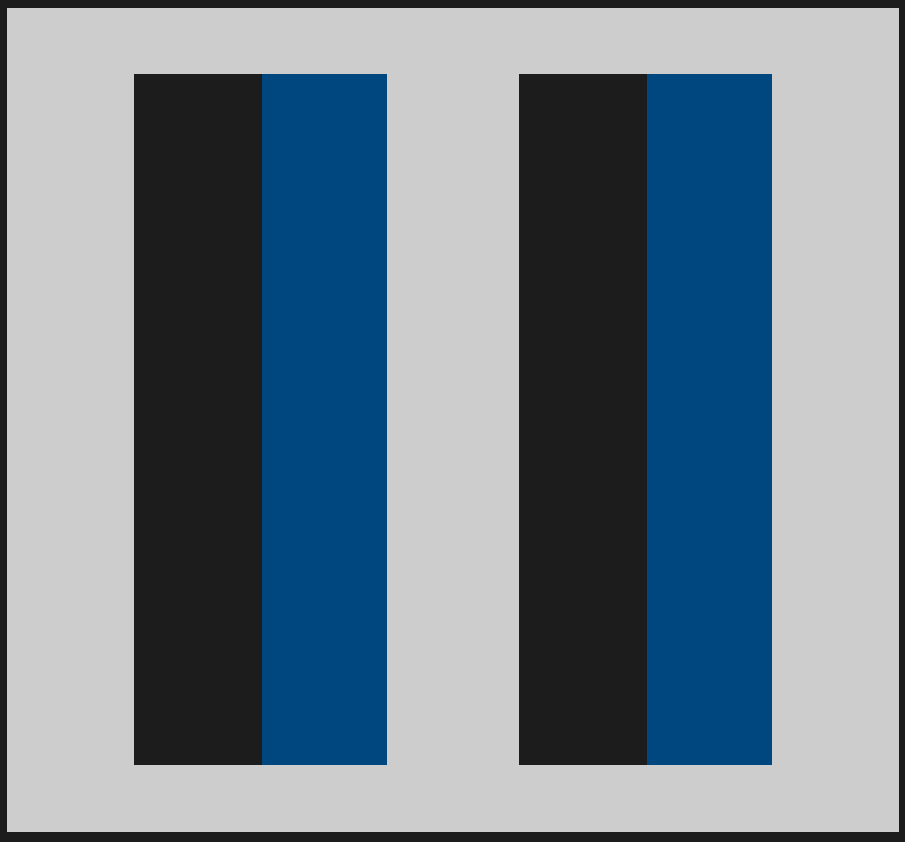
INTRODUCCIÓN A LA ECONOMETRÍA

UN PRIMER CONTACTO

JOSÉ ALBERTO MAURICIO



ucm.randomshock.com

Econometric techniques are usually developed and employed for answering practical questions. As the first five letters of the word "econometrics" indicate, these questions tend to deal with economic issues, although applications to other disciplines are widespread. The economic issues can concern macroeconomics, international economics, and microeconomics, but also finance, marketing, and accounting. The questions usually aim at a better understanding of an actually observed phenomenon and sometimes also at providing forecasts for future situations. Often it is hoped that these insights can be used to modify current policies or to put forward new strategies.

P.H. FRANSES
A Concise Introduction to Econometrics

Decision making in business and economics is often supported by the use of quantitative information. Econometrics is concerned with summarizing relevant data information by means of a model. Such econometric models help to understand the relation between economic and business variables and to analyse the possible effects of decisions.

C. HEIJ, P. DE BOER, P.H. FRANSES, T. KLOEK Y H.K. VAN DIJK
Econometric Methods with Applications in Business and Economics

Econometrics is used in all applied economics fields to test economic theories, to inform government and private policymakers, and to predict economic time series. Sometimes, an econometric model is derived from a formal economic model, but in other cases, econometric models are based on informal economic reasoning and intuition.

J.M. WOOLDRIDGE
Introductory Econometrics - A Modern Approach

Social scientists and policymakers alike seem driven to draw sharp conclusions, even when these can be generated only by imposing much stronger assumptions than can be defended. We need to develop a greater tolerance for ambiguity. We must face up to the fact that we cannot answer all of the questions that we ask.

