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 simulta-

neidad (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

$$colGPA = \beta_0 + \beta_1 alcohol + \beta_2 hsGPA + \beta_3 SAT + \beta_4 female + u \qquad \textbf{19.1}$$

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" (^) 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

$$\begin{aligned} thefts_t = \beta_0 + \beta_1 unem_t + \beta_2 unem_{t-1} + \beta_3 cars_t \\ + \beta_4 convrate_t + \beta_5 convrate_{t-1} + u_t, \end{aligned} \qquad \textbf{19.2}$$

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

$$\log(div/pop) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc/pop) + otros\ factores, \tag{19.3}$$

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

$$\log(div) = \gamma_0 + \gamma_1 mlb + \gamma_2 perCath + \gamma_3 \log(inc) + \gamma_4 \log(pop) + otros\ factores. \tag{19.4}$$

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,

$$\widehat{\log(salary)} = 2.45 + .236 \log(sales) + .008\, roe + .061\, ceoten$$
$$(0.93)\quad (.115)\qquad\qquad (.003)\qquad (.028)$$
$$n = 204,\ R^2 = .351,$$

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.