de que el rank($\mathbf{X}$) $= (k + 1)$, lo que significa que las columnas de $\mathbf{X}$ deben ser linealmente independientes. Ésta es la versión matricial de RLM.3 en el capítulo 3.

Antes de continuar, se debe advertir algo sobre (E8). Se puede caer en la trampa de simplificar la fórmula de $\hat{\boldsymbol{\beta}}$ de esta manera:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}.$$

El error en este razonamiento es que $\mathbf{X}$ suele no ser una matriz cuadrada, así que no se puede invertir. En otras palabras, no se puede escribir $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ a menos que $n = (k + 1)$, un caso que nunca surge en la práctica.

Los vectores $n \times 1$ de los valores ajustados y los residuales de MCO, están dados por

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},\ \hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ respectivamente.}$$

A partir de (E.6) y de la definición de $\hat{\mathbf{u}}$, se puede ver que la condición de primer orden para $\hat{\boldsymbol{\beta}}$ es la misma que

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \tag{E.9}$$

Como la primera columna de $\mathbf{X}$ consiste por completo en unos, (E.9) implica que los residuos MCO siempre suman cero cuando un intercepto se incluye en la ecuación y la covarianza muestral entre cada variable independiente y los residuales MCO es cero. (Se analizarán estas dos propiedades en el capítulo 3.)

La suma de los residuales cuadrados puede escribirse como

$$\text{SRC} = \sum_{t=1}^{n} \hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{E.10}$$

Todas las propiedades algebraicas del capítulo 3 se pueden determinar mediante álgebra matricial. Por ejemplo, se puede mostrar que la suma total de cuadrados es igual a la suma explicada de cuadrados más la suma de los residuales cuadrados [vea (3.27)]. El uso de matrices no ofrece una prueba más simple que la notación de suma, así que no se ofrece otra derivación.

El enfoque matricial para la regresión múltiple se puede usar como base para la interpretación geométrica de la regresión. Esto implica conceptos matemáticos que son incluso más avanzados que los que se cubrieron en el apéndice D. [Vea Goldberger (1991) o Greene (1997).]

E.2 Propiedades muestrales finitas de MCO

Determinar el valor esperado y la varianza del estimador $\hat{\boldsymbol{\beta}}$ de MCO es fácil con la ayuda del álgebra matricial, pero se debe tener cuidado al establecer los supuestos.

Supuesto E.1 (Linealidad en parámetros)

El modelo se puede escribir como en (E.3), donde $\mathbf{y}$ es un vector $n \times 1$ observado, $\mathbf{X}$ es una matriz observada $n \times (k + 1)$ y $\mathbf{u}$ es un vector $n \times 1$ de errores o perturbaciones inobservables.

Supuesto E.2 **(No colinealidad perfecta)**

La matriz $\mathbf{X}$ tiene rango $(k + 1)$.

Esta es una expresión cuidadosa del supuesto que descarta las dependencias lineales entre las variables explicativas. En el supuesto E.2, $\mathbf{X}'\mathbf{X}$ es no singular, así que $\hat{\boldsymbol{\beta}}$ es único y se puede escribir como en (E.8).

Supuesto E.3 **(Media condicional cero)**

La condicional sobre la matriz $\mathbf{X}$ total, cada error u_t tiene una media cero: $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots, n$.

En la forma vectorial, el supuesto E.3 se puede escribir como

$$\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \tag{E.11}$$

Este supuesto es una implicación de RLM.4 bajo el supuesto de muestreo aleatorio, RLM.2. En las aplicaciones de series de tiempo, el supuesto E.3 impone exogeneidad estricta en las variables explicativas, algo que se analizó a detalle en el capítulo 10. Esto descarta las variables explicativas cuyos valores futuros se correlacionan con u_t; en particular, elimina las variables dependientes rezagadas. Bajo el supuesto E.3, se puede condicionar x_{tj} cuando se calcule el valor esperado de $\hat{\boldsymbol{\beta}}$.

Teorema E.1 **(Insesgamiento de MCO)**

Bajo los supuestos E.1, E.2 y E.3, el estimador de MCO de $\hat{\boldsymbol{\beta}}$ está insesgado para $\boldsymbol{\beta}$.

PRUEBA: Utilice los supuestos E.1 y E2 y álgebra simple para escribir

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},\end{aligned} \tag{E.12}$$

donde se utiliza el hecho de que $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. Al determinar la expectativa condicional a $\mathbf{X}$ se tiene

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta},\end{aligned}$$

como $\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ bajo el supuesto E.3. Este argumento no depende del valor de $\boldsymbol{\beta}$, así que se ha mostrado que $\hat{\boldsymbol{\beta}}$ es insesgada.

Para obtener la forma más simple de la matriz varianza-covarianza de $\hat{\boldsymbol{\beta}}$, se imponen los supuestos de homocedasticidad y correlación no serial.

Supuesto E.4 (Homocedasticidad y correlación no serial)

i) $\text{Var}(u_t|\mathbf{X}) = \sigma^2$, $t = 1, 2, \ldots, n$. ii) $\text{Cov}(u_t, u_s|\mathbf{X}) = 0$, para toda $t \neq s$. De forma matricial se pueden escribir estos dos supuestos como

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}_n, \tag{E.13}$$

donde $\mathbf{I}_n$ es la matriz identidad $n \times n$.

La parte i) del supuesto E.4 es la homocedasticidad: la varianza de u_t no puede depender de ningún elemento de $\mathbf{X}$, y la varianza debe ser constante en todas las observaciones, t. La parte ii) es el supuesto de no correlación serial: los errores no se pueden correlacionar a través de las observaciones. En el muestreo aleatorio y en otros esquemas de muestreo de corte transversal con observaciones independientes, la parte ii) del supuesto E.4 aplica automáticamente. Para aplicaciones de series de tiempo, la parte ii) descarta la correlación en los errores con el paso del tiempo (tanto condicionales sobre $\mathbf{X}$ como incondicionales).

Debido a (E.13), con frecuencia se dice que $\mathbf{u}$ tiene una **matriz escalar de varianza-covarianza** cuando el supuesto E.4 aplica. Ahora se puede determinar la **matriz varianza-covarianza del estimador MCO**.

Teorema E.2 (Matriz varianza-covarianza del estimador MCO)

Bajo los supuestos E.1 a E.4,

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{E.14}$$

PRUEBA: A partir de la última fórmula en la ecuación (E.12), se tiene

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Ahora, se usa el supuesto E.4 para obtener

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

La fórmula (E.4) significa que la varianza de $\hat{\beta}_j$ (condicional a $\mathbf{X}$) se obtiene al multiplicar σ^2 por el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$. Para los coeficientes de pendiente, se tiene una fórmula interpretable en la ecuación (3.51). La ecuación (E.14) también indica cómo obtener la covarianza entre cualesquiera dos estimadores MCO: multiplicar σ^2 por el elemento diagonal apropiado de $(\mathbf{X}'\mathbf{X})^{-1}$. En el capítulo 4 se mostró cómo evitar la búsqueda explícita de covarianzas para obtener los intervalos de confianza y pruebas de hipótesis al escribir como es debido el modelo.

El teorema de Gauss Markov, se puede probar en términos generales.

Teorema E.3 (Teorema de Gauss-Markov)

Con base en los supuestos E.1 a E.4, $\hat{\boldsymbol{\beta}}$ es el mejor estimador lineal insesgado.

PRUEBA: Cualquier otro estimador lineal de $\boldsymbol{\beta}$ se puede escribir como

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \tag{E.15}$$

donde $\mathbf{A}$ es una matriz $n \times (k + 1)$. Con el fin de que $\tilde{\boldsymbol{\beta}}$ sea condicional insesgado en $\mathbf{X}$, $\mathbf{A}$ puede consistir en números y funciones no aleatorios de $\mathbf{X}$. (Por ejemplo, $\mathbf{A}$ no puede ser función de $\mathbf{y}$.) Para ver qué otras restricciones de $\mathbf{A}$ son necesarias, escribir

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \tag{E.16}$$

Entonces,

$$\begin{aligned}
\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathrm{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\
&= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathrm{E}(\mathbf{u}|\mathbf{X}) \text{ porque } \mathbf{A} \text{ es una función de } \mathbf{X} \\
&= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ porque } \mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}.
\end{aligned}$$

Para que $\tilde{\boldsymbol{\beta}}$ sea un estimador insesgado de $\boldsymbol{\beta}$, debe ser verdad que $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ para todo vector $(k + 1) \times 1$ de $\boldsymbol{\beta}$, es decir,

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ para toda } (k + 1) \times 1 \text{ vectores de } \boldsymbol{\beta}. \tag{E.17}$$

Como $\mathbf{A}'\mathbf{X}$ es una matriz $(k + 1) \times (k + 1)$, (E.17) se mantiene si, y sólo si, $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. Las ecuaciones (E.15) y (E.17) caracterizan la clase de estimadores lineales, insesgados de $\boldsymbol{\beta}$.

Después, de (E.16) se tiene

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\mathrm{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

por el supuesto E.4. Por tanto,

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\
&= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ porque } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\
&= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\
&\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A},
\end{aligned}$$

donde $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Debido a que $\mathbf{M}$ es simétrica e idempotente, $\mathbf{A}'\mathbf{M}\mathbf{A}$ es semindefinida positiva para toda matriz $\mathbf{A}$ de $n \times (k + 1)$. Esto establece que el estimador de MCO $\hat{\boldsymbol{\beta}}$ es BLUE. ¿Por qué es importante esto? Sea $\mathbf{c}$ cualquier vector $(k + 1) \times 1$ y considere la combinación lineal $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k$, que es escalar. Los estimadores insesgados de $\mathbf{c}'\boldsymbol{\beta}$ son $\mathbf{c}'\hat{\boldsymbol{\beta}}$ y $\mathbf{c}'\tilde{\boldsymbol{\beta}}$. pero

$$\mathrm{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\beta}|\mathbf{X})]\mathbf{c} \geq 0,$$

porque $[\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\beta}|\mathbf{X})]$ es semidefinida positiva. Por tanto, cuando se usa para estimar alguna combinación lineal de $\boldsymbol{\beta}$, MCO produce la menor variación. En particular, $\mathrm{Var}(\hat{\beta}_j|\mathbf{X}) \leq \mathrm{Var}(\tilde{\beta}_j|\mathbf{X})$ para cualquier otro estimador insesgado lineal de β_j.

El estimador insesgado de la varianza de error σ^2 puede escribirse como

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n - k - 1),$$

que es lo mismo que la ecuación (3.56).

Teorema E.4 (Insesgamiento de $\hat{\sigma}^2$)

Con base en los supuestos E.1 a E.4, $\hat{\sigma}^2$ es insesgado de σ^2: $\mathrm{E}(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ para toda $\sigma^2 > 0$.

PRUEBA: Escriba $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{My} = \mathbf{Mu}$, donde $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, y se desprende la última igualdad debido a que $\mathbf{MX} = \mathbf{0}$. Como $\mathbf{M}$ es simétrica e idempotente,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{Mu} = \mathbf{u}'\mathbf{Mu}.$$

Como $\mathbf{u}'\mathbf{Mu}$ es escalar, es igual a su traza. Por tanto,

$$\begin{aligned}\mathrm{E}(\mathbf{u}'\mathbf{Mu}|\mathbf{X}) &= \mathrm{E}[\mathrm{tr}(\mathbf{u}'\mathbf{Mu})|\mathbf{X}] = \mathrm{E}[\mathrm{tr}(\mathbf{Muu}')|\mathbf{X}] \\ &= \mathrm{tr}[\mathrm{E}(\mathbf{Muu}'|\mathbf{X})] = \mathrm{tr}[\mathbf{M}\mathrm{E}(\mathbf{uu}'|\mathbf{X})] \\ &= \mathrm{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\mathrm{tr}(\mathbf{M}) = \sigma^2(n - k - 1).\end{aligned}$$

La última igualdad se despende de $\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \mathrm{tr}\,(\mathbf{I}_{k+1}) = n - (k + 1) = n - k - 1$. Por tanto,

$$\mathrm{E}(\hat{\sigma}^2|\mathbf{X}) = \mathrm{E}(\mathbf{u}'\mathbf{Mu}|\mathbf{X})/(n - k - 1) = \sigma^2.$$

E.3 Inferencia estadística

Cuando se agrega el supuesto final del modelo lineal clásico, $\hat{\boldsymbol{\beta}}$ tiene una distribución multivariada normal, lo que conduce a las distribuciones t y F para las pruebas estadísticas estándar vistas en el capítulo 4.

Supuesto E.5 (Normalidad de errores)

Condicionadas a $\mathbf{X}$, las u_t son independientes y distribuidas idénticamente como Normal$(0,\sigma^2)$. Asimismo, $\mathbf{u}$ *dada* $\mathbf{X}$ se distribuye como normal multivariada con una media cero y matriz varianza-covarianza $\sigma^2\mathbf{I}_n$: $\mathbf{u} \sim \mathrm{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$.

En el supuesto E.5, cada u_t es independiente de las variables explicativas para toda t. En un escenario de series de tiempo, este es esencialmente el supuesto de exogeneidad estricta.

Teorema E.5 **(Normalidad de $\hat{\boldsymbol{\beta}}$)**

En los supuestos clásicos de modelo lineal E.1 a E.5, $\hat{\boldsymbol{\beta}}$ condicional a $\mathbf{X}$ se distribuye como normal multivariada con una media $\boldsymbol{\beta}$ y matriz varianza-covarianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

El teorema E.5 es la base para la inferencia estadística que implica a $\boldsymbol{\beta}$. De hecho, junto con las propiedades de *ji*-cuadrada, las distribuciones t y F que se resumieron en el apéndice D, se puede usar el teorema E.5 para establecer que las estadísticos t tienen una distribución t según los supuestos E.1 a E.5 (bajo la hipótesis nula) y de forma similar a los estadísticos F. Se ilustra con una prueba para los estadísticos t.

Teorema E.6

Según los supuestos E.1 a E.5,

$$(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j) \sim t_{n-k-1}, j = 0, 1, \ldots, k.$$

PRUEBA: La prueba requiere varios pasos; las siguientes expresiones son condicionales inicialmente a $\mathbf{X}$. Primero, por el teorema E.5, $(\hat{\beta}_j - \beta_j)/\text{de}(\hat{\beta}_j) \sim \text{Normal}(0,1)$ donde $\text{de}(\hat{\beta}_j) = \sigma\sqrt{c_{jj}}$ y c_{jj} es el j-ésimo del elemento diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$. Después, según los supuestos E.1 a E.5, condicional a $\mathbf{X}$,

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}. \quad \textbf{E.18}$$

Esto se desprende porque $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma)$, donde $\mathbf{M}$ es la matriz simétrica idempompotente $n \times n$ definida en el teorema E.4. Pero $\mathbf{u}/\sigma \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ por el supuesto E.5. Se desprende de la propiedad 1 de la distribución ji cuadrada en el apéndice D que $(\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma) \sim \chi^2_{n-k-1}$ (puesto que $\mathbf{M}$ tiene un rango $n - k - 1$).

También se necesita mostrar que $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ son independientes. Pero $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ y $\hat{\sigma}^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Ahora, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$ porque $\mathbf{X}'\mathbf{M} = \mathbf{0}$. Se desprende, de la propiedad 5 de la distribución normal multivariada en el apéndice D, que $\hat{\boldsymbol{\beta}}$ y $\mathbf{Mu}$ son independientes. Debido a que $\hat{\sigma}^2$ es una función de $\mathbf{Mu}$, $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ también son independientes.

$$(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{de}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

que es la razón de una variable aleatoria normal estándar y la raíz cuadrada de una variable aleatoria $\chi^2_{n-k-1}/(n - k - 1)$. Se acaba de mostrar que son independientes, así que, por definición de una variable aleatoria t, $(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j)$ tiene la distribución t_{n-k-1}. Debido a que esta distribución no depende de $\mathbf{X}$, es la distribución incondicional de $(\hat{\beta}_j - \beta_j)/\text{ee}(\hat{\beta}_j)$ también.

Para este teorema, se puede insertar cualquier valor hipotético de β_j y usar el estadístico t para probar las hipótesis como se acostumbra.

Según los supuestos E.1 a E.5, se puede calcular lo que se conoce como el límite inferior *Cramer-Rao* para la matriz varianza-covarianza de estimadores insesgados de $\boldsymbol{\beta}$ (condicional a $\mathbf{X}$, otra vez) [vea Green (1997, capítulo 4)]. Esto puede mostrarse como $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, que es exactamente la matriz varianza-covarianza del estimador MCO. Esto implica que $\hat{\boldsymbol{\beta}}$ es el **estimador insesgado de varianza mínima** de $\boldsymbol{\beta}$ (condicionada a $\mathbf{X}$): $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ es semidefinida positiva para cualquier otro estimador insesgado $\tilde{\boldsymbol{\beta}}$; ya no es necesario que restringir la atención a los estimadores lineales en $\mathbf{y}$.

Es fácil mostrar que el estimador de MCO es de hecho el estimador de máxima verisimilitud de $\boldsymbol{\beta}$ bajo el supuesto E.5. Para cada t, el estimador de y_t dada $\mathbf{X}$ es Normal$(\mathbf{x}_t\boldsymbol{\beta},\sigma^2)$. Debido a que y_t es la condicional independiente en $\mathbf{X}$, la función de probabilidad para la muestra se obtiene a partir del producto de las densidades:

$$\prod_{t=1}^{n}(2\pi\sigma^2)^{-1/2}\exp[-(y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)],$$

donde $\prod$ denota el producto. Maximice esta función con respecto a $\boldsymbol{\beta}$ y σ^2 es lo mismo que maximizar su logaritmo natural:

$$\sum_{t=1}^{n}[-(1/2)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

Para obtener $\hat{\boldsymbol{\beta}}$, esto es lo mismo que minimizar $\sum_{t=1}^{n}(y_t - \mathbf{x}_t\boldsymbol{\beta})^2$ la división entre $2\sigma^2$ no afecta la optimización, que es el problema que resuelven MCO. El estimador de σ^2 que se ha utilizado, $\text{SSR}/(n - k)$, resulta no ser el MLE de Estados Unidos; σ^2; el MLE ss SRC/n, que es un estimador sesgado. Debido a que el estimador insesgado de σ^2 genera estadísticos t y F con distribuciones t y F exactas bajo la nula, siempre se usa en lugar de MLE.

E.4 Algunos análisis asintóticos

El enfoque matricial del modelo de regresión múltiple también puede hacer derivaciones de las propiedades asintóticas más concisas. De hecho, se pueden dar pruebas generales de las afirmaciones hechas en el capítulo 11.

Se empezará por demostrar la consistencia del resultado del teorema 11.1. Recuerde que estos supuestos contienen, como caso especial, los supuestos para el análisis de corte transversal bajo el muestreo aleatorio.

Prueba del teorema 11.1. Como en el problema E. 1 y mediante los supuestos TS.1′, se escribe el estimador de MCO como

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'y_t\right) = \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'(\mathbf{x}_t\boldsymbol{\beta} + u_t)\right) \\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \qquad \textbf{E.19} \\ &= \boldsymbol{\beta} + \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right).\end{aligned}$$

Ahora, por la ley de los números grandes,

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ y } n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t \xrightarrow{p} \mathbf{0}, \qquad \textbf{E.20}$$

donde $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$ es una matriz no singular $(k + 1) \times (k + 1)$ bajo los supuestos TS.2′ y se ha utilizado el hecho de que $\mathrm{E}(\mathbf{x}_t'u_t) = \mathbf{0}$ bajo el supuesto TS.3′. Ahora, se debe usar una versión matricial de la propiedad PLIM.1 en el apéndice C. En otras palabras, porque $\mathbf{A}$ es no singular,

$$\left(n^{-1}\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}. \qquad \text{E.21}$$

[Wooldridge (2002, capítulo 3) contiene un análisis de este tipo de resultados de convergencia.] Se desprende de (E.19), (E.20) y (E.21) que

$$\mathrm{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}.$$

Esto completa la prueba.

Después, se diseña una prueba del resultado de normalidad asintótica en el teorema 11.2.

Prueba del teorema 11.2. De la ecuación (E.19), se puede escribir

$$\begin{aligned}\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(n^{-1}\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n} \mathbf{x}_t'u_t\right) \\ &= \mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n} \mathbf{x}_t'u_t\right) + o_p(1),\end{aligned} \qquad \text{E.22}$$

donde el término “$o_p(1)$” es un término recordatorio que converge en probabilidad cero. Este término es igual a $\left[\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} - \mathbf{A}^{-1}\right]\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$. El término entre corchetes converge en probabilidad cero (por el mismo argumento usado en la prueba del teorema 11.1), mientras que $\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$ está limitada en probabilidad pues converge a una distribución multivariada normal por el teorema del límite central. Un resultado bien conocido en la teoría asintótica es que el producto de tales términos converge en probabilidad en cero. Además, $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ hereda su distribución asintótica de $\mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$. Vea Wooldridge (2002, capítulo 3) para más detalle acerca de los resultados de convergencia utilizados en esta prueba.