C.F. MANSKI - *Identification Problems in the Social Sciences*
Citado en B.H. BALTAGI - *Econometrics*

Intro-Ectr.pdf

ucm.randomshock.com

Versión 5.0 - Julio 2021

CONTENIDO

PRÓLOGO

Las maneras de impartir y de recibir las enseñanzas universitarias correspondientes a un primer cuatrimestre de Econometría, han cambiado notablemente desde el curso académico 1986-87 (lo cual es bastante lógico, teniendo en cuenta que han transcurrido treinta y cinco años desde entonces ...), cuando algunos estudiantes de cuarto curso de la licenciatura en Ciencias Económicas y Empresariales de la UCM nacidos en torno a 1965 tuvimos nuestro primer contacto serio con ella. No obstante, el contenido de la materia sigue siendo esencialmente el mismo desde entonces (seguramente, incluso, desde mucho antes): métodos (teoría) y aplicaciones (práctica) del Análisis de Regresión Lineal con datos de sección cruzada y con datos de series temporales, con algunos añadidos importantes más recientes dedicados en particular al análisis de datos de series temporales no estacionarias.

Los cambios en las maneras de impartir y de recibir la docencia han estado motivados, sobre todo, por las enormes posibilidades asociadas con el desarrollo vertiginoso de la informática y las comunicaciones. Por ejemplo, en la actualidad es extraordinariamente sencillo acceder a grandes colecciones de datos y analizarlos cómoda y eficientemente con un ordenador portátil de gran potencia en prácticamente cualquier ubicación física. También es muy sencillo y rápido el intercambio de información de todo tipo entre quienes imparten la docencia y quienes la reciben, más allá del tradicional intercambio presencial que sigue siendo muy valioso y productivo.

Por su parte, el hecho de que el contenido de la materia en un primer cuatrimestre de Econometría siga siendo esencialmente el mismo que hace treinta y cinco años, está razonablemente justificado por la importancia del Análisis de Regresión Lineal no sólo como instrumento de gran utilidad por sí mismo, sino también como herramienta en la que se fundamenta gran parte de los métodos econométricos utilizados en la actualidad.

En cualquier caso, quizás lo realmente importante en un primer curso de Econometría no

tenga que ver del todo ni con el contenido de la materia ni con la forma de impartirla o de estudiarla, sino más bien con intentar explicar y entender claramente desde el principio del curso qué es la Econometría. La idea central a este respecto es muy simple: la Econometría es una **herramienta** de propósito general que puede contribuir a **resolver** muchos **problemas prácticos**, mediante la interpretación y el uso adecuado de la información contenida en una colección de **datos** resumida a través de uno o varios **modelos**.

Probablemente, la dificultad principal con la que nos hemos encontrado estudiantes y profesores de Econometría durante al menos los últimos treinta y cinco años, ha tenido que ver con no percibir ni transmitir claramente esta idea tan sencilla desde el principio del curso. En muchas ocasiones, nos hemos limitado a explicar y a entender la Econometría como un conjunto variado de fórmulas y técnicas dirigidas cada una por su cuenta, de forma inconexa, al análisis de distintos tipos de datos económicos o financieros, en vez de como un intento unificador y riguroso de analizar, resumir e interpretar la información no experimental que se puede obtener mediante la mera observación de fenómenos y entornos muy diversos de carácter tanto social como natural.

Con la presente *Introducción a la Econometría* se pretende ofrecer suficiente información (más de la concebida inicialmente, de hecho) para dejar lo más claras posible, desde el comienzo de un primer curso de Econometría, las ideas mencionadas en los dos párrafos anteriores. Con estas ideas bien transmitidas y entendidas, se espera que la parafernalia técnica (matemática, estadística e informática) asociada con un primer contacto con la Econometría teórica y aplicada, resulte fácilmente explicable y entendible dentro de un panorama amplio y progresivamente familiar en el que se pueden ir detallando, poco a poco y con sentido, todas las piezas que suelen conformar ese primer curso.

J.A.M.
Julio de 2021

D.AM.H.

Observación: Los archivos de datos (WF1) en el formato del programa EViews que se mencionan a lo largo del texto pueden obtenerse como se indica en el Apéndice A.1.

INTRODUCCIÓN A LA
ECONOMETRÍA
UN PRIMER CONTACTO

Wanderers in the land of Osten Ard are cautioned not to put blind trust in old rules and forms, and to observe all rituals with a careful eye, for they often mask being with seeming.

T. WILLIAMS
Memory, Sorrow and Thorn - The Dragonbone Chair

A pesar de sus "viejas reglas" y "rituales" y de lo que el término sugiere, la **econometría** no trata únicamente de cómo medir o evaluar teorías y relaciones económicas. Ante todo, la econometría moderna es una **herramienta** de propósito bastante general que puede contribuir a **resolver** una amplia gama de **problemas prácticos** mediante análisis de **datos**.

Las páginas siguientes están dedicadas en parte a presentar un muestrario breve de dichos problemas. También se describen con cierto detalle la metodología que se emplea en la práctica para contribuir a resolverlos y el tipo de soluciones que la econometría puede ofrecer en ámbitos variados. En conjunto, en estas páginas se presenta, desde un punto de vista introductorio, una panorámica global de la econometría a partir de la que podría desarrollarse el contenido de un primer curso sobre esta disciplina. Disponer de una visión global desde el principio puede resultar útil para entender, desde una perspectiva general, cuál es el interés específico de cada uno de los temas que suelen configurar un primer curso típico de econometría.

1 La Naturaleza de la Econometría

La **toma de decisiones** en economía y en finanzas, así como en muchos otros contextos de carácter tanto social como natural, se basa a menudo en la información cuantitativa contenida en unos **datos**. Los datos suelen reflejar, con mayor o menor precisión, ciertos aspectos del funcionamiento real de un **sistema** económico, financiero o de algún otro tipo,

cuyo entendimiento es esencial para tomar razonadamente una decisión entre varias alternativas posibles. En ocasiones, entender cuál es el funcionamiento de un sistema que aparece reflejado en unos datos es un asunto meramente científico, en el sentido de que, a veces, tan sólo se pretende **inferir** a partir de los datos y **divulgar** las pautas esenciales del funcionamiento del sistema considerado.

La **econometría aplicada** trata de cómo analizar datos para responder a preguntas diversas referidas a sistemas cuyo funcionamiento es, en buena medida, estocástico o aleatorio, es decir, imposible de caracterizar con total exactitud o de prever con absoluta certeza. Por este motivo, muchos de los métodos que se utilizan en la econometría aplicada para analizar datos están tomados directamente o adaptados de la estadística. Otros son métodos desarrollados específicamente para analizar ciertos tipos de datos que son especiales por algún motivo. En conjunto, todos estos métodos conforman lo que se denomina **econometría teórica**, teoría econométrica o métodos econométricos.

En esta Sección 1 se ofrece una panorámica general acerca de las **preguntas** que puede contribuir a responder la econometría aplicada, los tipos de **datos** que suelen emplearse para intentar responderlas, los **métodos** que se utilizan para analizar dichos datos y, en última instancia, las **respuestas** que pueden darse a las preguntas planteadas.

1.1 Preguntas

El punto de partida de cualquier análisis econométrico aplicado consiste en plantear con precisión una **pregunta concreta** sobre algún aspecto de un sistema estocástico, cuya **respuesta** se pretende obtener usando la **evidencia empírica** contenida en una colección de datos. Aunque hay otras posibilidades, muchas preguntas que se plantean en la econometría aplicada tratan sobre la **evaluación de efectos causales** entre determinadas **variables** y sobre la **previsión** de cantidades desconocidas.

1.1.1 Ejemplo - Algunas Preguntas Analizables con Datos

En este ejemplo se enuncian varios problemas prácticos potencialmente resolubles mediante análisis de datos. Cada problema va acompañado por una sugerencia sobre los datos que podrían emplearse para intentar resolverlo. EJERCICIO 1

P01: Evaluar la eficacia de un fertilizante. **Datos:** Rendimiento de varias zonas de cultivo y cantidad empleada de fertilizante en cada una de ellas.

P02: Evaluar el efecto de la asistencia a clase sobre las notas finales. **Datos:** Notas finales y horas de asistencia a clase de varios alumnos.

P03: Evaluar el efecto del consumo de tabaco durante el embarazo sobre el peso de un recién nacido. **Datos:** Peso de varios recién nacidos y consumo diario de cigarrillos por parte de sus madres durante el embarazo.