Por el teorema del límite central, $n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t$ tiene una distribución normal asintótica con media cero y, digamos $(k + 1) \times (k + 1)$ matriz $\mathbf{B}$ varianza-covarianza. Entonces, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tiene una distribución normal multivariada asintótica con media cero y matriz varianza-covarianza $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. Ahora se muestra que, según los supuestos TS.4′ y TS.5′, $\mathbf{B} = \sigma^2\mathbf{A}$. (La expresión general es útil pues en ella se basa la heterocedasticidad robusta y los errores estándar robustos de correlación serial para MCO, del tipo analizado en el capítulo 12.) Primero, bajo el supuesto TS.5′, $\mathbf{x}_t'u_t$ y $\mathbf{x}_s'u_s$ no se correlacionan para $t \neq s$. ¿Por qué? Suponga que $s < t$ concretamente. Entonces, por la ley de las expectativas iteradas, $\mathrm{E}(\mathbf{x}_t'u_tu_s\mathbf{x}_s) = \mathrm{E}[\mathrm{E}(u_tu_s\mathbf{x}_t'\mathbf{x}_s)|\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[\mathrm{E}(u_tu_s|\mathbf{x}_t'\mathbf{x}_s)\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[0 \cdot \mathbf{x}_t'\mathbf{x}_s] = \mathbf{0}$. Las covarianzas cero implican que la varianza de la suma es la suma de las varianzas. Pero $\mathrm{Var}(\mathbf{x}_t'u_t) = \mathrm{E}(\mathbf{x}_t'u_tu_t\mathbf{x}_t) = \mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$. Por la ley de las expectativas iteradas, $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t|\mathbf{x}_t)] = \mathrm{E}[\mathrm{E}(u_t^2|\mathbf{x}_t)\mathbf{x}_t'\mathbf{x}_t] = \mathrm{E}(\sigma^2\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathbf{A}$, donde se usa $\mathrm{E}(u_t^2|\mathbf{x}_t) = \sigma^2$ bajo los supuestos TS.3′ y TS.4′. Esto muestra que $\mathbf{B} = \sigma^2\mathbf{A}$, correlaciona, y así sucesivamente, según los supuestos TS.1′ y TS.5′, se tiene

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{a}{\sim} \text{Normal}\,(\mathbf{0},\sigma^2\mathbf{A}^{-1}). \qquad \text{E.23}$$

Esto completa la prueba.

De la ecuación (E.23), se trata a $\hat{\boldsymbol{\beta}}$ como si tuviera una distribución aproximadamente normal con media $\boldsymbol{\beta}$ y matriz $\sigma^2\mathbf{A}^{-1}/n$ varianza-covarianza. La división entre el tamaño muestral, n, se espera aquí: la aproximación a la matriz varianza-covarianza de $\hat{\boldsymbol{\beta}}$ se reduce a cero a la tasa $1/n$. Cuando se remplaza σ^2 con su estimador consistente, $\hat{\sigma}^2 = \text{SRC}/(n - k - 1)$, y se remplaza $\mathbf{A}$ con su estimador consistente, $n^{-1}\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t = \mathbf{X}'\mathbf{X}/n$, se obtiene un estimador para la varianza asintótica de $\hat{\boldsymbol{\beta}}$:

$$\widehat{\mathbf{A}var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{E.24}$$

Observe cómo las dos divisiones entre n se cancelan, y el lado derecho de (E.24) es la forma usual de estimar la matriz de varianza del estimador MCO bajo los supuestos de Gauss-Markov. Para resumir, se ha mostrado que, bajo los supuestos TS.1′ a TS.5′ que contienen RLM.1 a RLM.5 como casos especiales, los errores estándar y estadísticos t usuales son asintóticamente válidos. Es perfectamente legítimo utilizar la distribución t acostumbrada para obtener los valores críticos y los valores-p para probar una sola hipótesis. Es interesante ver que en la configuración general del capítulo 11, asumir la normalidad de los errores, —digamos, u_t dadas $\mathbf{x}_t$, u_{t-1}, $\mathbf{x}_{t-1}$, ..., u_1, $\mathbf{x}_1$ se distribuye como Normal$(0,\sigma^2)$— no necesariamente ayuda, puesto que el estadístico t generalmente no tendría los estadísticos t bajo este tipo de supuesto de normalidad. Cuando no se asume una exogeneidad estricta de las variables explicativas, los resultados distribucionales exactos son difíciles, si no imposibles de obtener.

Si modificamos el argumento anterior, se puede determinar una matriz de heterocedasticidad robusta de varianza-covarianza. La clave es que se debe estimar $\text{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$ por separado porque su matriz ya no es igual a $\sigma^2\text{E}(\mathbf{x}_t'\mathbf{x}_t)$. Pero si las $\hat{u}_t$ son los residuales MCO, un estimador consistente es

$$(n - k - 1)^{-1}\sum_{t=1}^{n} \hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t, \tag{E.25}$$

donde la división entre $n - k - 1$ y no entre n es un ajuste de grados de libertad que por lo general ayuda a las propiedades muestrales finitas del estimador. Cuando se usa la expresión en la ecuación (E.25), se obtiene

$$\widehat{\mathbf{A}var}(\hat{\boldsymbol{\beta}}) = [n/(n - k)](\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{t=1}^{n} \hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{E.26}$$

Las raíces cuadradas de los elementos diagonales de esta matriz son los mismos errores estándar de heterocedasticidad robusta que se obtuvieron en la sección 8.2 para el caso de corte transversal puro. Una extensión matricial de los errores estándar de correlación serial (y heterocedasticidad) robusta, se obtuvo en la sección 12.5 y también se puede obtener, pero la matriz que debe remplazar a (E.25) es complicada debido a la correlación serial. Vea, por ejemplo, Hamilton (1994, sección 10.5).

Estadístico de Wald para probar hipótesis múltiples

Argumentos similares se pueden utilizar para obtener la distribución asintótica del **estadístico de Wald** para probar hipótesis múltiples. Sea $\mathbf{R}$ una matriz $q \times (k + 1)$ con $q \leq (k + 1)$. Suponga que las restricciones q en el vector $(k + 1) \times 1$ de parámetros, $\boldsymbol{\beta}$, se puede expresar como

H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, donde $\mathbf{r}$ es un vector $q \times 1$ de constantes conocidas. Bajo los supuestos TS.1′ a TS.5′, se puede demostrar que, bajo H_0,

$$[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \overset{a}{\sim} \chi^2_q, \tag{E.27}$$

donde $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$, como en las pruebas de los teoremas 11.1 y 11.2. La lógica de la ecuación (E.25) es simple. Debido a que $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ por lo general se distribuye como Normal$(\mathbf{0},\sigma^2\mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\,\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ es aproximadamente Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ por la propiedad 3 de la distribución normal multivariada en el apéndice D. Bajo H_0, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, así que $\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ bajo H_0. Por la propiedad 3 de la distribución ji-cuadrada, $z'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}z \sim \chi^2_q$ si $z \sim$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. Para obtener el resultado final formalmente, se necesita usar una versión asintótica de esta propiedad, que se puede encontrar en Wooldridge (2002, capítulo 3).

Dado el resultado en (E.25), se obtiene un estadístico computable al remplazar $\mathbf{A}$ y σ^2 con sus estimadores consistentes; hacerlo no cambia la distribución asintótica. El resultado es el llamado estadístico Wald, que, después de cancelar los tamaños muestrales y hacer unos cuantos cálculos algebraicos, se puede escribir como

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \tag{E.28}$$

Bajo H_0, $W \overset{a}{\sim} \chi^2_q$, donde se recuerda que q es el número de restricciones que se están probando. Si $\hat{\sigma}^2 = \mathrm{SRC}/(n - k - 1)$, se puede mostrar que W/q es exactamente el estadístico F que se obtuvo en el capítulo 4 para probar restricciones lineales múltiples. [Vea, por ejemplo, Greene (1997, capítulo 7).] Por tanto, bajo los supuestos del modelo lineal clásico TS.1 a TS.6 en el capítulo 10, W/q tiene una distribución $F_{q,n-k-1}$ exacta. Bajo los supuestos TS.1′ a TS.5′, sólo se tiene el resultado asintótico en (E.26). Sin embargo, es apropiado y común tratar al estadístico F como poseedor de una distribución $F_{q,n-k-1}$ aproximada.

Un estadístico de Wald robusto para la heterocedasticidad de forma desconocida se obtiene mediante la matriz en (E.26) en lugar de $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, y de forma similar para un estadístico de prueba robusto para la hetorocedasticidad y la correlación serial. Las versiones robustas del estadístico de prueba no se pueden calcular mediante las sumas de los residuales cuadrados o R-cuadradas de las regresiones restringidas e irrestrictas.

RESUMEN

Este apéndice le ha proporcionado un breve análisis del modelo de regresión lineal mediante notación matricial. Este material se incluye para las clases más avanzadas que usen álgebra matricial, pero no se necesita leerlo. En efecto, este apéndice prueba algunos de los resultados que se pueden expresar sin prueba, probarse sólo en casos especiales o a través de un método más complicado. Otros temas, como las propiedades asintóticas, la estimación de las variables instrumentales y los modelos de panel de datos, pueden dar tratamientos concisos con uso de matrices. Mayores detalles se pueden encontrar en libros avanzados de econometría, como Davidson y MacKinnon (1993]), Greene (1997) y Wooldridge (2002).

TÉRMINOS CLAVE

Condición de primer orden
Estadístico de Wald
Estimador insesgado de varianza mínima
Matriz de varianza-covarianza del estimador MCO
Matriz escalar de varianza-covarianza
Notación matricial

PROBLEMAS

E.1 Sea $\mathbf{x}_t$ el vector $1 \times (k+1)$ de variables explicativas para cada observación t. Muestre que el estimador MCO de $\hat{\boldsymbol{\beta}}$ puede escribirse como

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1} \left(\sum_{t=1}^{n} \mathbf{x}_t'y_t\right).$$

Al dividir cada suma entre n se muestra que $\hat{\boldsymbol{\beta}}$ es una función de promedios muestrales.

E.2 Sea $\hat{\boldsymbol{\beta}}$ el vector $(k+1) \times 1$ de estimadores MCO.

i) Muestre que para cualquier vector $(k+1) \times 1$ de $\mathbf{b}$, se puede escribir la suma de los residuales cuadrados como

$$\text{SRC}(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

{*Sugerencia:* escriba $(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = [\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ y use el hecho de que $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.}

ii) Explique por qué la expresión de SRC($\mathbf{b}$) en la parte i) prueba que $\hat{\boldsymbol{\beta}}$ sólo minimiza SRC($\mathbf{b}$) a través de todos los valores posibles de $\mathbf{b}$, suponiendo que $\mathbf{X}$ tenga un rango $k+1$.

E.3 Sea $\hat{\boldsymbol{\beta}}$ el estimador MCO de la regresión de $\mathbf{y}$ sobre $\mathbf{X}$. Sea $\mathbf{A}$ una matriz no singular $(k+1) \times (k+1)$ y defina $z_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Por tanto, z_t es $1 \times (k+1)$ y es una combinación lineal no singular de $\mathbf{x}_t$. Sea $\mathbf{Z}$ la matriz $n \times (k+1)$ con filas z_t. Sea $\tilde{\boldsymbol{\beta}}$ el estimador MCO de una regresión de $\mathbf{y}$ en $\mathbf{Z}$.

i) Muestre que $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.

ii) Sea $\hat{y}_t$ los valores ajustados de la regresión original y sea $\tilde{y}_t$ los valores ajustados de regresar $\mathbf{y}$ sobre $\mathbf{Z}$. Muestre que $\tilde{y}_t = \hat{y}_t$, para toda $t = 1, 2, \ldots, n$. ¿Cómo se comparan los residuales de las dos regresiones?

iii) Muestre que la matriz de varianza estimada para $\tilde{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1\prime}$, donde $\hat{\sigma}^2$ es el estimado de varianza usual de regresar $\mathbf{y}$ sobre $\mathbf{X}$.

iv) Sea $\hat{\beta}_j$ los estimadores MCO de regresar y_t sobre $1, x_{t1}, \ldots, x_{tk}$, y sea $\tilde{\beta}_j$ los estimadores MCO de la regresión de y_t sobre $1, a_1x_{t1}, \ldots, a_kx_{tk}$, donde $a_j \neq 0$, $j = 1, \ldots, k$. Use los resultados de la parte i) para encontrar la relación entre $\tilde{\beta}_j$ y $\hat{\beta}_j$.

v) Suponga la configuración de la parte iv), use la parte iii) para mostrar que $\text{ee}(\tilde{\beta}_j) = \text{ee}(\hat{\beta}_j)/|a_j|$.

vi) Suponga la configuración de la parte iv), muestre que los valores absolutos del estadístico t para $\tilde{\beta}_j$ y $\hat{\beta}_j$ son idénticos.

E.4 Asuma que el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisface los supuestos de Gauss-Markov, sea $\mathbf{G}$ una matriz no aleatoria, no singular $(k+1) \times (k+1)$ y defina $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, de manera que $\boldsymbol{\delta}$ sea también un vector $(k+1) \times 1$. Sea $\hat{\boldsymbol{\beta}}$ el vector $(k+1) \times 1$ de estimadores MCO y defina $\hat{\boldsymbol{\delta}} = \mathbf{G}\hat{\boldsymbol{\beta}}$ como el estimador MCO de $\boldsymbol{\delta}$.

i) Muestre que $\mathrm{E}(\hat{\boldsymbol{\delta}}|\mathbf{X}) = \boldsymbol{\delta}$.

ii) Calcule $\mathrm{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ en términos de σ^2, $\mathbf{X}$ y $\mathbf{G}$.

iii) Utilice el problema E.3 para verificar que $\hat{\boldsymbol{\delta}}$ y el estimado apropiado de $\mathrm{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ se obtienen de la regresión de $\mathbf{y}$ sobre $\mathbf{XG}^{-1}$.

iv) Ahora sea $\mathbf{c}$ un vector $(k + 1) \times 1$ con al menos una entrada diferente de cero. En concreto, asuma que $c_k \neq 0$. Defina $\theta = \mathbf{c}'\boldsymbol{\beta}$, de manera que θ sea una escalar. Defina $\delta_j = \boldsymbol{\beta}_j$, $j = 0, 1, ..., k - 1$ y $\delta_k = \theta$. Muestre cómo definir una matriz $(k + 1) \times (k + 1)$ no singular $\mathbf{G}$ de manera que $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$. (*Sugerencia:* cada una de las primeras filas k de $\mathbf{G}$ deben contener k ceros y un uno. ¿Cuál es la última fila?)

v) Muestre que para la elección $\mathbf{G}$ en la parte iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ 0 & 0 & . & . & . & 1 & 0 \\ -c_0/c_k & -c_1/c_k & . & . & . & -c_{k-1}/c_k & 1/c_k \end{bmatrix}.$$

Use esta expresión para $\mathbf{G}^{-1}$ y la parte iii) para concluir que $\hat{\theta}$ y su error estándar se obtuvieron como el coeficiente en x_{tk}/c_k en la regresión de

$$y_t \text{ en } [1 - (c_0/c_k)x_{tk}],\ [x_{t1} - (c_1/c_k)x_{tk}],\ ...,\ [x_{t,k-1} - (c_{k-1}/c_k)x_{tk}],\ x_{tk}/c_k,\ t = 1, ..., n.$$

Esta regresión es exactamente la obtenida al escribir β_k en términos de θ y $\beta_0, \beta_1, ..., \beta_{k-1}$, al insertar el resultado en el modelo original y reordenar. Por tanto, se puede justificar formalmente el truco que se usa a través del libro para obtener el error estándar de una combinación lineal de parámetros.

E.5 Suponga que el modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisface los supuestos de Gauss-Markov y sea $\hat{\boldsymbol{\beta}}$ el estimador MCO de $\boldsymbol{\beta}$. Sea $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ una función matricial $n \times (k + 1)$ de $\mathbf{X}$ y suponga que $\mathbf{Z}'\mathbf{X}$ [una matriz $(k + 1) \times (k + 1)$] no es singular. Defina un nuevo estimador de $\boldsymbol{\beta}$ por $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

i) Muestre que $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, de manera que $\tilde{\boldsymbol{\beta}}$ también sea condicional insesgada en $\mathbf{X}$.

ii) Calcule $\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Asegúrese de que sea una matriz $(k + 1) \times (k + 1)$ simétrica que depende de $\mathbf{Z}$, $\mathbf{X}$ y σ^2.

iii) ¿Qué estimador prefiere, $\hat{\boldsymbol{\beta}}$ o $\tilde{\boldsymbol{\beta}}$? Explique.

Apéndice F

Respuestas a las preguntas del capítulo

Capítulo 2

Pregunta 2.1: Cuando la habilidad, motivación, edad y otros factores del estudiante en *u* no se relacionan con la asistencia, (2.6) sería válida. Pero parece que es improbable que éste sea el caso.

Pregunta 2.2: Aproximadamente \$11.05. Para verlo, de los salarios promedio medidos en dólares de 1976 y 2003, se puede obtener el factor causante de la deflación CPI como 19.06/5.90 $\approx$ 3.23. Cuando se multiplica 3.42 por 3.23, se obtiene alrededor de 11.05.

Pregunta 2.3: 54.65, como se puede ver al insertar $shareA = 60$ en la ecuación (2.28). Esto no es razonable: si el candidato A ejerce 60% del total del dinero gastado, se predice que él o ella recibirá casi 55% de los votos.

Pregunta 2.4: La ecuación será $\widehat{salaryhun} = 9{,}631.91 + 185.01\ roe$, como se puede ver fácilmente al multiplicar la ecuación (2.39) por 10.

Pregunta 2.5: La ecuación (2.58) se puede escribir como $\mathrm{Var}(\hat{\beta}_0) = (\sigma^2 n^{-1})\left(\sum_{i=1}^{n} x_i^2\right)\Big/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$, donde el término que multiplica a $\sigma^2 n^{-1}$ es mayor o igual que uno, pero es igual a uno si, y sólo si, $\bar{x} = 0$. En este caso, la varianza es la más pequeña posible: $\mathrm{Var}(\hat{\beta}_0) = \sigma^2/n$.

Capítulo 3

Pregunta 3.1: Tan sólo unos factores incluyen la edad y la distribución por género, el tamaño de la fuerza policiaca (o, más generalmente, los recursos dedicados a combatir la delincuencia), población y factores históricos generales. Estos factores, sin lugar a dudas, pueden correlacionarse con *prbcon* y *avgsen*, lo que significa que (3.5) no sería válida. Por ejemplo, el tamaño de la fuerza policiaca posiblemente está correlacionado tanto con *prbcon* como con *avgsen*, puesto que algunas ciudades se esfuerzan más en la prevención del delito y el cumplimiento de las leyes. Es necesario tratar de incluir en la ecuación tantos factores de éstos como sea posible.

Pregunta 3.2: Se usa la tercera propiedad de los MCO relativa a los valores y residuales predichos: cuando se insertan los valores promedio de todas las variables independientes en la línea de regresión MCO, se obtiene el valor promedio de la variable dependiente. Así que $\overline{colGPA} = 1.29 + .453\,\overline{hsGPA} + .0094\,\overline{ACT} = 1.29 + .453(3.4) + .0094(24.2) \approx 3.06$. Se puede verificar el promedio de *colGPA* en GPA1.RAW para verificar esto al segundo decimal.

Pregunta 3.3: No. La variable *shareA* no es una función lineal exacta de *expendA* y *expendB*, aunque es una función no lineal exacta *nonlinear*: $shareA = 100 \cdot [expendA/(expendA + expendB)]$. Por tanto, es legítimo tener *expendA*, *expendB*, y *shareA* como variables explicativas.

Pregunta 3.4: Como se analizó en la sección 3.4, si lo que interesa es el efecto de x_1 en y, la correlación entre las demás variables explicativas (x_2, x_3, etcétera) no afecta a $\mathrm{Var}(\hat{\beta}_1)$. Estas variables se incluyen como controles y no debe preocupar la colinealidad entre las variables de control. Por supuesto, se controlan principalmente porque se piensa que están correlacionadas con la asistencia, pero esto es necesario para llevar a cabo un análisis *ceteris paribus*.