P04: Evaluar la posible discriminación salarial por motivos de género. **Datos:** Salario, años de educación, años de experiencia y género de varias personas trabajadoras.

P05: Estimar el precio de venta de una vivienda de segunda mano en función de sus características. **Datos:** Precio de venta, superficie, número de habitaciones, estado de conservación y localización de varias viviendas puestas recientemente a la venta.

P06: Evaluar el efecto del gasto anual en publicidad sobre las ventas de una empresa. **Datos:** Volumen de ventas y gasto en publicidad referidos a varios años consecutivos.

P07: Describir la inercia observada en la evolución del crecimiento anual del PIB real y preverlo a corto plazo. **Datos:** Variación anual del PIB real en varios años consecutivos.

P08: Analizar y prever la rentabilidad de algunos valores de la deuda pública a corto plazo. **Datos:** Tipos de interés de las operaciones a seis meses y a tres meses en el mercado secundario de la deuda pública, referidos a varios trimestres consecutivos.

P09: Evaluar la relación entre el gasto en alimentación de una familia y su renta disponible. **Datos:** Gasto en alimentación y renta disponible de varias familias.

P10: Evaluar el efecto de la educación sobre los salarios. **Datos:** Salarios, años de educación y otras características laborales y personales de varios trabajadores.

■

Quizás las preguntas más difíciles de responder en la econometría aplicada son las que tratan sobre cómo evaluar algún **efecto causal** (como P01, P02, P03, P06 y P10 en el Ejemplo 1.1.1), es decir, cómo aislar y cuantificar la **influencia directa** o **parcial** de una variable sobre otra. La dificultad principal tiene que ver con que, en muchas ocasiones, los datos disponibles para intentar evaluar efectos causales son de tipo **observacional** o **no experimental**, es decir, son datos que no se han obtenido mediante la ejecución activa de

un **experimento controlado** diseñado específicamente para provocar, aislar y medir una reacción o una influencia, sino mediante la mera observación pasiva de un sistema dado. En estas condiciones, el problema reside en que las variables sobre las que se pretende evaluar algún efecto causal (el rendimiento de una zona de cultivo, la nota final de un alumno, el peso de un recién nacido, las ventas de una empresa, o el salario de un trabajador) suelen estar sometidas a un número elevado de influencias relacionadas entre sí, muchas de ellas no controlables por el analista (investigador), y algunas difíciles o imposibles de observar. Evaluar fiablemente el efecto aislado de una influencia dada (la cantidad empleada de un fertilizante, las horas de asistencia a clase, el consumo de tabaco durante el embarazo, el gasto en publicidad, o los años de educación) puede resultar especialmente difícil si la influencia que se pretende aislar está relacionada con otras que no se consideran, por imposibilidad o por error, explícitamente en el análisis.

1.1.2 Definición - Efectos Causales

A continuación se define con precisión un efecto causal y se describen las dificultades asociadas con la evaluación de efectos causales usando datos no experimentales. La posibilidad de ejecutar un experimento controlado elimina, en general, dichas dificultades.

En preguntas como P01 o P02 en el Ejemplo 1.1.1, se trata de cómo evaluar el efecto causal de una variable X (la cantidad empleada de un fertilizante, o el número de horas de asistencia a clase) sobre otra variable Y (el rendimiento de una zona de cultivo, o la nota final de un alumno en cierta asignatura). Todas las influencias posibles que recibe Y se pueden resumir mediante una expresión matemática (una función) del tipo

$$Y = F(X, W, V), \qquad [1]$$

donde W representa las influencias observables que no están recogidas en X, mientras que V recoge todas las influencias sobre Y difíciles o imposibles de observar. EJERCICIO 2

En relación con [1], el **efecto causal** (directo, parcial o "ceteris paribus") de X sobre Y puede definirse como la **variación** o la **respuesta** de Y **relativa** a una **variación** en X ($\Delta X \neq 0$) cuando W y V permanecen **constantes** ($\Delta W = \Delta V = 0$). Cualquier variación que tiene lugar en Y se puede descomponer, de acuerdo con [1], como

$$\Delta Y \cong [\, F_X \times \Delta X \,] + [\, F_W \times \Delta W \,] + [\, F_V \times \Delta V \,], \qquad [2]$$

donde F_X, F_W y F_V son las **derivadas parciales** de la función $F(X, W, V)$ con respecto a

cada uno de sus tres argumentos. La derivada F_X en [2] representa el efecto causal de X sobre Y en el sentido de que $\Delta Y/\Delta X \cong F_X$ cuando $\Delta X \neq 0$ y $\Delta W = \Delta V = 0$. La utilidad de una representación matemática como [1]-[2] reside justamente en que permite considerar de forma aislada un efecto de este tipo, a pesar de que en la práctica puede ser difícil o imposible observar qué ocurre cuando X varía ($\Delta X \neq 0$) a la vez que W y V permanecen constantes ($\Delta W = \Delta V = 0$).

En la práctica, cuando las influencias W y V en [1] presentan algún tipo de relación sistemática con X [por ejemplo, si la irrigación (W) y la calidad del suelo (V) de una zona de cultivo están relacionadas con la cantidad empleada de un fertilizante (X), o si el número de horas de estudio fuera de clase (W) y el grado de interés por una asignatura (V) están relacionados con el número de horas de asistencia a clase (X)], cualquier variación en X tiene asociadas ciertas variaciones en W y en V, que pueden representarse, respectivamente, como ΔW_X y ΔV_X. En este caso, la **respuesta total** de Y frente a una variación de cuantía ΔX en X puede escribirse, de acuerdo con [2], como

$$\Delta Y_X \cong \overbrace{\underbrace{F_X \times \Delta X}_{\text{Respuesta Directa}} + \underbrace{F_W \times \Delta W_X + F_V \times \Delta V_X}_{\text{Respuesta Indirecta}}}^{\text{Respuesta Total}},$$

o bien, de manera más explícita, como

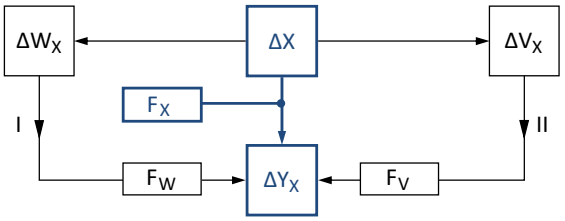
$$\Delta Y_X \cong \left[\overbrace{\underbrace{F_X}_{\text{Efecto Directo}} + \underbrace{F_W \times \frac{\Delta W_X}{\Delta X}}_{\text{Efecto Indirecto I}} + \underbrace{F_V \times \frac{\Delta V_X}{\Delta X}}_{\text{Efecto Indirecto II}}}^{\text{Efecto Total}} \right] \Delta X . \qquad [3]$$

En la Figura 1 de la página siguiente están representados el **efecto directo** (con trazo grueso) y los **efectos indirectos** (con trazos delgados) considerados en [3].

Tanto la expresión [3] como la Figura 1 ponen de manifiesto lo siguiente: la variación total relativa de Y asociada con X, $\Delta Y_X/\Delta X$, representa el efecto causal o directo F_X de X sobre Y sólo cuando (i) $F_W = F_V = 0$, de manera que ni W ni V influyen directamente sobre Y (una posibilidad, en general, más bien remota y, por lo tanto, descartable), o bien cuando (ii) $\Delta W_X = \Delta V_X = 0$, de manera que ni W ni V varían sistemáticamente con X. Desde el punto de vista de un análisis econométrico aplicado, esto significa que el grado de asociación (relativa) entre Y y X observado en una colección de datos sobre dichas

FIGURA 1

Efectos Directo e Indirectos de X sobre Y en la Ecuación [3]



variables, representará un efecto causal sólo si las influencias sobre Y que no se consideran de forma explícita en el análisis **no** están **relacionadas** con X. Por el contrario, si existen relaciones sistemáticas entre W y X, o entre V y X, entonces el grado de asociación observado entre Y y X puede deberse tan sólo a la existencia de dichas relaciones (es decir, a los efectos indirectos en [3], representados con trazos delgados en la Figura 1) y no a la existencia de un efecto causal de X sobre Y (el efecto directo en [3], representado con un trazo grueso en la Figura 1). En tal caso, suele decirse que la relación observada entre Y y X es una **relación espuria**, en el sentido de que dicha relación es tan sólo aparente, carente de legitimidad o de autenticidad.