Capítulo 4

Pregunta 4.1: En estos supuestos, los supuestos de Gauss-Markov se satisfacen: u es independiente de las variables explicativas, así que $\mathrm{E}(u|x_1, \ldots, x_k) = \mathrm{E}(u)$ y $\mathrm{Var}(u|x_1,\ldots, x_k) = \mathrm{Var}(u)$. Además, es fácil ver que $\mathrm{E}(u) = 0$. Por tanto RLM.4 y RLM.5 son válidas. Los supuestos del modelo lineal clásico no se satisfacen, pues u no se distribuye de formal normal (lo cual es una violación a RLM.6).

Pregunta 4.2: H_0: $\beta_1 = 0$, H_1: $\beta_1 < 0$.

Pregunta 4.3: Como $\hat{\beta}_1 = .56 > 0$ y se está probando contra H_1: $\beta_1 > 0$, el valor-*p* unilateral es la mitad del bilateral, o .043.

Pregunta 4.4: H_0: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. $k = 8$ y $q = 4$. La versión restringida del modelo es

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + u.$$

Pregunta 4.5: El estadístico F para probar la exclusión de *ACT* es $[(.291 - .183)/(1 - .291)](680 - 3) \approx 103.13$. Por tanto, el valor absoluto del estadístico t es aproximadamente 10.16. El estadístico t en *ACT* es negativo, debido a que $\hat{\beta}_{ACT}$ es negativo, de manera que $t_{ACT} = -10.16$.

Pregunta 4.6: No mucho. La prueba F para la significancia conjunta de *droprate* y *gradrate* se calcula fácilmente de las R-cuadradas en la tabla: $F = [(.361 - .353)/(1 - .361)](402/2) \approx 2.52$. El valor crítico de 10% se obtiene de la tabla G.3a como 2.30, mientras que el valor crítico de 5% de la tabla G.3b es 3. El valor-*p* es de aproximadamente .082. Por tanto, *droprate* y *gradrate* son significativamente conjuntas al nivel de 10%, pero no al nivel de 5%. En cualquier caso, controlar estas variables tiene un efecto menor en el coeficiente *b/s*.

Capítulo 5

Pregunta 5.1: Esto requiere algunos supuestos. Parece razonable asumir que $\beta_2 > 0$ (*score* depende positivamente de *priGPA*) y $\mathrm{Cov}(skipped, priGPA) < 0$ (*skipped* y *priGPA* están correlacionadas negativamente). Esto significa que $\beta_2\delta_1 < 0$, lo que significa que $\tilde{\beta}_1 < \beta_1$. Debido a que se piensa que β_1 es negativo (o al menos no positivo), es probable que una regresión simple sobrestime la importancia de no asistir a clases.

Pregunta 5.2: $\hat{\beta}_j \pm 1.96\text{ee}(\hat{\beta}_j)$ es el intervalo de confianza asintótico a 95%. O, se puede remplazar 1.96 por 2.

Capítulo 6

Pregunta 6.1: Debido a que *fincdol* = 1,000·*faminc*, el coeficiente en *fincdol* será el coeficiente de *faminc* dividido entre 1,000, o .0927/1,000 = .0000927. El error estándar también desciende en un factor de 1,000, así que el estadístico t no cambia, ni cualquiera de los demás estadísticos OLS. Para una fácil lectura, es mejor medir el ingreso familiar en miles de dólares.

Pregunta 6.2: Se puede hacer esto de forma general. La ecuación es

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \ldots,$$

donde x_2 es una porción más que un porcentaje. Entonces, *ceteris paribus*, $\Delta\log(y) = \beta_2 \Delta x_2$, $100\cdot\Delta\log(y) = \beta_2(100\cdot\Delta x_2)$, o $\%\Delta y \approx \beta_2(100\cdot\Delta x_2)$. Ahora, debido a que Δx_2 es el cambio en la proporción, $100\cdot\Delta x_2$ es un cambio de punto porcentual. En particular, si $\Delta x_2 = .01$. Entonces $100\cdot\Delta x_2 = 1$, lo cual corresponde a un cambio de punto porcentual. Pero entonces β_2 es el cambio porcentual en y cuando $100\cdot\Delta x_2 = 1$.

Pregunta 6.3: El nuevo modelo sería $stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA\cdot atndrte + \beta_7 ACT\cdot atndrte + u$. Por tanto, el efecto parcial de *atndrte* en *stndfnl* es $\beta_1 + \beta_6 priGPA + \beta_7 ACT$. Esto es lo que multiplicamos por $\Delta atndrte$ para obtener el cambio *ceteris paribus* en *stndfnl*.

Pregunta 6.4: De la ecuación (6.21), $\bar{R}^2 = 1 - \hat{\sigma}^2/[\text{SST}/(n-1)]$. Para una muestra y una variable determinada dadas, $\text{SST}/(n-1)$ es constante. Cuando se usan diferentes conjuntos de variables explicativas, sólo $\hat{\sigma}^2$ cambia. Conforme As $\hat{\sigma}^2$ disminuye, $\bar{R}^2$ aumenta. Si se hace que $\hat{\sigma}$, y por tanto $\hat{\sigma}^2$, sea lo más pequeña posible, se está haciendo que $\bar{R}^2$ sea lo más grande posible.

Pregunta 6.5: Una posibilidad es recabar datos acerca de las ganancias anuales para una muestra de actores, junto con la rentabilidad de las películas en las que cada uno aparece. En un análisis de regresión simple podríamos relacionar ganancias con rentabilidad. Pero quizá debemos controlar otros factores que pueden afectar el sueldo, tales como edad, género y el tipo de películas en que los actores participan. Los métodos para incluir factores cualitativos en los modelos de regresión se consideran en el capítulo 7.

Capítulo 7

Pregunta 7.1: No, porque no sería claro cuando *party* es uno y cuando es cero. Un nombre mejor sería algo como *Dem*, que es uno para los candidatos demócratas y cero para los republicanos. O *Rep*, que es uno para los republicanos y cero para los demócratas.

Pregunta 7.2: Con *outfield* como el grupo base, se incluirían las variables ficticias *frstbase*, *scndbase*, *thrdbase*, *shrtstop* y *catcher*.

Pregunta 7.3: La nula en este caso es H_0: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, así que hay cuatro restricciones. Como de costumbre, se usaría una prueba F (donde $q = 4$ y k dependen del número de las demás variables explicativas).

Pregunta 7.4: Debido a que *tenure* aparece como cuadrática, se deben permitir cuadráticas separadas para hombre y mujer. Es decir, se agregarían las variables explicativas $female \cdot tenure$ y $female \cdot tenure^2$.

Pregunta 7.5: Se inserta $pcnv = 0$, $avgsen = 0$, $tottime = 0$, $ptime86 = 0$, $qemp86 = 4$, $black = 1$, en $hispan = 0$ dentro de (7.31): $\widehat{arr86} = .380 - .038(4) + .170 = .398$, o casi .4. Es difícil saber si esto es "razonable". Para alguien sin condenas anteriores y que hubiera trabajado todo el año, esta estimación parece alta, pero recuerde que la población consiste en hombres que ya fueron arrestados al menos una vez antes de 1986.

Capítulo 8

Pregunta 8.1: Esta expresión es claramente falsa. Por ejemplo, en la ecuación (8.7), el error estándar usual para *black* is .147, mientras que el error estándar de heteroscedasticidad-robusta es .118.

Pregunta 8.2: La prueba F se obtendría al regresar $\hat{u}^2$ sobre *marrmale*, *marrfem* y *singfem* (*singmale* es el grupo base). Con $n = 526$ y tres variables independientes en esta regresión *df* son 3 y 522.

Pregunta 8.3: Sin lugar a dudas, el resultado de la prueba estadística sugiere algunas causas de interés. Un estadístico t de 2.96 es muy significativo e implica que existe heteroscedasticidad en la ecuación de la riqueza. Como cuestión práctica, se sabe que el error estándar WLS, .063, está sustancialmente por debajo del error estándar de heteroscedasticidad-robusta para MCO, .104, y por tanto la heteroscedasticidad parece ser prácticamente importante. (Además, el error estándar MCO no robusto es .061, que es demasiado optimista. Por consiguiente, aun cuando simplemente se ajuste al error estándar MCO para la heteroscedasticidad de forma desconocida, existen implicaciones no triviales.)

Pregunta 8.4: El valor crítico de 1% en la distribución F con $(2, \infty)$ df es 4.61. Un estadístico F de 11.15 está muy por encima del valor crítico de 1%, y por ello se rechaza de forma contundente la hipótesis nula de que los errores transformados, $u_i/\sqrt{h_i}$, son homoscedásticos. (De hecho, el valor-p es menor que .00002, que se obtiene de la distribución $F_{2,804}$.) Esto significa que el modelo para $\mathrm{Var}(u|\mathbf{x})$ es inadecuado para eliminar por completo la heteroscedasticidad en u.

Capítulo 9

Pregunta 9.1: Estas son variables binarias y elevarlas al cuadrado no tiene efecto: $black^2 = black$, y $hispan^2 = hispan$.

Pregunta 9.2: Cuando $educ \cdot IQ$ está en la ecuación, el coeficiente en *educ*, digamos, β_1, mide el efecto de *educ* en $\log(wage)$ cuando $IQ = 0$. (El efecto parcial de la educación es $\beta_1 + \beta_9 IQ$.) No hay uno en la población de interés con un IQ cercano a cero. En IQ promedio de la población, que es 100, el rendimiento estimado de la educación de la columna 3) es $.018 + .00034(100) = .052$, que es casi lo que se obtiene como el coeficiente en *educ* en la columna 2).

Pregunta 9.3: No. Si *educ** es un entero, que significa que alguien sin educación pasó el grado anterior, el error de medición en cero. Si *educ** no es un entero, $educ < educ^*$, de manera que el error de medición es negativo. A un mínimo, e_1 no puede tener media cero, y e_1 y *educ** probablemente estén correlacionados.

Pregunta 9.4: Una decisión de un candidato de no competir puede estar relacionada de forma sistemática con lo que espera hacer en la elección. Por tanto, sólo se tiene una muestra de los candidatos más fuertes, en promedio, que todos los posibles candidatos que participen. Esto genera un problema de selección muestral si la población de interés incluye a todos los candidatos. Si sólo interesan los efectos de los gastos de campaña sobre los resultados de la elección para los candidatos que buscan reelegirse, no hay problema de selección muestral.

Capítulo 10

Pregunta 10.1: El multiplicador de impacto es .48, mientras que el multiplicador de largo plazo es $.48 - .15 + .32 = .65$.

Pregunta 10.2: Las variables explicativas son $x_{t1} = z_t$ y $x_{t2} = z_{t-1}$. La ausencia de colinealidad perfecta significa que estas no pueden ser constantes y que no puede haber una relación lineal exacta entre ellas en la muestra. Esto descarta la posibilidad de que todas las $z_1, \ldots, z_n$ asuman el mismo valor o que las $z_0, z_1, \ldots, z_{n-1}$ asuman el mismo valor. Pero elimina también otros patrones. Por ejemplo, si $z_t = a + bt$ para las constantes a y b, entonces $z_{t-1} = a + b(t - 1) = (a + bt) - b = z_t - b$, que es una función lineal perfecta de z_t.

Pregunta 10.3: Si $\{z_t\}$ se mueve lentamente con el paso del tiempo, como el caso de los niveles o logaritmos en muchas series de tiempo económicas, entonces z_t y z_{t-1} pueden estar altamente correlacionadas. Por ejemplo, la correlación entre $unem_t$ y $unem_{t-1}$ en PHILLIPS.RAW es .75.

Pregunta 10.4: No, porque una tendencia lineal con $\alpha_1 < 0$ se vuelve cada vez más negativa conforme t aumenta. Dado que *gfr* no puede ser negativa, una tendencia temporal lineal con un coeficiente de tendencia negativo no representaría *gfr* en todos los periodos futuros.

Pregunta 10.5: El intercepto para marzo es $\beta_0 + \delta_2$. Las variables estacionales ficticias son estrictamente exógenas pues siguen un patrón determinista. Por ejemplo, los meses no cambian con base en si lo hacen las variables explicativas o la variable dependiente.

Capítulo 11

Pregunta 11.1: i) No, porque $E(y_t) = \delta_0 + \delta_1 t$ depende de t. ii) Sí, porque $y_t - E(y_t) = e_t$ es una secuencia i.i.d.

Pregunta 11.2: Se inserta $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$ en $inf_t - inf_t^e = \beta_1(unem_t - \mu_0) + e_t$ y se reacomoda: $inf_t - (1/2)(inf_{t-1} + inf_{t-2}) = \beta_0 + \beta_1 unem_t + e_t$, donde $\beta_0 = -\beta_1\mu_0$, como antes. Por tanto, se regresaría y_t sobre $unem_t$, donde $y_t = inf_t - (1/2)(inf_{t-1} + inf_{t-2})$. Observe que se pierden las primeras dos observaciones al construir y_t.

Pregunta 11.3: No, porque u_t y u_{t-1} están correlacionadas. En particular, $\mathrm{Cov}(u_t, u_{t-1}) = E[(e_t + \alpha_1 e_{t-1})(e_{t-1} + \alpha_1 e_{t-2})] = \alpha_1 E(e_{t-1}^2) = \alpha_1\sigma_e^2 \neq 0$ si $\alpha_1 \neq 0$. Si los errores se correlacionan serialmente, el modelo no puede ser dinámicamente completo.

Capítulo 12

Pregunta 12.1: Se usa la ecuación (12.4). Ahora sólo están correlacionados los términos adyacentes. En particular, la covarianza entre $x_t u_t$ y $x_{t+1}u_{t+1}$ es $x_t x_{t+1}\text{Cov}(u_t,u_{t+1}) = x_t x_{t+1}\alpha\sigma_e^2$. Por consiguiente, la fórmula es

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \text{SST}_x^{-2}\left(\sum_{t=1}^{n} x_t^2\text{Var}(u_t) + 2\sum_{t=1}^{n-1} x_t x_{t+1}\text{E}(u_t u_{t+1})\right)\\ &= \sigma^2/\text{SST}_x + (2/\text{SST}_x^2)\sum_{t=1}^{n-1}\alpha\sigma_e^2 x_t x_{t+1}\\ &= \sigma^2/\text{SST}_x + \alpha\sigma_e^2(2/\text{SST}_x^2)\sum_{t=1}^{n-1} x_t x_{t+1},\end{aligned}$$

donde $\sigma^2 = \text{Var}(u_t) = \sigma_e^2 + \alpha_1^2\sigma_e^2 = \sigma_e^2(1 + \alpha_1^2)$. A menos que x_t y x_{t+1} no se correlacionen en la muestra, el término segundo es diferente de cero siempre que $\alpha \neq 0$. Observe que si x_t y x_{t+1} se correlacionan positivamente y $\alpha < 0$, la verdadera varianza es en realidad *menor* que la varianza usual. Cuando la ecuación está en niveles (contrario a estar diferenciada), el caso típico es $\alpha > 0$, con una correlación positiva entre x_t y x_{t+1}.

Pregunta 12.2: $\hat{\rho} \pm 1.96\text{ee}(\hat{\rho})$, mide $\text{ee}(\hat{\rho})$ es el error estándar reportado en la regresión. O, se puede usar el error estándar de heteroscedasticidad robusta. Mostrar que esto es asintóticamente válido es complicado debido a que los residuales MCO dependen de $\hat{\beta}_j$, pero se puede hacer.

Pregunta 12.3: El modelo que se tiene en mente es $u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + e_t$, y se quiere probar H_0: $\rho_1 = 0$, $\rho_4 = 0$ contra la alternativa de que H_0 sea falsa. Se ejecutaría la regresión de $\hat{u}_t$ en $\hat{u}_{t-1}$ y $\hat{u}_{t-4}$ para obtener el estadístico F usual para la significancia conjunta de los dos rezagos (Se están probando dos restricciones.)

Pregunta 12.4: Probablemente se estimaría la ecuación mediante las primeras diferencias, puesto que $\hat{\rho} = .92$ es lo suficientemente cercano a 1 para plantear dudas acerca de la regresión en niveles. Vea el capítulo 18 para un mayor análisis.

Pregunta 12.5: Debido a que sólo hay una variable explicativa, la prueba White es fácil de calcular. Simplemente realiza la regresión de $\hat{u}_t^2$ sobre $return_{t-1}$ y $return_{t-1}^2$ (con un intercepto, como siempre) y se calcula la prueba F para la significancia conjunta de $return_{t-1}$ y $return_{t-1}^2$. Si éstos son significativos en conjunto a un nivel de significancia lo bastante pequeño, se rechaza la nulidad de la homoscedasticidad.

Capítulo 13

Pregunta 13.1: Sí, suponiendo que se han controlado todos los factores relevantes. El coeficiente en *black* es 1.076, y, con un error estándar de .174, no es estadísticamente diferente de 1. El intervalo de confianza a 95% es de casi .735 a 1.417.

Pregunta 13.2: El coeficiente en *highearn* muestra que, en ausencia de algún cambio en el límite de ingresos, quienes ganan más pasan mucho más tiempo (en el orden de 29.2% en promedio debido a que $\exp(.256) - 1 \approx .292$], en la compensación de los trabajadores.

Pregunta 13.3: Primero, $\text{E}(v_{i1}) = \text{E}(a_i + u_{i1}) = \text{E}(a_i) + \text{E}(v_{i1}) = 0$. De forma similar, $\text{E}(v_{i2}) = 0$. Por tanto, la covarianza entre v_{i1} y v_{i2} es simplemente $\text{E}(v_{i1}v_{i2}) = \text{E}[(a_i + u_{i1})(a_i + u_{i2})] = \text{E}(a_i^2) + \text{E}(a_i u_{i1}) + \text{E}(a_i u_{i2}) + \text{E}(u_{i1}u_{i2}) = \text{E}(a_i^2)$, debido a que todos los términos de covarianza son cero por suposición. Pero $E(a_i^2) = Var(a_i)$, debido a que $E(a_i) = 0$. Esto ocasiona una correlación serial positiva a través del tiempo en los errores dentro de cada *i*, lo cual sesga los errores estándar MCO acostumbrados en una regresión de MCO combinada.

Pregunta 13.4: Debido a que $\Delta admn = admn_{90} - admn_{85}$ es la diferencia en indicadores binarios, puede ser -1 si, y sólo si, $admn_{90} = 0$ y $admn_{85} = 1$. En otras palabras, el estado de Washington tenía una ley administrativa propiamente dicha en 1985, pero fue derogada en 1990.

Pregunta 13.5: No, sólo así no causa sesgo ni inconsistencia en una regresión de series de tiempo con variables explicativas estrictamente exógenas. Hay dos razones por lo que es un problema. Primero, la correlación serial en los errores en cualquier ecuación suele sesgar los errores estándar MCO y los estadísticos de prueba habituales. Segundo, significa que MCO combinado no es tan eficaz como los estimadores que toman en cuenta la correlación serial (como en el capítulo 12).

Capítulo 14

Pregunta 14.1: Ya sea que se use la primera diferencia o la transformación within (o intra), se tendrán problemas para estimar el coeficiente en $kids_{it}$. Por ejemplo, usar la transformación within, si $kids_{it}$ no varía para la familia *i*, entonces $\ddot{kids}_{it} = kids_{it} - \overline{kids_i} = 0$ para $t = 1,2,3$. Siempre y cuando algunas familias tengan variación en $kids_{it}$, entonces se puede calcular el estimador de efectos fijos, pero el coeficiente *kids* puede estimarse con gran imprecisión. Ésta es una forma de multicolinealidad en la estimación de efectos fijos (o estimación de primeras diferencias).

Pregunta 14.2: Si una empresa no recibe un subsidio el primer año, puede o no recibir un subsidio en el segundo año. Pero si una empresa sí recibió el subsidio el primer año, puede no obtenerlo en el segundo año. Es decir si $grant_{-1} = 1$, entonces $grant = 0$. Esto induce una correlación negativa entre *grant* y $grant_{-1}$. Se puede verificar esto al calcular una regresión de *grant* sobre $grant_{-1}$, mediante los datos en JTRAIN.RAW para 1989. Al usar todas las empresas en la muestra, se obtiene

$$\begin{aligned} \widehat{grant} &= \underset{(.035)}{.248} - \underset{(.072)}{.248}\, grant_{-1} \\ n &= 157,\ R^2 = .070. \end{aligned}$$

El coeficiente en $grant_{-1}$ debe ser el negativo del intercepto debido a que $\widehat{grant} = 0$ cuando $grant_{-1} = 1$.

Pregunta 14.3: Sugiere que el efecto inobservable a_i está positivamente correlacionado con $union_{it}$. Recuerde que los MCO combinados dejan a a_i en el término de error, mientras que los efectos fijos eliminan a a_i. Por definición, a_i tiene un efecto positivo sobre log(*wage*). Por el análisis de las variables omitidas habitual (vea el capítulo 3), MCO tiene un sesgo ascendente cuando la variable explicativa (*union*) está correlacionada positivamente con la variable omitida (a_i). Por tanto, pertenecer a un sindicato parece tener una relación positiva con los factores inobservables constantes en el tiempo que afectan el salario.

Pregunta 14.4: No si todas la hermanas dentro de una familia tienen un mismo padre y madre. Entonces, como las variables de la raza de los progenitores no cambiarían por hermanas quedarían diferenciadas en (14.13).

Capítulo 15

Pregunta 15.1: Probablemente no. En la ecuación simple (15.18), los años de educación son parte del término de error. Si algunos hombres a quienes se asignaron números de sorteo de alistamiento bajos obtenían educación adicional, entonces el número del sorteo y la educación tienen una correlación negativa, lo que viola el primer requisito de una variable instrumental en la ecuación (15.4).

Pregunta 15.2: i) Para (15.27), se requiere que los efectos del grupo de compañeros de bachillerato se extiendan a la universidad. Es decir, para una calificación en SAT determinada, un estudiante que fue al bachillerato donde fumar marihuana era común, lo hará también en la universidad. Aun si la condición de identificación (15.27) es válida, el vínculo sería débil.

ii) Se tiene que asumir que el porcentaje de estudiantes que usan marihuana en el bachillerato no se correlaciona con factores inobservables que afectan el promedio de calificaciones universitarias. Aunque se está controlando en cierto grado la calidad del bachillerato al incluir el *SAT* en la ecuación, esto puede no ser suficiente. Quizá las preparatorias que hicieron un mejor trabajo preparando a los estudiantes para el bachillerato, también tenían un menor número de estudiantes consumidores de marihuana. O el uso de la marihuana podía estar correlacionado con los niveles de ingreso promedio. Éstas son, desde luego, cuestiones empíricas que podremos o no responder.

Pregunta 15.3: Aunque la influencia de la Asociación Nacional de Portadores de Rifles y los suscriptores a revistas de armas quizás esté correlacionada con la presencia de legislación de control de armas, no es evidente que no se correlacionen con factores inobservables que afecten la tasa de delitos violentos. De hecho, se podría argumentar que una población interesada en armas es un reflejo de las altas tasas delictivas, y con controlar las variables económicas y demográficas no basta para captarlo. Sería difícil argumentar de manera convincente que éstas sean verdaderamente exógenas en la ecuación de delitos violentos.

Pregunta 15.4: Como es habitual, existen dos requisitos. Primero, podría ocurrir que el aumento en el gasto gubernamental esté relacionado sistemáticamente con el partido presidencial, después de totalizar la tasa de inversión y el aumento en la fuerza de trabajo. En otras palabras, el instrumento debe correlacionarse de forma parcial con la variable explicativa endógena. Si bien se puede pensar que el gasto gubernamental aumenta más lentamente con presidentes republicanos, esto sin lugar a dudas no siempre ha sido cierto en Estados Unidos y tendría que demostrarse mediante el estadístico t en REP_{t-1} en la forma reducida $gGOV_t = \pi_0 + \pi_1 REP_{t-1} + \pi_2 INVRAT_t + \pi_3 gLAB_t + v_t$. Se debe asumir que el partido presidencial no tiene efecto separado sobre *gGDP*. Esto se violaría, por ejemplo, si las políticas monetarias difieren sistemáticamente según el partido presidencial y tiene un efecto diferente en el crecimiento del producto interno bruto (*GDP*).

Capítulo 16

Pregunta 16.1: Probablemente no. Esto se debe a que las empresas eligen de manera conjunta el precio y los gastos publicitarios, no interesa el experimento en el que, digamos, la publicidad cambie de manera exógena y se quiere saber el efecto en el precio. Por el contrario, el precio y la publicidad se modelarían en función de las variables de la demanda y los costos. Esto es lo que se desprende de la teoría económica.

Pregunta 16.2: Se asumen dos cosas. Primero, el aumento en la oferta de dinero debe aparecer en la ecuación (16.22), de manera que se correlaciona parcialmente con *inf*. Segundo, se debe asumir que el aumento en la oferta de dinero no aparece en la ecuación (16.23). Si se piensa que se debe incluir el aumento en la oferta de dinero en la ecuación (16.23), entonces aún se carecerá de un instrumento para *inf*. Desde luego, el supuesto de que el crecimiento en la oferta de dinero es exógeno también se puede poner en duda.

Pregunta 16.3: Se usa la prueba de Hausman del capítulo 15. En particular, sean $\hat{v}_2$ los residuales MCO de la regresión en forma reducida de *open* on log(*pcinc*) y log(*land*). Por tanto, utilice una regresión MCO de *inf* sobre *open*, log(*pcinc*), y $\hat{v}_2$ y se calcula el estadístico t para la significancia de $\hat{v}_2$. Si $\hat{v}_2$ es significativo, los estimadores de 2SLS y MCO son estadísticamente diferentes.

Pregunta 16.4: La ecuación de la demanda sería

$$\log(fish_t) = \beta_0 + \beta_1\log(prcfish_t) + \beta_2\log(inc_t)$$
$$+ \beta_3\log(prcchick_t) + \beta_4\log(prcbeef_t) + u_{t1},$$

donde los logaritmos se usan de manera que todas las elasticidades sean constantes. Por suposición, la función de la demanda no contiene estacionalidad alguna, de tal modo que la ecuación no contiene variables ficticias mensuales (digamos, feb_t, mar_t, …, dec_t, con enero como el mes base). Por otra parte, por suposición, la oferta del pescado es estacional, lo cual significa que la función de la oferta no depende de al menos algunas de las variables binarias mensuales. Incluso sin resolver la forma reducida para log(*prcfish*), se concluye que depende de las variables binarias mensuales. Dado que son exógenas, se pueden usar como instrumentos para log(*prcfish*) en la ecuación de la demanda. Por tanto, se puede estimar la ecuación de la demanda del pescado mediante variables binarias mensuales como IV para log(*prcfish*). La identificación requiere que al menos una variable binaria mensual aparezca con un coeficiente diferente de cero en la forma reducida para log(*prcfish*).

Capítulo 17

Pregunta 17.1: H_0: $\beta_4 = \beta_5 = \beta_6 = 0$, así que hay tres restricciones y por consiguiente tres *df* en la prueba *LR* o de Wald.

Pregunta 17.2: Se requiere la derivada parcial de $\Phi(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \hat{\beta}_3 exper + \hat{\beta}_4 exper^2 + \ldots)$ con respecto a *exper*, que es $\phi(\cdot)(\hat{\beta}_3 + 2\hat{\beta}_4 exper)$, donde $\phi(\cdot)$ se evalúa en los valores dados y el nivel inicial de experiencia. Por ello, se debe evaluar la densidad de la probabilidad normal estándar en $.270 - .012(20.13) + .131(12.3) + .123(10) - .0019(10^2) - .053(42.5) - .868(0) + .036(1) \approx .463$, donde se inserta el nivel inicial de experiencia (10). Pero $\phi(.463) = (2\pi)^{-1/2}\exp[-(.463^2)/2] \approx .358$. Después, se multiplica esto por $\hat{\beta}_3 + 2\hat{\beta}_4 exper$, que se evalúa en $exper = 10$. El efecto parcial mediante la aproximación de cálculo es $.358[.123 - 2(.0019)(10)] \approx .030$. En otras palabras, en los valores dados de las variables explicativas e iniciando en $exper = 10$, el siguiente año de experiencia aumenta la probabilidad de participación de la fuerza laboral en aproximadamente .03.

Pregunta 17.3: No. El número de aventuras extramaritales es un entero no negativo, que presumiblemente asume números de cero o pequeños para una fracción sustancial de la población. No es realista usar un modelo Tobit, que, si bien permite una acumulación en cero, trata a y como si estuviera distribuida de forma continua sobre valores positivos. Si se asume que $y = \max(0, y^*)$, donde y^* está distribuida de manera normal, no concuerda con la discreción del número de aventuras extramaritales cuando $y > 0$.

Pregunta 17.4: Los errores estándar ajustados son los errores estándar usuales de la estimación de Poisson MLE multiplicados por $\hat{\sigma} = \sqrt{2} \approx 1.41$, así que los errores estándar ajustados serán alrededor de 41% o superiores. El cuasi-estadístico de *LR* es el estadístico habitual de *LR* dividido entre $\hat{\sigma}^2$, de manera que será la mitad del estadístico usual de *LR*.

Pregunta 17.5: Por suposición, $mvp_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i$, en donde, como siempre, $\mathbf{x}_i\boldsymbol{\beta}$ denota una función lineal de las variables exógenas. Ahora, el salario observado es el mayor del salario mínimo y el producto del valor marginal, así que $wage_i = \max(minwage_i, mvp_i)$, que es muy similar a la ecuación (17.34), salvo que el operador max se ha remplazado por el operador min.

Capítulo 18

Pregunta 18.1: Se pueden insertar estos valores directamente en la ecuación (18.1) y se observa qué sucede. Primero, debido a que $z_s = 0$, para toda $s < 0$, $y_{-1} = \alpha + u_{-1}$. Entonces, $z_0 = 1$, de manera que $y_0 = \alpha + \delta_0 + u_0$. Para $h \geq 1$, $y_h = \alpha + \delta_{h-1} + \delta_h + u_h$. Debido a que los errores tienen valores esperados de cero, $\mathrm{E}(y_{-1}) = \alpha$, $\mathrm{E}(y_0) = \alpha + \delta_0$ y $\mathrm{E}(y_h) = \alpha + \delta_{h-1} + \delta$, para toda $h \geq 1$. Conforme $h \to \infty$, $\delta_h \to 0$. Se desprende que $\mathrm{E}(y_h) \to \alpha$ cuando $h \to \infty$, es decir, el valor esperado de y_h regresa al valor esperado antes del incremento en *z*, en el tiempo cero. Esto es lógico, aunque el incremento en *z* duró dos periodos, sigue siendo un incremento temporal.

Pregunta 18.2: En el contexto descrito, Δy_t y Δx_t son secuencias i.i.d., independientes entre sí. En particular, Δy_t y Δx_t no se correlacionan. Si $\hat{\gamma}_1$ es la pendiente de la regresión de Δy_t sobre Δx_t, $t = 1, 2, \ldots, n$, entonces plim $\hat{\gamma}_1 = 0$. Esto es como se esperaba, puesto se está haciendo la regresión de un proceso I(0) sobre otro proceso I(0), y no se correlacionan. Se escribe la ecuación $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + e_t$, donde $\gamma_0 = \gamma_1 = 0$. Debido a que $\{e_t\}$ es independiente de $\{\Delta x_t\}$, el supuesto de exogeneidad estricta se mantiene. Además, $\{e_t\}$ no se correlaciona serialmente y es homoscedástica. Por el teorema 11.2 en el capítulo 11, el estadístico *t* para $\hat{\gamma}_1$ tiene una distribución normal estándar aproximada. Si e_t se distribuye normalmente, los supuestos del modelo lineal clásico se mantienen, y el estadístico *t* tiene una distribución *t* exacta.

Pregunta 18.3: Escriba $x_t = x_{t-1} + a_t$, donde $\{a_t\}$ es I(0). Por suposición, existe una combinación lineal, digamos, $s_t = y_t - \beta x_t$, que es I(0). Ahora, $y_t - \beta x_{t-1} = y_t - \beta(x_t - a_t) = s_t + \beta a_t$. Debido a que s_t y a_t son I(0) por suposición, así también $s_t + \beta a_t$.

Pregunta 18.4: Sólo use la suma de los residuales cuadrados de la prueba *F* y asuma homoscedasticidad. La SSR restringida se obtiene al hacer la regresión de $\Delta hy6_t - \Delta hy3_{t-1} + (hy6_{t-1} - hy3_{t-2})$ sobre una constante. Observe que α_0 es el único parámetro para estimar en $\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2})$ cuando se imponen las restricciones. La suma irrestricta de los residuales cuadrados se obtiene de la ecuación (18.39).

Pregunta 18.5: Se están ajustando dos ecuaciones: $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$ y $\hat{y}_t = \hat{\gamma} + \hat{\delta}year_t$. Se puede obtener la relación entre los parámetros al observar que $year_t = t + 49$. Al insertar esto en la segunda ecuación se produce $\hat{y}_t = \hat{\gamma} + \hat{\delta}(t + 49) = (\hat{\gamma} + 49\hat{\delta}) + \hat{\delta}t$. Al unir la pendiente y el intercepto con la primera ecuación se tiene $\hat{\delta} = \hat{\beta}$, de manera que las pendientes en *t* y $year_t$ son idénticas y $\hat{\alpha} = \hat{\gamma} + 49\hat{\delta}$. En general, cuando se usa *year* en lugar de *t*, el intercepto cambiará, pero la pendiente no. (Puede verificar esto usando una de las bases de datos de series de tiempo, como HSEINV.RAW o INVEN.RAW.) Si se usa *t* o alguna otra medida de *year* no cambian los valores ajustados y, naturalmente no cambian los pronósticos de los valores futuros. El intercepto simplemente se ajusta de forma apropiada a diferentes formas de incluir una tendencia en la regresión.

Apéndice G

Tablas estadísticas

TABLA G.1

Áreas acumuladas bajo la distribución normal estándar

z	0	1	2	3	4	5	6	7	8	9
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483

(*continúa*)

TABLA G.1 (*continuación*)

z	0	1	2	3	4	5	6	7	8	9
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Ejemplos: Si $Z \sim$ Normal(0,1), entonces $P(Z \leq -1.32) = .0934$ y $P(Z \leq 1.84) = .9671$.
Fuente: Esta tabla se generó mediante la función normprob de Stata®.

TABLA G.2

Valores críticos de la distribución *t*

		Nivel de significancia				
1-Cola:		.10	.05	.025	.01	.005
2-Cola:		.20	.10	.05	.02	.01
Grados de libertad	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Ejemplos: El valor crítico al 1% para una prueba de una cola con 25 grados de libertad es 2.485. El valor crítico al 5% para una prueba de dos colas con grados de libertad grandes (> 120) es de 1.96.
Fuente: Esta tabla se generó mediante la función invttail de Stata®.

TABLA G.3a

Valores críticos al 10% de la distribución *F*

	Grados de libertad del numerador									
Grado de libertad del denominador	1	2	3	4	5	6	7	8	9	10
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

Ejemplo: El valor crítico al 10% de los grados de libertad del numerador = 2 y de los grados de libertad del denominador = 40 es 2.44.
Fuente: Esta tabla se generó mediante la función invFtail de Stata®.

TABLA G.3b

Valores críticos al 5% de la distribución *F*

Grados de libertad del denominador	Grados de libertad del numerador: 1	2	3	4	5	6	7	8	9	10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Ejemplo: El valor crítico al 5% de los grados de libertad del numerador = 4 y de los grados de libertad del denominador (∞) es 2.37.

Fuente: Esta tabla se generó mediante la función invFtail de Stata®.

TABLA G.3c

Valores críticos al 1% de la distribución *F*

Grados de libertad del denominador	Grados de libertad del numerador: 1	2	3	4	5	6	7	8	9	10
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Ejemplo: El valor crítico al 1% de los grados de libertad del numerador = 3 y de los grados de libertad del denominador = 60 es 4.13.

Fuente: Esta tabla se generó mediante la función invFtail de Stata®.

TABLA G.4

Valores críticos de la distribución *ji*-cuadrada

Grados de libertad	Nivel de significancia		
	.10	.05	.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

Ejemplo: El valor crítico al 5% de los grados de libertad del numerador = 8 es 15.51.
Fuente: Esta tabla se generó mediante la función invchi2tail de Stata®.

Referencias

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review* 80, 313-336.

Angrist, J. D., and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106, 979-1014.

Ashenfelter, O., and A. B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84, 1157-1173.

Averett, S., and S. Korenman (1996), "The Economic Reality of the Beauty Myth," *Journal of Human Resources* 31, 304-330.

Ayres, I., and S. D. Levitt (1998), "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics* 108, 43-77.

Banerjee, A., J. Dolado, J. W. Galbraith, and D. F. Hendry (1993), *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bartik, T. J. (1991), "The Effects of Property Taxes and Other Local Public Policies on the Intrametropolitan Pattern of Business Location," in *Industry Location and Public Policy.* Ed. H. W. Herzog and A. M. Schlottmann, 57-80. Knoxville: University of Tennessee Press.

Becker, G. S. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy* 76, 169-217.

Belsley, D., E. Kuh, and R. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Berk, R. A. (1990), "A Primer on Robust Regression," in *Modern Methods of Data Analysis*. Ed. J. Fox and J. S. Long, 292-324. Newbury Park, CA: Sage Publications.

Betts, J. R. (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics* 77, 231-250.

Biddle, J. E., and D. S. Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy* 98, 922-943.

Biddle, J. E., and D. S. Hamermesh (1998), "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics* 16, 172-201.

Blackburn, M., and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics* 107, 1421-1436.

Blomström, M., R. E. Lipsey, and M. Zejan (1996), "Is Fixed Investment the Key to Economic Growth?" *Quarterly Journal of Economics* 111, 269-276.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics* 52, 5-59.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994), "ARCH Models," Chapter 49 in *Handbook of Econometrics*, Volume 4. Ed. R. F. Engle and D. L. McFadden, 2959-3038. Amsterdam: North-Holland.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and Endogenous Explanatory Variables Is Weak," *Journal of the American Statistical Association* 90, 443-450.

Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica* 47, 987-1007.

Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Campbell, J. Y., and N. G. Mankiw (1990), "Permanent Income, Current Income, and Consumption," *Journal of Business and Economic Statistics* 8, 265-279.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Ed. L. N. Christophides, E. K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press.

Card, D., and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100, 1-40.

Castillo-Freeman, A. J., and R. B. Freeman (1992), "When the Minimum Wage Really Bites: The Effect of the U.S.-Level Minimum on Puerto Rico," in *Immigration and the Work Force*. Ed. G. J. Borjas and R. B. Freeman, 177-211. Chicago: University of Chicago Press.

Clark, K. B. (1984), "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review* 74, 893-919.

Cloninger, D. O. (1991), "Lethal Police Response as a Crime Deterrent: 57-City Study Suggests a Decrease in Certain Crimes," *American Journal of Economics and Sociology* 50, 59-69.

Cloninger, D. O., and L. C. Sartorius (1979), "Crime Rates, Clearance Rates and Enforcement Effort: The Case of Houston, Texas," *American Journal of Economics and Sociology* 38, 389-402.

Cochrane, J. H. (1997), "Where Is the Market Going? Uncertain Facts and Novel Theories," *Economic Perspectives* 21, Federal Reserve Bank of Chicago, 3-37.

Cornwell, C., and W. N. Trumbull (1994), "Estimating the Economic Model of Crime Using Panel Data," *Review of Economics and Statistics* 76, 360-366.

Currie, J. (1995), *Welfare and the Well-Being of Children*. Chur, Switzerland: Harwood Academic Publishers.

Currie, J., and N. Cole (1993), "Welfare and Child Health: The Link between AFDC Participation and Birth Weight," *American Economic Review* 83, 971-983.

Currie, J., and D. Thomas (1995), "Does Head Start Make a Difference?" *American Economic Review* 85, 341-364.

Davidson, R., and J. G. MacKinnon (1981), "Several Tests of Model Specification in the Presence of Alternative Hypotheses," *Econometrica* 49, 781-793.

Davidson, R., and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Long, J. B., and L. H. Summers (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics* 106, 445-502.

Dickey, D. A., and W. A. Fuller (1979), "Distributions of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association* 74, 427-431.

Diebold, F. X. (2001), *Elements of Forecasting*. 2d ed. Cincinnati, OH: South-Western.

Downes, T. A., and S. M. Greenstein (1996), "Understanding the Supply Decisions of Nonprofits: Modeling the Location of Private Schools," *Rand Journal of Economics* 27, 365-390.

Draper, N., and H. Smith (1981), *Applied Regression Analysis*. 2d ed. New York: Wiley.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association* 78, 605-610.

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regressions when Some of the Regressors Are Lagged Dependent Variables," *Econometrica* 38, 410-421.

Durbin, J., and G. S. Watson (1950), "Testing for Serial Correlation in Least Squares Regressions I," *Biometrika* 37, 409-428.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 59-82. Berkeley: University of California Press.

Eide, E. (1994), *Economics of Crime: Deterrence and the Rational Offender*. Amsterdam: North-Holland.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* 50, 987-1007.

Engle, R. F., and C. W. J. Granger (1987), "Cointegration and Error Correction: Representation, Estimation, and Testing," *Econometrica* 55, 251-276.

Evans, W. N., and R. M. Schwab (1995), "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 110, 941-974.

Fair, R. C. (1996), "Econometrics and Presidential Elections," *Journal of Economic Perspectives* 10, 89-102.