■

1.1.3 Ejemplo - Cómo Obtener Datos Experimentales

En la práctica, para garantizar la independencia de W (por ejemplo, irrigación o número de horas de estudio fuera de clase) y de V (por ejemplo, calidad del suelo o grado de interés por una asignatura) con respecto a X (cantidad empleada de un fertilizante o número de horas de asistencia a clase), se podría intentar obtener datos sobre Y y X ejecutando alguno de los dos experimentos siguientes: EJERCICIO 3

Experimento I. **Paso 1:** Escoger una colección de datos sobre X tal que los valores correspondientes de W y de V sean idénticos en todos los casos. **Paso 2:** Observar los valores correspondientes de Y. Esta estrategia garantiza que W y V no varían en absoluto en los casos considerados ($\Delta W = \Delta V = 0$), por lo que cualquier asociación observada entre Y y X representa ciertamente el efecto parcial o "ceteris paribus" de X sobre Y. El problema asociado con esta estrategia reside en que es muy difícil asegurar que W y V tengan exactamente los mismos valores en todos los casos, especialmente en relación con

las influencias difíciles o imposibles de observar contenidas en V.

Experimento II. **Paso 1:** Asignar a X una colección de valores predeterminados sin tener en cuenta (ignorando o independientemente de) los valores correspondientes de W y de V (por ejemplo, asignando valores a X de acuerdo con una secuencia de números aleatorios generados por un ordenador). **Paso 2:** Observar los valores correspondientes de Y. Esta estrategia garantiza que W y V varían independientemente de X ($\Delta W_X = \Delta V_X = 0$), por lo que (igual que en el Experimento I anterior) cualquier asociación observada entre Y y X representa el efecto directo o causal de X sobre Y. El problema asociado con esta estrategia reside en que asignar a X unos valores controlados o decididos de antemano puede resultar impracticable por diversos motivos (por ejemplo, porque dicha asignación sea prohibitivamente costosa, moralmente inaceptable o simplemente imposible).

Cuando no es posible obtener datos a través de alguna de estas estrategias, la evaluación fiable en la práctica de efectos causales puede resultar difícil, especialmente si no se tiene confianza en que las influencias sobre Y que no se consideran de forma explícita son suficientemente independientes de los datos observados sobre X. Por este motivo, dos **elementos centrales** en el diseño y la elaboración de cualquier análisis econométrico aplicado son los que tienen que ver con qué **variables** se **incluyen** explícitamente en el análisis y cuál es su posible **relación** con otras variables que se **omiten**.

Actualmente, la econometría puede emplearse para contribuir a resolver problemas muy variados, que no se limitan obligatoriamente a cuestiones causales de tipo económico o financiero. No obstante, esta amplitud de posibilidades sí obliga a plantear el problema que se pretende resolver en cada caso de la manera más concreta y específica posible. Esto es imprescindible para decidir adecuadamente qué **datos** utilizar (de qué tipo, sobre qué colectivo, o sobre qué período temporal) y qué **métodos** emplear para analizarlos. Sin un planteamiento inicial preciso sobre el objetivo y el interés de un análisis, es prácticamente imposible saber siquiera por dónde empezar.

1.2 Datos

Cada dato que se emplea en un análisis econométrico aplicado es un valor numérico de cierta característica de una **entidad observable** (un individuo, una organización, un objeto

TABLA 1

Datos sobre Algunas Características de 80 Viviendas Unifamiliares Vendidas en el Área Metropolitana de Boston en 1990

Número de Observación	Precio de Venta (miles de dólares)	Superficie (metros cuadrados)	Número de Dormitorios
1	335.0	226.5	4
2	405.0	192.9	3
3	226.0	127.6	3
⋮	⋮	⋮	⋮
78	365.5	203.1	3
79	281.0	179.1	4
80	260.0	120.2	3

o un lugar, en el sentido más amplio de cualquiera de estos términos) que forma parte del sistema social o natural al que se refiere la pregunta planteada al comienzo del análisis. En función de cómo se haya llevado a cabo la observación de dicho sistema, los datos resultantes pueden ser, al menos, de tres tipos (con algunas variaciones dentro de cada uno de ellos): **datos de sección cruzada** o transversales, **datos de series temporales** y **datos de panel** o longitudinales. Cada uno de estos tipos de datos resulta adecuado para resolver unas cuestiones determinadas, por lo que la pregunta que se plantea al comienzo de un análisis suele indicar el tipo de datos que es relevante en cada caso.