Franses, P. H., and R. Paap (2001), *Quantitative Models in Marketing Research*. Cambridge: Cambridge University Press.

Friedman, B. M., and K. N. Kuttner (1992), "Money, Income, Prices, and Interest Rates," *American Economic Review* 82, 472-492.

Geronimus, A. T., and S. Korenman (1992), "The Socioeconomic Consequences of Teen Childbearing Reconsidered," *Quarterly Journal of Economics* 107, 1187-1214.

Goldberger, A. S. (1991), *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Graddy, K. (1995), "Testing for Imperfect Competition at the Fulton Fish Market," *Rand Journal of Economics* 26, 75-92.

Graddy, K. (1997), "Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?" *Journal of Business and Economic Statistics* 15, 391-401.

Granger, C. W. J., and P. Newbold (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics* 2, 111-120.

Greene, W. (1997), *Econometric Analysis.* 3d ed. New York: MacMillan.

Griliches, Z. (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics* 39, 8-20.

Grogger, J. (1990), "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts," *Journal of the American Statistical Association* 410, 295-303.

Grogger, J. (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297-309.

Hall, R. E. (1988), "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy* 96, 921-948.

Hamermesh, D. S., and J. E. Biddle (1994), "Beauty and the Labor Market," *American Economic Review* 84, 1174-1194.

Hamilton, J. D. (1994), *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hanushek, E. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 24, 1141-1177.

Harvey, A. (1990), *The Econometric Analysis of Economic Time Series*. 2d ed. Cambridge, MA: MIT Press.

Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251-1271.

Hausman, J. A., and D. A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica* 45, 319-339.

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475-492.

Herrnstein, R. J., and C. Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hersch, J., and L. S. Stratton (1997), "Housework, Fixed Effects, and Wages of Married Workers," *Journal of Human Resources* 32, 285-307.

Hines, J. R. (1996), "Altered States: Taxes and the Location of Foreign Direct Investment in America," *American Economic Review* 86, 1076-1094.

Holzer, H. (1991), "The Spatial Mismatch Hypothesis: What Has the Evidence Shown?" *Urban Studies* 28, 105-122.

Holzer, H., R. Block, M. Cheatham, and J. Knott (1993), "Are Training Subsidies Effective? The Michigan Experience," *Industrial and Labor Relations Review* 46, 625-636.

Horowitz, J. (2001), "The Bootstrap," Chapter 52 in *Handbook of Econometrics*, Volume 5. E. Leamer and J. L. Heckman, 3159-3228. Amsterdam: North Holland.

Hoxby, C. M. (1994), "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper Number 4978.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221-233. Berkeley: University of California Press.

Hunter, W. C., and M. B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate Finance and Economics* 13, 57-70.

Hylleberg, S. (1992), *Modelling Seasonality*. Oxford: Oxford University Press.

Imbens, G. W., and J. M. Wooldridge (2007), "What's New in Econometrics?" Lecture Notes, National Bureau of Economic Research Summer Institute, 2007.

Kane, T. J., and C. E. Rouse (1995), "Labor-Market Returns to Two- and Four-Year Colleges," *American Economic Review* 85, 600-614.

Kiefer, N. M., and T. J. Vogelsang (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory* 21, 1130-1164.

Kiel, K. A., and K. T. McClain (1995), "House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation," *Journal of Environmental Economics and Management* 28, 241-255.

Kleck, G., and E. B. Patterson (1993), "The Impact of Gun Control and Gun Ownership Levels on Violence Rates," *Journal of Quantitative Criminology* 9, 249-287.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics* 17, 107-112.

Koenker, R. (2005), *Quantile Regression*. Cambridge: Cambridge University Press.

Korenman, S., and D. Neumark (1991), "Does Marriage Really Make Men More Productive?" *Journal of Human Resources* 26, 282-307.

Korenman, S., and D. Neumark (1992), "Marriage, Motherhood, and Wages," *Journal of Human Resources* 27, 233-255.

Krueger, A. B. (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989," *Quarterly Journal of Economics* 108, 33-60.

Krupp, C. M., and P. S. Pollard (1996), "Market Responses to Antidumping Laws: Some Evidence from the U.S. Chemical Industry," *Canadian Journal of Economics* 29, 199-227.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54, 159-178.

Lalonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604-620.

Larsen, R. J., and M. L. Marx (1986), *An Introduction to Mathematical Statistics and Its Applications.* 2d ed. Englewood Cliffs, NJ: Prentice-Hall.

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review* 73, 31-43.

Levine, P. B., A. B. Trainor, and D. J. Zimmerman (1996), "The Effect of Medicaid Abortion Funding Restrictions on Abortions, Pregnancies, and Births," *Journal of Health Economics* 15, 555-578.

Levine, P. B., and D. J. Zimmerman (1995), "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economics Statistics* 13, 137-149.

Levitt, S. D. (1994), "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House," *Journal of Political Economy* 102, 777-798.

Levitt, S. D. (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation," *Quarterly Journal of Economics* 111, 319-351.

Low, S. A., and L. R. McPheters (1983), "Wage Differentials and the Risk of Death: An Empirical Analysis," *Economic Inquiry* 21, 271-280.

Lynch, L. M. (1992), "Private Sector Training and the Earnings of Young Workers," *American Economic Review* 82, 299-312.

MacKinnon, J. G., and H. White (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29, 305-325.

Maloney, M. T., and R. E. McCormick (1993), "An Examination of the Role that Intercollegiate Athletic Participation Plays in Academic Achievement: Athletes' Feats in the Classroom," *Journal of Human Resources* 28, 555-570.

Mankiw, N. G. (1994), *Macroeconomics*. 2d ed. New York: Worth.

McCarthy, P. S. (1994), "Relaxed Speed Limits and Highway Safety: New Evidence from California," *Economics Letters* 46, 173-179.

McClain, K. T., and J. M. Wooldridge (1995), "A Simple Test for the Consistency of Dynamic Linear Regression in Rational Distributed Lag Models," *Economics Letters* 48, 235-240.

McCormick, R. E., and M. Tinsley (1987), "Athletics versus Academics: Evidence from SAT Scores," *Journal of Political Economy* 95, 1103-1116.

McFadden, D. L. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics.* Ed. P. Zarembka, 105-142. New York: Academic Press.

Meyer, B. D. (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* 13, 151-161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85, 322-340.

Mizon, G. E., and J. F. Richard (1986), "The Encompassing Principle and Its Application to Testing Nonnested Hypotheses," *Econometrica* 54, 657-678.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica* 55, 765-799.

Mullahy, J., and P. R. Portney (1990), "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health," *Journal of Health Economics* 9, 193-205.

Mullahy, J., and J. L. Sindelar (1994), "Do Drinkers Know When to Say When? An Empirical Analysis of Drunk Driving," *Economic Inquiry* 32, 383-394.

Netzer, D. (1992), "Differences in Reliance on User Charges by American State and Local Governments," *Public Finance Quarterly* 20, 499-511.

Neumark, D. (1996), "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics* 111, 915-941.

Neumark, D., and W. Wascher (1995), "Minimum Wage Effects on Employment and School Enrollment," *Journal of Business and Economic Statistics* 13, 199-206.

Newey, W. K., and K. D. West (1987), "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703-708.

Papke, L. E. (1987), "Subnational Taxation and Capital Mobility: Estimates of Tax-Price Elasticities," *National Tax Journal* 40, 191-203.

Papke, L. E. (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37-49.

Papke, L. E. (1995), "Participation in and Contributions to 401(k) Pension Plans: Evidence from Plan Data," *Journal of Human Resources* 30, 311-325.

Papke, L. E. (1999), "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data," *Journal of Human Resources*, 34, 346-368.

Papke, L. E. (2005), "The Effects of Spending on Test Pass Rates: Evidence from Michigan," *Journal of Public Economics* 89, 821-839.

Park, R. (1966), "Estimation with Heteroskedastic Error Terms," *Econometrica* 34, 888.

Peek, J. (1982), "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review* 72, 980-991.

Pindyck, R. S., and D. L. Rubinfeld (1992), *Microeconomics*. 2d ed. New York: MacMillan.

Ram, R. (1986), "Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data," *American Economic Review* 76, 191-203.

Ramanathan, R. (1995), *Introductory Econometrics with Applications*. 3d ed. Fort Worth: Dryden Press.

Ramey, V. (1991), "Nonconvex Costs and the Behavior of Inventories," *Journal of Political Economy* 99, 306-334.

Ramsey, J. B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Analysis," *Journal of the Royal Statistical Association*, Series B, 71, 350-371.

Romer, D. (1993), "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics* 108, 869-903.

Rose, N. L. (1985), "The Incidence of Regulatory Rents in the Motor Carrier Industry," *Rand Journal of Economics* 16, 299-318.

Rose, N. L., and A. Shepard (1997), "Firm Diversification and CEO Compensation: Managerial Ability or Executive Entrenchment?" *Rand Journal of Economics* 28, 489-514.

Rouse, C. E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics* 113, 553-602.

Sander, W. (1992), "The Effect of Women's Schooling on Fertility," *Economic Letters* 40, 229-233.

Savin, N. E., and K. J. White (1977), "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica* 45, 1989-1996.

Shea, J. (1993), "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11, 145-155.

Shughart, W. F., and R. D. Tollison (1984), "The Random Character of Merger Activity," *Rand Journal of Economics* 15, 500-509.

Solon, G. (1985), "The Minimum Wage and Teenage Employment: A Re-analysis with Attention to Serial Correlation and Seasonality," *Journal of Human Resources* 20, 292-297.

Staiger, D., and J. H. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, 557-586.

Stigler, S. M. (1986), *The History of Statistics*. Cambridge, MA: Harvard University Press.

Stock, J. H., and M. W. Watson (1989), "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics* 40, 161-181.

Stock, J. H., and M. W. Watson (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica* 61, 783-820.

Sydsaeter, K., and P. J. Hammond (1995), *Mathematics for Economic Analysis.* Englewood Cliffs, NJ: Prentice Hall.

Vella, F., and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics* 13, 163-183.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *Annals of Mathematical Statistics* 11, 284-300.

Wallis, K. F. (1972), "Testing for Fourth-Order Autocorrelation in Quarterly Regression Equations," *Econometrica* 40, 617-636.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817-838.

White, H. (1984), *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, M. J. (1986), "Property Taxes and Firm Location: Evidence from Proposition 13," in *Studies in State and Local Public Finance*. Ed. H. S. Rosen, 83-112. Chicago: University of Chicago Press.

Whittington, L. A., J. Alm, and H. E. Peters (1990), "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States," *American Economic Review* 80, 545-556.

Wooldridge, J. M. (1989), "A Computationally Simple Heteroskedasticity and Serial Correlation-Robust Standard Error for the Linear Regression Model," *Economics Letters* 31, 239-243.

Wooldridge, J. M. (1991a), "A Note on Computing *R*-Squared and Adjusted *R*-Squared for Trending and Seasonal Data," *Economics Letters* 36, 49-54.

Wooldridge, J. M. (1991b), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics* 47, 5-46.

Wooldridge, J. M. (1994a), "A Simple Specification Test for the Predictive Ability of Transformation Models," *Review of Economics and Statistics* 76, 59-65.

Wooldridge, J. M. (1994b), "Estimation and Inference for Dependent Processes," Chapter 45 in *Handbook of Econometrics*, Volume 4. Ed. R. F. Engle and D. L. McFadden, 2639-2738. Amsterdam: North-Holland.

Wooldridge, J. M. (1995), "Score Diagnostics for Linear Models Estimated by Two Stage Least Squares," in *Advances in Econometrics and Quantitative Economics*. Ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66-87. Oxford: Blackwell.

Wooldridge, J. M. (2002), Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Glosario

A

Ajuste estacional: Eliminación de los componentes estacionales de una serie de tiempo mensual o trimestral.

Ajustados estacionalmente: Datos de series de tiempo trimestrales o mensuales donde se ha usado un procedimiento estadístico (posiblemente regresión sobre variables binarias estacionales) para eliminar el componente estacional.

Altamente persistente: Proceso de series de tiempo donde los resultados en el futuro distante están altamente correlacionados con los resultados actuales.

Alternativa de dos colas: Alternativa donde el parámetro poblacional puede ser mayor o menor que el valor establecido bajo la hipótesis nula.

Alternativa de una cola: Hipótesis alternativa que establece que el parámetro es mayor (o menor) que el valor hipotético de la nula.

Análisis de duración: Una aplicación del modelo de regresión censurada, en la que la variable dependiente es el tiempo que transcurre hasta que sucede cierto evento, como el tiempo que pasa antes de que una persona desempleada vuelva a trabajar.

Análisis de error de especificación: Proceso de determinar los sesgos probables que pudieran surgir de variables omitidas, error de medición, simultaneidad y otros tipos de errores de especificación del modelo.

Análisis de políticas: Análisis empírico que utiliza métodos econométricos para evaluar los efectos de cierta política.

Análisis de regresión múltiple: Tipo de análisis usado para describir la estimación e inferencia en el modelo de regresión lineal múltiple.

Análisis de residuales: Tipo de análisis que estudia el signo y tamaño de los residuales para observaciones particulares después de haberse estimado un modelo de regresión múltiple.

Análisis de sensibilidad: Proceso de verificar si los efectos estimados y significancia estadística de las variables explicativas claves son sensibles a la inclusión de otras variables explicativas, otra forma funcional, a la eliminación de posibles observaciones atípicas o diferentes métodos de estimación.

Análisis empírico: Estudio que usa los datos en un análisis econométrico formal para poner a prueba la teoría, estimar una relación o determinar la efectividad de una política.

Archivo de texto (ASCII): Formato universal de archivos que se puede transportar a través de numerosas plataformas de cómputo.

Asimetría: Medida de qué tan lejos está una distribución de ser simétrica, con base en el tercer momento de la variable aleatoria estandarizada.

Asintóticamente eficiente: Para estimadores consistentes con distribuciones asintóticamente normales, el estimador con la menor varianza asintótica.

Autocorrelación: *Vea* correlación serial.

Autocorrelación de primer orden: Para un proceso de series de tiempo ordenado cronológicamente, el coeficiente de correlación entre pares de observaciones adyacentes.

Autoselección: Decidir sobre una acción con base en beneficios probables, o costos, de emprender esa acción.

B

Base de datos de corte transversal: Base de datos de una población en un momento determinado.

Bases de datos en línea: Bases de datos a las que se puede acceder a través de una red de cómputo.

Bootstrap: Método de remuestreo que extrae muestras aleatorias con remplazo del conjunto original de datos.

C

Cambio en puntos porcentuales: Cambio en una variable que está medida como un porcentaje.

Cambio porcentual: Cambio proporcional en una variable, multiplicado por 100.

Cambio proporcional: Cambio en una variable en relación con su valor inicial; matemáticamente, el cambio dividido entre el valor inicial.

Cambio relativo: *Vea* cambio proporcional.

Caminata aleatoria: Proceso de series de tiempo donde el valor del periodo siguiente se obtiene como el valor de este periodo, más un término de error independiente (o al menos uno no correlacionado).

Caminata aleatoria con tendencia estocástica: Caminata aleatoria que tiene una constante (o tendencia) en cada periodo.

Causalidad de Granger: Noción limitada de la causalidad donde los valores pasados de una serie (x_t) son útiles para predecir valores futuros de otra (y_t) después de controlar por los valores pasados de y_t.

Censura de datos: Situación que surge cuando no siempre se observa el resultado en la variable dependiente debido a que en el límite superior (o inferior) sólo se sabe que el resultado estaba por encima (o debajo) del límite. (*Vea también* modelo de regresión censurada.)

Ceteris paribus: Todos los demás factores relevantes permanecen constantes.
Codificación superior: Forma de censura de datos en donde el valor de una variable no se reporta cuando está por encima de determinado umbral; sólo se sabe que es al menos de la misma magnitud que el umbral.
Coeficiente de correlación: Medida de dependencia lineal entre dos variables aleatorias que no depende de unidades de medida y que está limitada entre -1 y 1.
Coeficiente de correlación muestral: Estimación del coeficiente de correlación (poblacional) a partir de una muestra de datos.
Coeficientes beta: *Vea* coeficientes estandarizados.
Coeficiente de determinación: *Vea* *R*-cuadrada.
Coeficientes estandarizados: Coeficientes de regresión que miden el cambio medido en desviaciones estándar de la variable dependiente, dado un incremento de una desviación estándar en una variable independiente.
Cointegración: Noción de que una combinación lineal de dos series, cada una integrada de orden uno, es integrada de orden cero.
Colinealidad perfecta: En regresión múltiple, una variable independiente es una función lineal exacta de una o más variables independientes diferentes.
Condición de orden: Condición necesaria para identificar los parámetros en un modelo con una o más variables explicativas endógenas: el número total de variables exógenas debe ser al menos igual al número total de variables explicativas.
Condiciones de primer orden: Conjunto de ecuaciones lineales usadas para encontrar los estimadores de MCO.
Condición de rango: Condición suficiente para identificar un modelo con una o más variables explicativas endógenas.
Combinación independiente de cortes transversales: Base de datos obtenida al combinar muestras aleatorias independientes provenientes de diferentes puntos en el tiempo.
Conjuntamente no significativas: Fracaso en rechazar, mediante una prueba F a un nivel específico de significancia, que todos los coeficientes de un grupo de variables explicativas son cero.
Conjunto de información: En pronósticos, el conjunto de variables que se pueden observar antes de formar un pronóstico.
Consistencia: Conforme el tamaño de la muestra crece, un estimador converge en probabilidad al valor poblacional correcto.
Contemporáneamente exógeno: En aplicaciones de series de tiempo o datos de panel, un regresor es contemporáneamente exógeno si no está correlacionado con el término de error del mismo periodo de tiempo, aunque puede estar correlacionado con los errores de otros periodos de tiempo.
Contemporáneamente homocedástico: En series de tiempo o aplicaciones de datos de panel, la varianza del término de error, condicional en los regresores del mismo periodo de tiempo, es constante.
Correlación espuria: Correlación entre dos variables que no se debe a la causalidad, sino quizás a la dependencia que tienen ambas de un factor inobservable.
Correlación muestral: Para resultados en dos variables aleatorias, la covarianza muestral dividida entre el producto de las desviaciones estándar muestrales.
Correlación serial: En un modelo de series de tiempo o de datos de panel, la correlación entre los errores de diferentes periodos de tiempo.
Correlación serial AR(1): Los errores en un modelo de regresión de series de tiempo siguen un modelo AR(1).
Cortes transversales combinados: Configuración de datos donde cortes transversales independientes, usualmente recolectados en distintos puntos del tiempo, se combinan para producir una sola base de datos.
Covariada: *Vea* variable explicativa.
Covarianza: Medida de la dependencia lineal entre dos variables aleatorias.
Covarianza muestral: Estimador insesgado de la covarianza poblacional entre dos variables aleatorias.
Criterios dentro de la muestra: Criterios para elegir modelos de pronóstico, con base en la bondad de ajuste dentro de la muestra que fue usada para obtener las estimaciones de los parámetros.
Criterios fuera de la muestra: Criterios usados para elegir modelos de pronóstico basados en una parte de la muestra que no se usó para obtener las estimaciones de los parámetros.
Cuasi experimento: *Vea* experimento natural.
Curtosis: Medida del espesor de las colas de una distribución con base en el cuarto momento de la variable aleatoria estandarizada; la medida suele compararse con el valor de la distribución normal estándar, que es tres.

D

Datos con el tiempo deducido: Datos de panel donde, para cada unidad de corte transversal, a los datos en cada periodo de tiempo se les resta su promedio a través de los periodos de tiempo.
Datos cuasi deducidos: En la estimación de efectos aleatorios para datos de panel, son los datos originales en cada periodo menos una fracción de su promedio a lo largo del tiempo; estos cálculos se realizan para cada unidad de corte transversal.
Datos cuasi diferenciados: En la estimación de un modelo de regresión con correlación serial AR(1), es la diferencia entre el periodo actual y un múltiplo del periodo anterior, donde el múltiplo es el parámetro en el modelo AR(1).
Datos de panel: Base de datos formada por cortes transversales repetidos a través del tiempo. Con un panel *balanceado*, las mismas unidades aparecen en cada periodo de tiempo. Con panel *no balanceado*, algunas unidades no aparecen en cada periodo de tiempo, con frecuencia debido a su desaparición.
Datos de series de tiempo: Datos recolectados a través del tiempo acerca de una o más variables.
Datos experimentales: Datos que se han obtenido al realizar un experimento controlado.
Datos faltantes: Problema de datos que ocurre cuando no se observan los valores de algunas variables para ciertas observaciones (individuos, ciudades, periodos de tiempo, etc.) en la muestra.
Datos longitudinales: *Vea* datos de panel.
Datos no experimentales: Datos que no se han obtenido al realizar un experimento controlado.
Datos observacionales: *Vea* datos no experimentales.

Datos retrospectivos: Datos recolectados basados en información pasada más que actual.

Débilmente dependiente: Término que describe un proceso de series de tiempo donde alguna medida de dependencia entre dos variables aleatorias de dos momentos del tiempo, como la correlación, disminuye en la medida en que aumenta el intervalo entre los dos puntos del tiempo.

Definida positiva: Matriz simétrica en la que todas las formas cuadráticas, salvo la trivial que debe ser cero, son estrictamente positivas.

Derivada: Pendiente de una función suave, como se define en cálculo.

Derivada parcial: Para una función suave de más de una variable, la pendiente de la función en una dirección.

Desplazamiento del intercepto: El intercepto en un modelo de regresión es diferente por grupo o por periodo de tiempo.

Desviación estándar: Medida común de la dispersión de la distribución de una variable aleatoria.

Desviación estándar de $\hat{\beta}_j$: Medida común de la dispersión de la distribución de muestreo de $\hat{\beta}_j$.

Desviación estándar de muestreo: Desviación estándar de un estimador, esto es, la desviación estándar de una distribución de muestreo.

Desviación estándar muestral: Estimador consistente de la desviación estándar poblacional.

Diferencia en pendientes: Descripción de un modelo donde algunos parámetros de pendiente pueden diferir entre grupos o periodos de tiempo.

Distribución binomial: La distribución de probabilidad del número de éxitos ocurridos en n eventos independientes de Bernoulli, donde cada evento tiene la misma probabilidad de éxito.

Distribución condicional: Distribución de probabilidad de una variable aleatoria, dados los valores de una o más variables aleatorias diferentes.

Distribución conjunta: Distribución de probabilidad que determina las probabilidades de resultados que impliquen a dos o más variables aleatorias.

Distribución de Dickey-Fuller: Distribución límite del estadístico t al probar la hipótesis nula de una raíz unitaria.

Distribución de muestreo: Distribución de probabilidad de un estimador sobre todos los posibles resultados de la muestra.

Distribución de rezagos: En un modelo de rezagos distribuidos finitos o infinitos, los coeficientes de los rezagos graficados como una función de la extensión del rezago.

Distribución *F*: Distribución de probabilidad obtenida al formar la razón de dos variables aleatorias ji-cuadrada independientes, donde cada una está dividida entre sus grados de libertad.

Distribución ji-cuadrada: Distribución de probabilidad obtenida mediante la suma de los cuadrados de variables aleatorias normal estándar independientes. El número de términos en la suma es igual a los grados de libertad en la distribución.

Distribución normal: Distribución de probabilidad que se usa comúnmente en estadística y econometría para modelar una población. Su función de distribución de probabilidad tiene forma de campana.

Distribución normal estándar: Distribución normal con media cero y varianza uno.

Distribución normal multivariada: Distribución para múltiples variables aleatorias donde cada combinación lineal de las variables aleatorias tiene una distribución normal univariada (unidimensional).

Distribución Poisson: Distribución de probabilidad para las variables de conteo.

Distribución simétrica: Distribución de probabilidad caracterizada por una función de densidad de probabilidad simétrica en torno a su mediana, que también debe ser el valor de la media (siempre que la media exista).

Distribución *t*: Distribución de la razón formada por una variable aleatoria normal estándar y la raíz cuadrada de una variable aleatoria ji-cuadrada, donde la variable aleatoria ji-cuadrada se divide primero entre sus *gl*.

E

Ecuación en la forma reducida: Ecuación lineal donde una variable endógena es una función de las variables exógenas y los errores inobservables.

Ecuación en primeras diferencias (PD): En modelos de series de tiempo o datos de panel, ecuación donde a todas las variables dependientes e independientes se les ha aplicado la primera diferencia.

Ecuación estructural: Ecuación derivada de la teoría económica o de un razonamiento económico menos formal.

Ecuación exactamente identificada: Para modelos con variables explicativas endógenas, una ecuación que es identificada, pero que no se habría identificado con una variable instrumental menos.

Ecuación identificada: Ecuación cuyos parámetros se pueden estimar consistentemente, en especial en modelos con variables explicativas endógenas.

Ecuación no identificada: Ecuación con una o más variables explicativas endógenas donde no existen suficientes variables instrumentales para identificar los parámetros.

Ecuación sobreidentificada: En modelos con variables explicativas endógenas, una ecuación donde el número de variables instrumentales es estrictamente mayor que el número de variables explicativas endógenas.

Editor de texto: Software de cómputo que se puede usar para editar archivos de texto.

Efecto causal: Un cambio *ceteris paribus* en una variable tiene un efecto en otra.

Efecto de grupo: Efecto inobservable común a todas las unidades, por lo general personas, en un grupo.

Efecto de interacción: En regresión múltiple, el efecto parcial de una variable explicativa depende del valor de otra variable explicativa diferente.

Efecto fijo: *Vea* efecto inobservable.

Efecto inobservable: En un modelo de datos de panel, una variable inobservable es el término de error que no cambia con el tiempo. Para muestras de agrupamientos, una variable inobservable que es común a todas las unidades en el grupo.

Efecto marginal: Efecto en la variable dependiente que resulta de cambiar un poco una variable independiente.

Efecto marginal decreciente: El efecto marginal de una variable explicativa disminuye a medida que el valor de la variable explicativa aumenta.

Efecto parcial: El efecto de una variable explicativa en la variable dependiente, manteniendo fijos los demás factores en el modelo de regresión.

Efecto parcial en los valores promedio (EPeP): En modelos con efectos parciales no constantes, el efecto parcial evaluado con base en los valores promedio de las variables explicativas.

Efecto parcial promedio: Para efectos parciales no constantes, el efecto parcial promediado sobre la población especificada.

Efecto promedio del tratamiento: Efecto de un tratamiento o política promediado sobre la población.

Elasticidad: Cambio porcentual en una variable, dado un incremento *ceteris paribus* de 1% en otra variable.

Elasticidad de corto plazo: La propensión de impacto en un modelo de rezagos distribuidos cuando las variables dependiente e independiente están en forma logarítmica.

Elasticidad de impacto: En un modelo de rezagos distribuidos, el cambio porcentual inmediato en la variable dependiente, dado un incremento de 1% en la variable independiente.

Elasticidad de largo plazo: La propensión de largo plazo en un modelo de rezagos distribuidos con las variables dependiente e independiente en forma logarítmica; por tanto, la elasticidad de largo plazo es el incremento porcentual eventual en la variable dependiente, dado un incremento permanente de 1% en la variable explicativa.

Eliminación de la tendencia: Práctica de eliminar la tendencia en una serie de tiempo.

Endogeneidad: Término usado para describir la presencia de una variable explicativa endógena.

Error absoluto medio (EAM): Medida del desempeño en el pronóstico, calculada como el promedio de los valores absolutos de los errores de pronóstico.

Error cuadrático medio (ECM): Distancia cuadrada esperada a la que un estimador está del valor poblacional; es igual a la varianza más el cuadrado del sesgo.

Error de la forma reducida: Término de error que aparece en una ecuación en la forma reducida.

Error de medición multiplicativo: Error de medición donde la variable observada es el producto de la variable inobservable verdadera y un error positivo de medición.

Error de medición: Diferencia entre la variable observada y la variable que pertenece a la ecuación de regresión múltiple.

Error de predicción: Diferencia entre el resultado real y una predicción de ese resultado.

Error de pronóstico: Diferencia entre el resultado real y el pronóstico.

Error del Tipo I: Rechazo de la hipótesis nula cuando esta es verdadera.

Error del Tipo II: No rechazar la hipótesis nula cuando esta es falsa.

Error estándar: Genéricamente, una estimación de la desviación estándar de un estimador.

Error estándar asintótico: Error estándar válido en muestras grandes.

Error estándar bootstrap: Error estándar obtenido como la desviación estándar muestral de una estimación a través de todas las muestras del remuestreo.

Error estándar de $\hat{\beta}_j$: Estimación de la desviación estándar de la distribución de muestreo de $\hat{\beta}_j$.

Error estándar de la estimación: *Vea* error estándar de la regresión.

Error estándar de la regresión (EER): En el análisis de regresión múltiple, la estimación de la desviación estándar del error poblacional, obtenida como la raíz cuadrada de la suma de los residuales cuadrados sobre los grados de libertad.

Error estándar robusto a la correlación serial: Error estándar para un estimador que es (asintóticamente) válido aunque los errores en el modelo estén o no serialmente correlacionados.

Error estándar robusto a la heterocedasticidad: Error estándar que es (asintóticamente) robusto a la heterocedasticidad de la forma desconocida.

Error estructural: Término de error en una ecuación estructural, la cual podría ser una ecuación en un modelo de ecuaciones simultáneas.

Error idiosincrático: En modelos de datos de panel, el error que cambia con el tiempo, así como a través de las unidades (por ejemplo, individuos, empresas o ciudades).

Errores clásicos en las variables (ECV): Modelo de error de medición en donde la medida observada es igual a la variable real más un error de medición independiente o al menos no correlacionado.

Errores en las variables: Situación donde la variable dependiente o algunas variables independientes se miden con errores.

Especificación incorrecta de la forma funcional: Problema que ocurre cuando un modelo omite funciones de las variables explicativas (como cuadráticas) o usa funciones equivocadas de la variable dependiente o de algunas variables explicativas.

Esperanza condicional: Valor esperado o promedio de una variable aleatoria, llamada variable dependiente o explicada, que depende de los valores de una o más variables diferentes, llamadas variables independientes o explicativas.

Estacionalidad: Característica de las series de tiempo mensuales o trimestrales donde el valor promedio es sistemáticamente diferente por temporada del año.

Estacionario en covarianza: Proceso de series de tiempo con una media y varianza constantes donde la covarianza entre dos variables aleatorias cualesquiera en la secuencia depende sólo de la distancia entre ellas.

Estadística descriptiva: Estadística que se usa para resumir un conjunto de números; el promedio de la muestra, la mediana de la muestra y la desviación estándar de la muestra son las más comunes.

Estadísticamente diferente de cero: *Vea* estadísticamente significativo.

Estadísticamente no significativo: No rechazar la hipótesis nula de que un parámetro poblacional es igual a cero, al nivel de significancia elegido.

Estadísticamente significativas conjuntamente: La hipótesis nula de que dos o más variables explicativas tienen coeficientes poblacionales de cero se rechaza al nivel de significancia elegido.

Estadísticamente significativo: Rechazar la hipótesis nula de que un parámetro es igual a cero contra la alternativa especificada, al nivel de significancia elegido.

Estadístico de Chow: Estadístico *F* para probar la igualdad de los parámetros de regresión a través de diferentes grupos (hombres y mujeres, por ejemplo) o periodos de tiempo (por ejemplo, antes y después del cambio de una política).

Estadístico de prueba: Regla usada para probar hipótesis donde cada resultado de la muestra produce un valor numérico.

Estadístico de puntuación (o estadístico score): *Vea* estadístico del multiplicador de Lagrange.

Estadístico de razón de cuasi verosimilitudes: Modificación del estadístico de razón de verosimilitudes que toma en cuenta la especificación incorrecta de la distribución, como en el modelo de regresión Poisson.

Estadístico de razón de verosimilitudes: Estadístico que se puede usar para probar una o múltiples hipótesis cuando los modelos restringido y no restringido se han estimado mediante máxima verosimilitud. El estadístico es dos veces la diferencia entre la log-verosimilitud no-restringida y la restringida.

Estadístico de Wald: Estadístico general de prueba para probar hipótesis en una variedad de escenarios econométricos; por lo general, el estadístico de Wald tiene una distribución ji-cuadrada asintótica.

Estadístico del multiplicador de Lagrange (ML): Estadístico de prueba justificado para muestras grandes que sirve para probar variables omitidas, heteroscedasticidad y correlación serial, entre otros problemas en la especificación de modelos.

Estadístico Durbin-Watson (DW): Estadístico usado para probar la correlación serial de primer orden en los errores de un modelo de regresión de series de tiempo, bajo los supuestos clásicos del modelo lineal.

Estadístico *F*: Estadístico usado para probar hipótesis múltiples acerca de los parámetros en un modelo de regresión múltiple.

Estadístico *F* robusto a la heterocedasticidad: Estadístico de tipo *F* que es (asintóticamente) robusto a la heterocedasticidad de la forma desconocida.

Estadístico ML robusto a la heterocedasticidad: Estadístico ML robusto a la heterocedasticidad de la forma desconocida.

Estadístico *n-R*-cuadrada: *Vea* estadístico del multiplicador de Lagrange.

Estadístico *t*: Estadístico usado para comprobar una hipótesis simple acerca de los parámetros en un modelo econométrico.

Estadístico *t* asintótico: Estadístico *t* que tiene una distribución normal estándar aproximada en muestras grandes.

Estadístico *t* robusto a la heterocedasticidad: Estadístico de tipo *t* que es (asintóticamente) robusto a la heterocedasticidad de la forma desconocida.

Estimación: Valor numérico que un estimador adopta para una determinada muestra de datos.

Estimación combinada de MCO: La estimación de MCO con cortes transversales, datos de panel o muestras de agrupamientos independientes combinados, donde las observaciones se combinan en el tiempo (o entre grupos) así como entre las unidades de corte transversal.

Estimación de Cochrane-Orcutt (CO): Método para estimar un modelo de regresión lineal múltiple con errores AR(1) y variables explicativas estrictamente exógenas; a diferencia de Prais-Winsten, Cochrane-Orcutt no estima la ecuación para el primer periodo de tiempo.

Estimación de cuasi máxima verosimilitud (ECMV): Estimación de máxima verosimilitud, pero donde la función de log-verosimilitud quizá no corresponda a la verdadera distribución condicional de la variable dependiente.

Estimación de la pendiente de MCO: Pendiente en una línea de regresión de MCO.

Estimación de máxima verosimilitud (EMV): Método de estimación ampliamente aplicable donde las estimaciones de los parámetros se eligen para maximizar la función de log-verosimilitud.

Estimación de Prais-Winsten (PW): Método para estimar un modelo de regresión lineal múltiple con errores AR(1) y variables explicativas estrictamente exógenas; a diferencia de Cochrane-Orcutt, Prais-Winsten estima la ecuación para el primer periodo de tiempo.

Estimación del intercepto de MCO: Intercepto en una línea de regresión de MCO.

Estimador: Regla para combinar datos que produce un valor numérico para un parámetro poblacional; la forma de la regla no depende de la muestra particular obtenida.

Estimador consistente: Conforme el tamaño de la muestra aumenta, el estimador que converge en probabilidad al parámetro poblacional.

Estimador de adelantos y rezagos: Estimador de un parámetro cointegrador en una regresión con variables I(1), donde las primeras diferencias actuales, algunas pasadas y algunas futuras de la variable explicativa se incluyen como regresores.

Estimador de diferencia-en-diferencias: Estimador que surge en el análisis de políticas con datos para dos periodos de tiempo. Una versión del estimador aplica a cortes transversales combinados independientemente y otra a bases de datos de panel.

Estimador de efectos aleatorios: Estimador de MCG factibles en el modelo de efectos inobservables, donde se asume que el efecto inobservable no está correlacionado con las variables explicativas en cada periodo de tiempo.

Estimador de efectos fijos: Para el modelo de datos de panel de efectos inobservables, el estimador obtenido al aplicar la estimación combinada de MCO a una ecuación con el tiempo deducido.

Estimador de intervalo: Regla que usa datos para obtener límites inferiores y superiores para un parámetro poblacional. (*Vea también* intervalo de confianza.)

Estimador de máxima verosimilitud: Estimador que maximiza la función de (log-)verosimilitud.

Estimador de mínimos cuadrados en dos etapas (MC2E): Estimador de variables instrumentales donde la VI para una variable explicativa endógena se obtiene a partir del valor ajustado de la regresión de la variable explicativa endógena sobre todas las variables exógenas.

Estimador de mínimos cuadrados generalizados (MCG): Estimador que toma en cuenta la estructura conocida de la varianza del error (heterocedasticidad), el patrón de correlación serial en los errores, o ambos, a través de la transformación del modelo original.

Estimador de mínimos cuadrados ponderados (MCP): Estimador usado para ajustar una forma conocida de heterocedasticidad, donde cada residual cuadrado se pondera por la inversa de la varianza (estimada) del error.

Estimador de mínimos cuadrados: Estimador que minimiza una suma de residuales cuadrados.

Estimador de primeras diferencias (PD): En datos de panel, la estimación combinada de MCO aplicada a las primeras diferencias de los datos a través del tiempo.

Estimador de variables instrumentales (VI): Estimador en un modelo lineal usado cuando hay variables instrumentales disponibles para una o más variables explicativas endógenas.

Estimador del método de momentos: Estimador obtenido al usar el análogo muestral de los momentos poblacionales; mínimos cuadrados ordinarios y mínimos cuadrados en dos etapas son ambos estimadores del método de momentos.

Estimador insesgado: Estimador cuyo valor esperado (o media de su distribución de muestreo) es igual al valor poblacional (sin importar el valor poblacional).

Estimador insesgado de varianza mínima: Estimador con la menor varianza en la clase de todos los estimadores insesgados.

Estimador intragrupal (*within*): *Vea* estimador de efectos fijos.

Estimador lineal insesgado: En el análisis de regresión múltiple, un estimador insesgado que es una función lineal de las observaciones de la variable dependiente.

Estimador MCG factibles (MCGF): Procedimiento de MCG donde la varianza o los parámetros de correlación son desconocidos y, por tanto, se deben estimar primero. (*Vea también* estimador de mínimos cuadrados generalizados.)

Estimador no paramétrico de retransformación (estimador *smearing*): Método de retransformación útil particularmente para predecir el nivel de una variable de respuesta cuando un modelo lineal se ha estimado para el logaritmo natural de la variable de respuesta.

Estimador sesgado: Estimador cuya esperanza, o media muestral, es diferente del valor de la población que se debe estimar.

Estrictamente exógeno: Característica de las variables explicativas en un modelo de series de tiempo o datos de panel donde el término de error en cualquier periodo de tiempo tiene una expectativa de cero, condicional en las variables explicativas en todos los periodos; una versión menos restrictiva se establece en términos de correlaciones cero.

Estudio de evento: Análisis econométrico de los efectos de un evento, como un cambio en la regulación gubernamental o la política económica, sobre un resultado.

Evaluación de programa: Análisis de un programa en particular, privado o público, mediante métodos econométricos para obtener el efecto causal del programa.

Exclusión de una variable relevante: En el análisis de regresión múltiple, omitir una variable que tiene un efecto parcial diferente de cero sobre la variable dependiente.

Exogeneidad de los instrumentos: En la estimación de variables instrumentales, el requisito de que una variable instrumental no esté correlacionada con el término de error.

Exogeneidad estricta: Supuesto que se aplica en un modelo de series de tiempo o datos de panel cuando las variables explicativas son estrictamente exógenas.

Experimento: En probabilidad, un término general usado para denotar un evento cuyo resultado es incierto. En el análisis econométrico denota una situación en que los datos se recaban asignando aleatoriamente individuos a grupos de control y tratamiento.

Experimento natural: Situación donde el entorno económico, resumido algunas veces por una variable explicativa, cambia exógenamente y quizá de forma inadvertida, debido a una política o cambio institucional.

Extracción o minería de datos (*data mining*): Práctica de usar la misma base de datos para estimar numerosos modelos en una búsqueda por encontrar el "mejor" modelo.

F

Factor inflacionario de la varianza: En el análisis de regresión múltiple bajo los supuestos de Gauss-Markov, el término en la varianza de muestreo afectado por la correlación entre las variables.

Forma cuadrática: Función matemática donde el argumento vectorial premultiplica y postmultiplica a una matriz cuadrada simétrica.

Forma *R*-cuadrada del estadístico *F*: Estadístico *F* para probar restricciones de exclusión expresadas en términos de las *R*-cuadradas de los modelos restringido y no restringido.

Frecuencia de datos: Intervalo al cual se recolectan los datos de series de tiempo. Anualmente, trimestralmente y mensualmente son las frecuencias de datos más comunes.

Fuertemente dependiente: *Vea* altamente persistente.

Función de densidad de probabilidad (fdp): Función que, para variables aleatorias discretas, proporciona la probabilidad de que la variable aleatoria adopte cada valor; para variables aleatorias continuas, el área bajo la fdp indica la probabilidad de diferentes sucesos.

Función de distribución acumulada (fda): Función que da la probabilidad de que una variable aleatoria sea menor o igual que cualquier número real dado.

Función de log-verosimilitud: Suma de las log-verosimilitudes, donde la log-verosimilitud para cada observación es el log de la densidad de la variable dependiente dadas las variables explicativas; la función de log-verosimilitud se considera como función de los parámetros a estimarse.

Función de pérdida: Función que mide la pérdida cuando un pronóstico difiere del resultado real; los ejemplos más comunes son la pérdida en valor absoluto y la pérdida cuadrada.

Función de regresión muestral (FRM): *Vea* línea de regresión de MCO.

Función de regresión poblacional: *Vea* esperanza condicional.

Función exponencial: Función matemática definida para todos los valores, que tiene una pendiente creciente pero un cambio proporcional constante.

Función lineal: Función donde el cambio en la variable dependiente, dado un cambio de una unidad en una variable independiente, es constante.

Función log: Función matemática, definida sólo para argumentos estrictamente positivos, con una pendiente positiva pero decreciente.

Función logarítmica: Función matemática definida para argumentos positivos que tiene una pendiente positiva, pero decreciente.

Función no lineal: Función cuya pendiente no es constante.

Funciones cuadráticas: Funciones que contienen cuadrados de una o más variables explicativas; capturan los efectos decrecientes o crecientes sobre la variable dependiente.

G

Grados de libertad (*gl*): En el análisis de regresión múltiple, el número de observaciones menos el número de parámetros estimados.
Grados de libertad del denominador: En una prueba *F*, los grados de libertad en el modelo no restringido.
Grados de libertad del numerador: En una prueba *F*, el número de restricciones que se están probando.
Grupo base: Grupo representado por el intercepto general en un modelo de regresión múltiple que incluye variables explicativas binarias.
Grupo de comparación: *Vea* grupo base.
Grupo de control: En una evaluación de programa, el grupo que no participa en el programa.
Grupo de tratamiento: En la evaluación de programas, el grupo que participa en el programa.
Grupo experimental: *Vea* grupo de tratamiento.

H

Heterocedasticidad: La varianza del término de error, dadas las variables explicativas, no es constante.
Heterocedasticidad condicional autorregresiva (ARCH): Modelo de heterocedasticidad dinámica en que la varianza del término de error, dada la información pasada, depende linealmente de los errores cuadrados.
Heterocedasticidad de forma desconocida: Heterocedasticidad que puede depender de las variables explicativas en una forma desconocida y arbitraria.
Heterogeneidad inobservable: *Vea* efecto inobservable.
Hipótesis alternativa: Hipótesis contra la que se compara la hipótesis nula.
Hipótesis nula: En una prueba de hipótesis clásica, se toma esta hipótesis como verdadera y requiere que los datos proporcionen evidencia sustancial contra ella.
Hoja de datos: Software de cómputo usado para ingresar y manipular datos.
Homocedasticidad: Los errores en un modelo de regresión tienen una varianza constante, condicional en las variables explicativas.

I

Identificación: Un parámetro poblacional, o conjunto de parámetros, se puede estimar consistentemente.
Inclusión de una variable irrelevante: Inclusión de una variable explicativa en un modelo de regresión que tiene un parámetro poblacional igual a cero en la estimación de una ecuación por MCO.
Inconsistencia: Diferencia entre el límite en probabilidad de un estimador y el valor del parámetro.
Inconsistente: Conforme el tamaño de la muestra crece, el estimador no converge (en probabilidad) al valor poblacional correcto.
Inferencia estadística: El acto de probar hipótesis acerca de los parámetros poblacionales.
Instrumentos débiles: Variables instrumentales que están sólo ligeramente correlacionadas con la variable o variables explicativas endógenas.
Integrada de orden cero [I(0)]: Proceso estacionario, débilmente dependiente de series de tiempo, que cuando se usa en el análisis de regresión, satisface la ley de los grandes números y el teorema del límite central.
Integrada de orden uno [I(1)]: Proceso de series de tiempo al que se necesita aplicar la primera diferencia con el fin de producir un proceso I(0).
Intercepto: En la ecuación de una recta, el valor de la variable *y* cuando la variable *x* es cero.
Internet: Red global de computadoras que se puede usar para acceder a información y descargar bases de datos.
Intervalo de confianza (IC): Regla usada para construir un intervalo aleatorio de manera que un cierto porcentaje de todos los conjuntos de datos, determinado por el nivel de confianza, produzca un intervalo que contenga el valor de la población.
Intervalo de confianza asintótico: Intervalo de confianza que es aproximadamente válido en muestras grandes.
Intervalo de predicción: Intervalo de confianza para un resultado desconocido de una variable dependiente en un modelo de regresión múltiple.
Intervalo de pronóstico: En pronósticos, intervalo de confianza para un valor futuro aún no realizado de una variable de series de tiempo. (*Vea también* intervalo de predicción.)
Inversa: Para una matriz $n \times n$, su inversa (si existe) es la matriz $n \times n$ para la cual la pre- y post-multiplicación por la matriz original genera la matriz identidad.

L

Ley de las esperanzas iteradas: Resultado de la teoría de probabilidad que relaciona las esperanzas incondicionales y las condicionales.
Ley de los grandes números (LGN): Teorema que dice que el promedio de una muestra aleatoria converge en probabilidad al promedio poblacional; la LGN también aplica a las series de tiempo débilmente dependientes, estacionarias.
Límite en probabilidad: Valor al cual un estimador converge conforme el tamaño de la muestra crece sin límite.
Línea de regresión de MCO: Ecuación que relaciona el valor predicho de la variable dependiente con las variables independientes, donde las estimaciones de los parámetros se han obtenido mediante MCO.
Logaritmo natural: *Vea* función logarítmica.

M

Martingala: Proceso de series de tiempo cuyo valor esperado, dados los resultados anteriores de la serie, simplemente iguala al valor más reciente.
Matriz: Arreglo de números.
Matriz cero: Matriz donde todas las entradas son cero.
Matriz cuadrada: Matriz con el mismo número de filas que de columnas.
Matriz de varianza-covarianza del estimador de MCO: Matriz de las varianzas y covarianzas de muestreo para el vector de los coeficientes de MCO.
Matriz de varianza-covarianza: Para un vector aleatorio, la matriz semidefinida positiva que se define al colocar las varianzas en la diagonal y las covarianzas en las entradas adecuadas fuera de la diagonal.

Matriz diagonal: Matriz con ceros en todas las entradas fuera de la diagonal.

Matriz escalar de varianza-covarianza: Matriz de varianza-covarianza donde todos los términos fuera de la diagonal son cero y los términos de la diagonal son la misma constante positiva.

Matriz idempotente: Matriz (cuadrada) donde la multiplicación de la matriz por sí misma es igual a sí misma.

Matriz identidad: Matriz cuadrada donde todos los elementos de la diagonal son uno y todos los elementos fuera de la diagonal son cero.

Matriz simétrica: Matriz (cuadrada) que es igual a su transpuesta.

MCO: *Vea* mínimos cuadrados ordinarios.

Media: *Vea* valor esperado.

Media independiente: Requisito clave en el análisis de regresión múltiple que dice que el error inobservable tiene una media que no cambia a través de subconjuntos poblacionales, definidos por valores diferentes de las variables explicativas.

Mediana: En una distribución de probabilidad, es el valor donde hay una probabilidad de 50% de estar por debajo del valor y una probabilidad de 50% de estar por encima de él. En una muestra de números es el valor medio después de que los números se han ordenado.

Mediana condicional: Mediana de una variable de respuesta, condicional en algunas variables explicativas.

Medida de bondad de ajuste: Estadístico que resume qué tan bien explica un conjunto de variables explicativas una variable dependiente o variable de respuesta.

Mejor estimador lineal insesgado (MELI): Entre todos los estimadores lineales insesgados, el estimador con la menor varianza. Los estimadores de MCO son MELI, de forma condicional en los valores muestrales de las variables explicativas, bajo los supuestos de Gauss-Markov.

MELI: *Vea* mejor estimador lineal insesgado.

Método Heckit: Procedimiento econométrico utilizado para corregir el sesgo de la selección muestral debido a un truncamiento incidental u otra forma de no aleatoriedad en los datos faltantes.

Método de remuestreo: Técnica para aproximar errores estándar (y distribuciones de estadísticos de prueba) en la que se obtiene una serie de muestras del conjunto de datos originales y se calculan las estimaciones en cada submuestra.

Micronumerosidad: Término introducido por Arthur Goldberger para describir las propiedades de estimadores econométricos de muestras pequeñas.

Mínimas desviaciones absolutas (MDA): Método para estimar los parámetros de un modelo de regresión múltiple basado en la minimización de la suma de los valores absolutos de los residuales.

Mínimos cuadrados ordinarios (MCO): Método para estimar los parámetros de un modelo de regresión lineal múltiple. Las estimaciones de mínimos cuadrados ordinarios se obtienen mediante la minimización de la suma de los residuales cuadrados.

Modelo de coeficiente (pendiente) aleatorio: Modelo de regresión múltiple donde se permite que los parámetros de las pendientes dependan de variables inobservables específicas a cada unidad.

Modelo de corrección de errores: Modelo de series de tiempo en primeras diferencias que también contiene un término de corrección de error, que sirve para que dos series I(1) regresen al equilibrio de largo plazo.

Modelo de ecuaciones simultáneas (MES): Modelo que determina conjuntamente dos o más variables endógenas, donde cada variable endógena puede estar en función de otras variables endógenas, así como de variables exógenas y un término de error.

Modelo de efectos aleatorios: Modelo de datos de panel de efectos inobservables donde se asume que el efecto inobservable no está correlacionado con las variables explicativas en cada periodo de tiempo.

Modelo de efectos fijos: Modelo de datos de panel de efectos inobservables donde se permite que los efectos inobservables estén arbitrariamente correlacionados con las variables explicativas en cada periodo de tiempo.

Modelo de efectos inobservables: Modelo para datos de panel o muestras de agrupamientos donde el término de error contiene un efecto inobservable.

Modelo de elasticidad constante: Modelo donde la elasticidad de la variable dependiente con respecto a una variable explicativa es constante; en regresión múltiple, ambas variables aparecen en forma logarítmica.

Modelo de probabilidad lineal (MPL): Modelo de respuesta binaria donde la probabilidad de respuesta es lineal en sus parámetros.

Modelo de regresión bivariada: *Vea* modelo de regresión lineal simple.

Modelo de regresión censurada: Modelo de regresión múltiple donde la variable dependiente se ha censurado por encima o debajo de cierto límite conocido.

Modelo de regresión censurada normal: El caso especial del modelo de regresión censurada donde el modelo de población subyacente satisface los supuestos clásicos del modelo lineal.

Modelo de regresión lineal múltiple (RLM): Modelo lineal en sus parámetros, donde la variable dependiente es una función de las variables independientes más un término de error.

Modelo de regresión lineal simple: Modelo donde la variable dependiente es una función lineal de una variable independiente única, además de un término de error.

Modelo de regresión normal truncada: Caso especial del modelo de regresión truncada donde el modelo de población subyacente satisface los supuestos del modelo lineal clásico.

Modelo de regresión Poisson: Modelo para una variable dependiente de conteo donde se asume nominalmente que la variable dependiente, condicionada a las variables explicativas, tiene una distribución Poisson.

Modelo de regresión truncada: Modelo de regresión lineal para datos de corte transversal en el cual todo el esquema de muestreo excluye por completo a una parte de la población, con base en los resultados de la variable dependiente.

Modelo de respuesta binaria: Modelo para una variable dependiente binaria (*dummy*).

Modelo de rezagos distribuidos: Modelo de series de tiempo que relaciona la variable dependiente con los valores actuales y pasados de una variable explicativa.

Modelo de rezagos distribuidos finitos (RDF): Modelo dinámico donde se permite que una o más variables explicativas tengan efectos rezagados sobre la variable dependiente.

Modelo de rezagos distribuidos geométricos (o de Koyck): Modelo de rezagos distribuidos infinitos donde los coeficientes de los rezagos disminuyen a una tasa geométrica.

Modelo de rezagos distribuidos infinitos (RDI): Modelo de rezagos distribuidos donde un cambio en la variable explicativa puede tener impacto en la variable dependiente a lo largo de un futuro indefinido.

Modelo de rezagos distribuidos racionales (RDR): Tipo de modelo de rezagos distribuidos infinitos donde la distribución de rezagos depende de relativamente pocos parámetros.

Modelo de variable latente: Modelo donde se asume que la variable dependiente observada está en función de una variable subyacente latente o inobservable.

Modelo de vectores autorregresivos (VAR): Modelo para dos o más series de tiempo donde cada variable está modelada como una función lineal de los valores pasados de todas las variables, más las perturbaciones que tienen media cero, dados todos los valores pasados de las variables obervadas.

Modelo dinámicamente completo: Modelo de series de tiempo donde más rezagos de la variable dependiente o de las variables explicativas no ayudan a explicar la media de la variable dependiente.

Modelo econométrico: Ecuación que relaciona la variable dependiente con un conjunto de variables explicativas y perturbaciones no observadas, donde los parámetros desconocidos de la población determinan el efecto *ceteris paribus* de cada variable explicativa.

Modelo económico: Relación derivada de la teoría económica o de un razonamiento económico menos formal.

Modelo estático: Modelo de series de tiempo donde sólo las variables explicativas contemporáneas afectan a la variable dependiente.

Modelo lineal clásico: Modelo de regresión lineal múltiple bajo el conjunto completo de supuestos del modelo lineal clásico.

Modelo logit: Modelo para variables de respuesta binaria donde la probabilidad de respuesta es la función logit evaluada en una función lineal de las variables explicativas.

Modelo log-log: Modelo de regresión donde la variable dependiente y (al menos una de) las variables explicativas se encuentran en su forma logarítmica.

Modelo log-nivel: Modelo de regresión donde la variable dependiente está en forma logarítmica y las independientes se encuentran en su nivel (original).

Modelo nivel-log: Modelo de regresión donde la variable dependiente se encuentra en su nivel y (al menos una de) las variables independientes están en su forma logarítmica.

Modelo nivel-nivel: Modelo de regresión donde la variable dependiente y las independientes se encuentran en su nivel (original).

Modelo no restringido: En pruebas de hipótesis, el modelo que no tiene restricciones sobre sus parámetros.

Modelo parsimonioso: Modelo con tan pocos parámetros como es posible para capturar cualquier característica deseada.

Modelo poblacional: Modelo, en particular un modelo de regresión lineal múltiple, que describe a una población.

Modelo probit: Modelo para respuestas binarias donde la probabilidad de respuesta es la función fda normal estándar evaluada en una función lineal de las variables explicativas.

Modelo restringido: En pruebas de hipótesis, el modelo que se obtiene después de imponer todas las restricciones requeridas de la hipótesis nula.

Modelo Tobit: Modelo para una variable dependiente que adopta el valor cero con probabilidad positiva, pero que casi continuamente se distribuye a través de valores estrictamente positivos. (*Vea también* respuesta de solución de esquina.)

Modelo verdadero: Modelo poblacional real que relaciona la variable dependiente con las variables independientes, más una perturbación, donde aplica el supuesto de media condicional cero.

Modelos no anidados: Dos (o más) modelos que no se pueden escribir como caso especial del otro mediante la imposición de restricciones en los parámetros.

Muestra aleatoria: Muestra obtenida mediante el muestreo aleatorio de la población especificada.

Muestra de agrupamientos: Muestra compuesta de agrupamientos o grupos naturales que suelen estar compuestos por personas.

Muestra de datos apareados: Muestra donde cada observación está apareada con otra, como en una muestra compuesta por esposo y esposa o un conjunto de dos hermanos.

Muestra no aleatoria: Muestra obtenida de otra forma que mediante el muestreo aleatorio de la población de interés.

Muestra seleccionada: Muestra de datos obtenida no por muestreo aleatorio sino mediante la selección basada en alguna característica observada o inobservable.

Muestreo aleatorio: Esquema de muestreo donde cada observación se elige al azar de la población. En particular, ninguna unidad tiene más probabilidad de salir elegida que cualquier otra, y cada elección es independiente de las demás.

Muestreo estratificado: Esquema de muestreo no aleatorio donde la población se divide primero en varios estratos no superpuestos y exhaustivos, y después las muestras aleatorias se eligen de cada estrato.

Multicolinealidad: Término que se refiere a la correlación entre variables independientes en un modelo de regresión múltiple; suele invocarse cuando algunas correlaciones son "grandes", pero su magnitud real no está bien definida.

Multiplicación escalar: Algoritmo para multiplicar un (número) escalar por un vector o matriz.

Multiplicación matricial: Algoritmo para multiplicar dos matrices concordantes.

Multiplicador de impacto: *Vea* propensión de impacto.

Multiplicador de largo plazo: *Vea* propensión de largo plazo.

N

Nivel de confianza: Porcentaje de muestras en el que se quiere que el intervalo de confianza comprenda el valor poblacional; 95% es el nivel de confianza más común, pero también se utilizan 90% y 99%.

Nivel de significancia: Probabilidad del error Tipo I en pruebas de hipótesis.

No correlacionado asintóticamente: Proceso de series de tiempo en el cual la correlación entre variables aleatorias de dos puntos en el tiempo tiende a cero mientras el intervalo entre ellas aumenta. (*Vea también* débilmente dependientes.)

No correlacionado serialmente: Errores en un modelo de series de tiempo o datos de panel que no están correlacionados uno con otro en el tiempo.

Normalidad asintótica: La distribución de muestreo de un estimador apropiadamente normalizado converge a la distribución normal estándar.

Notación matricial: Notación matemática conveniente, basada en el álgebra matricial, para expresar y manipular el modelo de regresión múltiple.

Número índice: Estadístico que agrega información sobre la actividad económica, como la producción o precios.

O

Observaciones atípicas (*outliers*): Observaciones en un conjunto de datos que son sustancialmente diferentes de la mayoría de los datos, quizá debido a errores o debido a que algunos datos están generados por un modelo diferente que la mayoría de los demás datos.

Observaciones influyentes: *Vea* observaciones atípicas.

Operador suma: Notación, denotada por $\sum$, usada para definir la suma de un conjunto de números.

P

Panel balanceado: Conjunto de datos de panel en el que están disponibles los datos de todos los años (o periodos) para todas las unidades de corte transversal.

Panel no balanceado: Conjunto de datos de panel en el que faltan los datos de todos los años (o periodos) para algunas unidades de corte transversal.

Parámetro: Valor desconocido que describe una relación poblacional.

Parámetro de la pendiente: Coeficiente de una variable independiente en un modelo de regresión múltiple.

Parámetro del intercepto: Parámetro en un modelo de regresión lineal múltiple que proporciona el valor esperado de la variable dependiente cuando todas las variables independientes son iguales a cero.

Parámetros de la forma reducida: Parámetros que aparecen en una ecuación en la forma reducida.

Parámetros estructurales: Parámetros que aparecen en una ecuación estructural.

Pendiente: En la ecuación de una recta, el cambio en la variable y cuando la variable x aumenta en una unidad.

Periodo base: Para números índice, como índices de precios o producción, el periodo contra el cual todos los demás periodos de tiempo se miden.

Perturbación: *Vea* término de error.

Población: Grupo bien definido (de personas, empresas, ciudades, etc.) en el que se enfoca un análisis estadístico o econométrico.

Porcentaje predicho correctamente: En un modelo de respuesta binaria, el porcentaje de veces que la predicción de cero o uno coincide con el resultado real.

Potencia de una prueba: Probabilidad de rechazar la hipótesis nula cuando esta es falsa; el poder depende de los valores de los parámetros poblacionales bajo la alternativa.

Predicción: La estimación de un resultado obtenido al insertar valores específicos de las variables explicativas en un modelo estimado, que suele ser un modelo de regresión múltiple.

Primeras diferencias: Transformación de una serie de tiempo construida al determinar la diferencia entre los periodos de tiempo adyacentes, donde el periodo antecedente se le resta al periodo subsecuente.

Probabilidad de respuesta: En un modelo de respuesta binaria, la probabilidad de que la variable dependiente adopte el valor uno, condicional en las variables explicativas.

Problema de regresión espuria: Problema que surge cuando un análisis de regresión indica una relación entre dos o más procesos de series de tiempo no relacionados, simplemente debido a que cada uno tiene una tendencia, es una serie de tiempo integrada (como una caminata aleatoria), o ambas.

Procedimiento de dos pasos de Engle-Granger: Método de dos pasos para estimar modelos de corrección de errores, donde el parámetro cointegrador se estima en la primera etapa, y los parámetros de corrección de errores se estiman en la segunda.

Proceso autorregresivo de orden uno [AR(1)]: Modelo de series de tiempo cuyo valor actual depende linealmente de su valor más reciente más una perturbación impredecible.

Proceso con tendencia: Proceso de series de tiempo cuyo valor esperado es una función creciente o decreciente del tiempo.

Proceso de promedio móvil de orden uno [MA(1)]: Proceso de series de tiempo generado como una función lineal del valor actual y un valor rezagado de un proceso estocástico de media cero, varianza constante y sin correlación serial.

Proceso de raíz unitaria: Proceso de series de tiempo altamente persistente donde el valor actual es igual al valor del último periodo más una perturbación débilmente dependiente.

Proceso de series de tiempo: *Vea* proceso estocástico.

Proceso estable AR(1): Proceso AR(1) donde el parámetro del rezago es menor que uno en valor absoluto. La correlación entre dos variables aleatorias de la secuencia declina a cero a una tasa geométrica conforme aumenta la distancia entre las variables aleatorias, y por tanto un proceso AR(1) estable es débilmente dependiente.

Proceso estacionario con (o alrededor de una) tendencia: Proceso que es estacionario una vez que se elimina la tendencia en el tiempo; suele ser implícito que la serie sin tendencia es débilmente dependiente.

Proceso estacionario en diferencias: Secuencia de series de tiempo que es I(0) en sus primeras diferencias.

Proceso estacionario: Proceso de series de tiempo donde las distribuciones marginales y conjuntas son invariables en el tiempo.

Proceso estocástico: Secuencia de variables aleatorias indexadas con base en el tiempo.

Proceso no estacionario: Proceso de series de tiempo cuyas distribuciones conjuntas no son constantes a través de diferentes épocas.

Promedio: La suma de n números dividida entre n.

Promedio muestral: Suma de n números dividida entre n; medida de la tendencia central.

Pronóstico condicional: Pronóstico que asume que los valores futuros de algunas variables explicativas se conocen con certidumbre.

Pronóstico de múltiples pasos hacia delante: Pronóstico de series de tiempo para más de un periodo en el futuro.

Pronóstico de un paso hacia delante: Pronóstico de series de tiempo de un solo periodo en el futuro.
Pronóstico incondicional: Pronóstico que no depende de los valores conocidos o supuestos de las variables explicativas futuras.
Pronóstico puntual: Valor pronosticado de un resultado futuro.
Propensión de impacto: En un modelo de rezagos distribuidos, el cambio porcentual inmediato en la variable dependiente, dado un incremento de una unidad en la variable independiente.
Propensión de largo plazo (PLP): En un modelo de rezagos distribuidos, el cambio eventual en la variable dependiente, dado un incremento permanente de una unidad en la variable independiente.
Propiedades asintóticas: Propiedades de los estimadores y estadísticos de prueba que aplican cuando el tamaño de la muestra crece sin límite.
Propiedades de muestra grande: *Vea* propiedades asintóticas.
Prueba consistente: Prueba donde, bajo la hipótesis alternativa, la probabilidad de rechazar la hipótesis nula converge a uno, a medida que la muestra aumenta sin límite.
Prueba de Breusch-Godfrey: Prueba asintóticamente justificada para correlación serial AR(p), donde AR(1) es la más popular; la prueba permite variables dependientes rezagadas así como otros regresores que no sean estrictamente exógenos.
Prueba de Breusch-Pagan: Prueba de heterocedasticidad donde los residuales cuadrados de MCO se regresan sobre las variables explicativas del modelo.
Prueba de Davidson-MacKinnon: Prueba que se utiliza para probar un modelo contra una alternativa no anidada; se puede implementar como una prueba t para los valores ajustados del modelo contrario.
Prueba de Dickey-Fuller (DF): Prueba t de la hipótesis nula de raíz unitaria en un modelo AR(1). (*Vea también* prueba de Dickey Fuller aumentada.)
Prueba de Dickey-Fuller aumentada: Prueba para una raíz unitaria que incluye los cambios rezagados de la variable como regresores.
Prueba de dos colas: Prueba para una alternativa de dos colas.
Prueba de Engle-Granger: Prueba de la hipótesis nula de que dos series de tiempo no están cointegradas; el estadístico se obtiene como el estadístico de Dickey-Fuller usando los residuales de MCO.
Prueba de error de especificación de la regresión (RESET, por sus siglas en inglés): Prueba general para la forma funcional en un modelo de regresión múltiple; es una prueba F de significancia conjunta de los cuadrados, cubos y quizá potencias superiores de los valores ajustados de la estimación inicial de MCO.
Prueba de hipótesis: Prueba estadística de la hipótesis nula, o sostenida, contra una hipótesis alternativa.
Prueba de hipótesis conjunta: Prueba que implica más de una restricción sobre los parámetros de un modelo.
Prueba de hipótesis múltiple: Prueba de una hipótesis nula que implica más de una restricción en los parámetros.
Prueba de una cola: Prueba de hipótesis contra una alternativa de una cola.
Prueba de White: Prueba de heterocedasticidad que implica regresar los residuales cuadrados de MCO sobre los valores ajustados de MCO y sobre los cuadrados de los valores ajustados; en su forma más general, los residuales cuadrados de MCO se regresan sobre las variables explicativas, los cuadrados de las variables explicativas y todas las interacciones no redundantes de las variables explicativas.
Pseudo *R*-cuadrada: Cualquier número de mediciones de la bondad de ajuste para modelos con variables dependientes limitadas.

R

Raíz del error cuadrático medio (RECM): Otro nombre para el error estándar de la regresión en el análisis de regresión múltiple.
Rango de una matriz: Número de columnas linealmente independientes en una matriz.
Razón inversa de Mills: Término que se puede sumar a un modelo de regresión múltiple para eliminar el sesgo de selección muestral.
Razón *t*: *Vea* estadístico t.
***R*-cuadrada:** En un modelo de regresión múltiple, la proporción de la variación muestral total de la variable dependiente que es explicada por la variable independiente.
***R*-cuadrada ajustada:** Medida de bondad de ajuste en el análisis de regresión múltiple que penaliza las variables explicativas adicionales mediante el ajuste de grados de libertad en la estimación de la varianza del error.
***R*-cuadrada barra:** *Vea* *R*-cuadrada ajustada.
***R*-cuadrada descentrada:** *R*-cuadrada calculada sin restar el promedio muestral de la variable dependiente cuando se obtiene la suma total de cuadrados (STC).
***R*-cuadrada poblacional:** En la población, la fracción de la variación en la variable independiente explicada por las variables explicativas.
Región de rechazo: Conjunto de valores de una prueba estadística que lleva a rechazar la hipótesis nula.
Regla de rechazo: En la comprobación de la hipótesis, la regla que determina cuándo se rechaza la hipótesis nula a favor de la hipótesis alternativa.
Regresando: *Vea* variable dependiente.
Regresión a través del origen: Análisis de regresión donde el intercepto se fija en cero; las pendientes se obtienen minimizando la suma de los residuales cuadrados, como se acostumbra.
Regresión auxiliar: Regresión usada para calcular un estadístico de prueba, como los estadísticos de prueba para la heterocedasticidad y la correlación serial, o cualquier otra regresión que no estime el modelo de interés principal.
Regresión de variables binarias: En una base de datos de panel, la regresión que incluye una variable binaria por cada unidad de corte transversal, junto con las variables explicativas restantes. Produce el estimador de efectos fijos.
Regresor: *Vea* variable explicativa.
Relevancia de los instrumentos: En la estimación de variables instrumentales, el requisito de que una variable instrumental ayude a explicar parcialmente la variación en la variable explicativa endógena.
Residual: Diferencia entre el valor real y el valor ajustado (o predicho); existe un residual para cada observación en la muestra, y se usa para obtener una línea de regresión de MCO.

Residuales estudentizados: Residuales calculados al excluir cada observación, una por una por separado, de la estimación, divididos entre la desviación estándar estimada del error.

Respuesta de solución de esquina: Variable dependiente no negativa que es aproximadamente continua sobre valores estrictamente positivos, pero que asume el valor de cero con cierta regularidad.

Restricciones de exclusión: Restricciones que establecen que ciertas variables se excluyen del modelo (o que tienen coeficientes poblacionales iguales a cero).

Restricciones de sobreidentificación: Las condiciones de momentos adicionales que son resultado de tener más variables instrumentales que variables explicativas endógenas en un modelo lineal.

Restricciones múltiples: Más de una restricción en los parámetros en un modelo econométrico.

S

Secuencia de una martingala en diferencia: La primera diferencia de una martingala. Es impredecible (o tiene una media de cero), dados los valores pasados de la secuencia.

Secuencialmente exógeno: Característica de una variable explicativa en modelos de series de tiempo (o datos de panel) donde el término de error en el periodo de tiempo actual tiene una media condicional cero en todas las variables explicativas actuales y pasadas; una versión más débil se expresa en términos de cero correlaciones.

Selección muestral endógena: Selección muestral no aleatoria donde la selección está relacionada con la variable dependiente, ya sea directamente o a través del término de error de la ecuación.

Selección muestral exógena: Selección muestral que depende de variables explicativas exógenas o que es independiente del término de error en la ecuación en cuestión.

Semidefinida positiva: Matriz simétrica en la que todas las formas cuadráticas son no negativas.

Semielasticidad: Cambio porcentual en la variable dependiente, dado un incremento unitario en una variable independiente.

Servicios de búsqueda en línea: Software de cómputo que permite la búsqueda por tema, nombre, título o palabras clave en Internet o bases de datos en Internet.

Sesgado hacia cero: Descripción de un estimador cuya esperanza en valor absoluto es menor que el valor absoluto del parámetro poblacional.

Sesgo: Diferencia entre el valor esperado de un estimador y el valor poblacional que se supone está estimando.

Sesgo asintótico: *Vea* inconsistencia.

Sesgo de atenuación: Sesgo en un estimador que siempre tiende a cero; por tanto, el valor esperado de un estimador con un sesgo de atenuación es menor en magnitud que el valor absoluto del parámetro.

Sesgo de heterogeneidad: Sesgo en MCO debido a la heterogeneidad omitida (o variables omitidas).

Sesgo de selección muestral: Sesgo en el estimador de MCO inducido mediante el uso de datos que surgen de una selección muestral endógena.

Sesgo de simultaneidad: Sesgo que surge de usar MCO para estimar una ecuación en un modelo de ecuaciones simultáneas.

Sesgo de variable omitida: Sesgo que surge en los estimadores de MCO cuando una variable relevante se omite en la regresión.

Sesgo hacia abajo: El valor esperado de un estimador está por debajo del valor poblacional del parámetro.

Sesgo hacia arriba: El valor esperado de un estimador es mayor que el valor del parámetro de la población.

Significancia económica: *Vea* significancia práctica.

Significancia estadística: Importancia de una estimación medida por el tamaño de un estadístico de prueba, por lo general un estadístico *t*.

Significancia general de una regresión: Prueba de significancia conjunta de todas las variables explicativas que aparecen en una ecuación de regresión múltiple.

Significancia práctica: Importancia económica o práctica de una estimación, que se mide por su signo y magnitud, a diferencia de su significancia estadística.

Simultaneidad: Término que significa que al menos una variable explicativa en un modelo de regresión lineal múltiple está determinada conjuntamente con la variable dependiente.

Sobrecontrol: En un modelo de regresión múltiple, incluir variables explicativas que no deben mantenerse fijas cuando se estudia el efecto *ceteris paribus* de una o más de otras variables explicativas; esto puede ocurrir cuando variables que son por sí mismas resultado de una intervención o política se incluyen entre los regresores.

Sobredispersión: En los modelos de una variable de conteo, cuando la varianza es mayor que la media.

Sobreespecificación de un modelo: *Vea* inclusión de una variable irrelevante.

Solución suplente al problema de variables omitidas: Variable proxy que sustituye a una variable omitida inobservable en una regresión de MCO.

Suavización exponencial: Método simple para pronosticar una variable que implica ponderar todos sus resultados previos.

Subespecificación de un modelo: *Vea* exclusión de una variable relevante.

Suma de residuales cuadrados (SRC): En el análisis de regresión múltiple, la suma de los residules cuadrados de MCO de todas las observaciones.

Suma explicada de cuadrados (SEC): Variación muestral total de los valores ajustados en un modelo de regresión múltiple.

Suma residual de cuadrados: *Vea* suma de residuales cuadrados.

Suma total de cuadrados (STC): Variación total muestral en una variable dependiente con respecto al promedio muestral.

Supuesto de la media condicional cero: Supuesto clave utilizado en el análisis de regresión múltiple que afirma que, dado cualquier valor en las variables explicativas, el valor esperado del error es igual a cero. (*Vea* los supuestos RLM.4, ST.3 y ST.3′).

Supuesto de normalidad: Supuesto del modelo lineal clásico que establece que el error (o variable dependiente) tiene una distribución normal, condicional en las variables explicativas.

Supuestos de Gauss-Markov: Conjunto de supuestos (supuestos RLM.1 a RLM.5 o ST.1 a ST.5) bajo los cuales MCO es MELI.

Supuestos del modelo lineal clásico (MLC): Conjunto ideal de supuestos para el análisis de regresión múltiple: para el análisis de corte transversal, supuestos RLM.1 a RLM.6 y para el análisis de series de tiempo, supuestos ST.1 a ST.6. Los supuestos incluyen la linealidad en los parámetros, la no colinealidad perfecta, el supuesto de media condicional cero, la homocedasticidad, la no correlación serial y la normalidad de los errores.

T

Tasa de crecimiento: Cambio proporcional en una serie de tiempo en relación con el periodo anterior. Puede aproximarse como la diferencia en logaritmos o reportarse como un porcentaje.

Tendencia en el tiempo: El valor esperado, expresado como una función del tiempo, de un proceso de series de tiempo que muestra una tendencia.

Tendencia exponencial: Tendencia con una tasa de crecimiento constante.

Tendencia lineal en el tiempo: Tendencia que es una función lineal del tiempo.

Teorema de Gauss-Markov: Teorema que establece que, bajo los cinco supuestos de Gauss-Markov (para modelos de corte transversal y de series de tiempo), el estimador de MCO es MELI (condicional sobre los valores muestrales de las variables explicativas).

Teorema del límite central (TLC): Resultado clave de la teoría de probabilidad que implica que la suma de las variables aleatorias independientes, o incluso de las variables aleatorias débilmente dependientes, cuando se estandariza por su desviación estándar, tiene una distribución que tiende a la normal estándar conforme el tamaño de la muestra aumenta.

Término de error: Variable en una ecuación de regresión simple o múltiple que contiene factores inobservados que afectan la variable dependiente. El término de error también puede incluir errores de medición en las variables dependientes o independientes observadas.

Término de error compuesto: En un modelo de datos de panel, la suma del efecto inobservable constante en el tiempo y el error idiosincrático.

Término de interacción: Variable independiente en un modelo de regresión que es el producto de dos variables explicativas.

Trampa de la variable binaria: Error de incluir demasiadas variables binarias entre las variables independientes; ocurre cuando en el modelo hay un intercepto general y se incluye una variable binaria por cada grupo.

Transformación de efectos fijos: Para los datos de panel, los datos con el tiempo deducido.

Transformación intragrupal (*within*): *Vea* transformación de efectos fijos.

Transpuesta: Para cualquier matriz, la nueva matriz obtenida al intercambiar sus filas y columnas.

Traza de una matriz: Para una matriz cuadrada, la suma de sus elementos diagonales.

Truncamiento incidental: Problema de selección muestral donde una variable, que suele ser la variable dependiente, es observada sólo para ciertos resultados de otra variable.

V

Valor base: Valor asignado al periodo base para construir un número índice; por lo general el valor base es 1 o 100.

Valor crítico: En prueba de hipótesis, el valor contra el cual se compara un estadístico de prueba para determinar si la hipótesis nula se rechaza o no.

Valor esperado: Medida de la tendencia central en la distribución de una variable aleatoria, incluso un estimador.

Valores ajustados: Valores estimados de la variable dependiente cuando los valores de las variables independientes para cada observación están insertos en la línea de regresión de MCO.

Valor-*p*: El menor nivel de significancia al cual se puede rechazar la hipótesis nula. De manera equivalente, el mayor nivel de significancia al cual no se puede rechazar la hipótesis nula.

Variable aleatoria: Variable cuyo resultado es incierto.

Variable aleatoria continua: Variable aleatoria que asume un valor particular cualquiera con cero de probabilidad.

Variable aleatoria de Bernoulli (o binaria): Variable aleatoria que asume los valores cero o uno.

Variable aleatoria discreta: Variable aleatoria que adopta un número de valores finito o infinito contable.

Variable aleatoria estandarizada: Variable aleatoria transformada al restarle su valor esperado y dividir el resultado entre su desviación estándar; la nueva variable aleatoria tiene una media cero y una desviación estándar uno.

Variable aleatoria *F*: Variable aleatoria con una distribución *F*.

Variable aleatoria ji-cuadrada: Variable aleatoria con una distribución ji-cuadrada.

Variable binaria: *Vea* variable dummy.

Variable cualitativa: Variable que describe una característica no cuantitativa de un individuo, empresa, ciudad, etcétera.

Variable de conteo: Variable que asume valores enteros no negativos.

Variable de control: *Vea* variable explicativa.

Variable de respuesta: *Vea* variable dependiente.

Variable dependiente: Variable a explicarse en un modelo de regresión múltiple (y en una variedad de otros modelos).

Variable dependiente binaria (*dummy*): *Vea* modelo de respuesta binaria.

Variable dependiente limitada (VDL): Variable dependiente o de respuesta cuyo rango está restringido de forma importante.

Variable dependiente rezagada: Variable explicativa que es igual a la variable dependiente de un periodo de tiempo anterior.

Variable *dummy* (binaria): Variable que adopta el valor de cero o uno.

Variable endógena rezagada: En un modelo de ecuaciones simultáneas, un valor rezagado de una de las variables endógenas.

Variable exógena: Cualquier variable no correlacionada con el término de error en el modelo de interés.

Variable explicada: *Vea* variable dependiente.

Variable explicativa: En análisis de regresión, la variable que se usa para explicar la variación en la variable dependiente.

Variable explicativa endógena: Variable explicativa en un modelo de regresión múltiple que está correlacionada con el término de error, ya sea debido a una variable omitida, a un error de medición o a la simultaneidad.

Variable explicativa exógena: Variable explicativa que no está correlacionada con el término de error.

Variable independiente: *Vea* variable explicativa.

Variable instrumental (VI): En una ecuación con una variable explicativa endógena, una VI es una variable que no aparece en la ecuación, no está relacionada con el error de la ecuación y se correlaciona (parcialmente) con la variable explicativa endógena.

Variable nominal: Variable medida en dólares nominales o corrientes.

Variable ordinal: Variable donde el orden de los valores transmite información, pero la magnitud de los valores no.

Variable predeterminada: En un modelo de ecuaciones simultáneas, una variable endógena rezagada o una variable exógena rezagada.

Variable predicha: *Vea* variable dependiente.

Variable predictora: *Vea* variable explicativa.

Variable proxy: Variable observada relacionada, pero no idéntica a una variable explicativa inobservable en el análisis de regresión múltiple.

Variable real: Valor monetario medido en términos de un periodo base.

Variables aleatorias independientes: Variables aleatorias cuya distribución conjunta es el producto de las distribuciones marginales.

Variables aleatorias no correlacionadas por pares: Conjunto de dos o más variables aleatorias donde cada par no está correlacionado.

Variables aleatorias no correlacionadas: Variables aleatorias que no están relacionadas linealmente.

Variables binarias anuales: Para bases de datos con un componente de series de tiempo, las variables binarias son iguales a uno en el año relevante y a cero en todos los demás años.

Variables binarias estacionales: Conjunto de variables binarias usado para denotar los trimestres o meses del año.

Variables endógenas: En modelos de ecuaciones simultáneas, las variables que están determinadas por las ecuaciones en el sistema.

Variables omitidas: Una o más variables, las cuales nos hubiera gustado controlar, se han omitido en la estimación de un modelo de regresión.

Varianza: Medida de la dispersión de la distribución de una variable aleatoria.

Varianza asintótica: Cuadrado del valor entre el que se debe dividir un estimador con el fin de obtener una distribución normal estándar asintótica.

Varianza condicional: Varianza de una variable aleatoria, dada una o más de otras variables aleatorias.

Varianza de muestreo: Varianza en la distribución de muestreo de un estimador; mide la dispersión de la distribución de muestreo.

Varianza del error de predicción: Varianza en el error que surge cuando se predice el valor futuro de una variable dependiente con base en una ecuación de regresión múltiple estimada.

Varianza del error: Varianza del término de error en un modelo de regresión múltiple.

Varianza muestral: Estimador consistente e insesgado de la varianza poblacional.

Vector aleatorio: Vector que consiste en variables aleatorias.

Vector columna: Vector de números ordenados como una columna.

Vector fila: Vector de números ordenados como una fila.

Vectores linealmente independientes: Conjunto de vectores tal que ningún vector se puede escribir como una combinación lineal de los demás vectores del conjunto.

Índice

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Z

NOTAS

NOTAS

NOTAS