1.2.1 Definición - Datos de Sección Cruzada o Transversales

Una **sección cruzada** es una colección de datos sobre una o varias características comunes de distintas entidades observables en un momento dado.

1.2.2 Ejemplo - Una Sección Cruzada de Viviendas Unifamiliares

La Tabla 1 contiene parte de una colección de datos de sección cruzada que se podría utilizar para intentar resolver un problema como P05 en el Ejemplo 1.1.1. La sección cruzada completa se encuentra en el archivo SC01-VIVIENDAS.WF1.

La Tabla 1 está organizada de manera que cada fila se refiere a una entidad observada (una vivienda) y cada columna a una característica (precio de venta, superficie y número de dormitorios). Aunque conviene asignar a cada entidad observada un número de orden (como en la primera columna de la Tabla 1), el orden en el que están dispuestas las

TABLA 2

Datos sobre Ventas y Gasto en Publicidad Anuales de la Empresa Lydia E. Pinkham's Medicine Co. desde 1907 hasta 1960

Número de Observación	Fecha (año)	Volumen de Ventas (miles de dólares)	Gasto en Publicidad (miles de dólares)
1	1907	1016	608
2	1908	921	451
3	1909	934	529
⋮	⋮	⋮	⋮
53	1959	1387	644
54	1960	1289	564

observaciones es, en general, irrelevante para el análisis.

En muchas ocasiones, una sección cruzada se refiere a un grupo de entidades observables que es tan sólo un subconjunto de un colectivo más amplio. Por este motivo, una sección cruzada suele interpretarse como una **muestra aleatoria** procedente de una **población** bien definida. Si en una sección cruzada están suficientemente bien representadas todas las entidades observables de una población, entonces puede esperarse que las conclusiones obtenidas del análisis de dicha sección cruzada (muestra) sean aplicables a todo el colectivo (población). EJERCICIO 4

Los datos de sección cruzada suelen emplearse para investigar posibles relaciones entre ciertas características o variables de una población examinando las diferencias observadas entre distintas entidades de una muestra en un momento dado.

1.2.3 Definición - Datos de Series Temporales

Una **serie temporal** es una secuencia de datos ordenados cronológicamente sobre una o varias características de una única entidad observable en diferentes momentos.

1.2.4 Ejemplo - Series Temporales de Ventas y Publicidad

La Tabla 2 contiene parte de una colección de datos de series temporales que podría utilizarse para intentar resolver un problema como P06 en el Ejemplo 1.1.1. Las series completas se encuentran en el archivo ST01-PINKHAM.WF1.

La Tabla 2 está organizada de manera que cada fila se refiere a una fecha (un año) y cada columna a una característica (volumen de ventas y gasto en publicidad) de la entidad observada (una empresa). A diferencia de lo que suele ocurrir con una sección cruzada, el orden en el que figuran los datos en una serie temporal es crucial para detectar posibles inercias o dinámicas en la evolución de las características a las que se refieren los datos.

Una serie temporal está asociada con un período muestral que es sólo una parte de la historia de la entidad considerada. En este sentido, una serie temporal suele interpretarse como una **muestra ordenada** (no aleatoria) extraída de un **proceso estocástico** (desarrollo histórico) bien definido. Si las circunstancias sociales o naturales del período muestral al que se refiere la serie temporal considerada se mantienen relativamente estables después de dicho período, entonces puede esperarse que las conclusiones obtenidas del análisis de dicha serie sean aplicables también a momentos posteriores, al menos a corto plazo.

El análisis de series temporales permite investigar posibles **relaciones dinámicas** entre variables examinando las variaciones recogidas en los datos entre momentos consecutivos de la historia de una entidad observable. Adicionalmente, el análisis de series temporales se emplea con mucha frecuencia para **prever** la evolución futura de variables económicas, financieras, y de muchos otros tipos. EJERCICIO 5

En ocasiones, puede resultar útil analizar información que combine datos de sección cruzada con datos de series temporales. En particular, el análisis de una combinación de distintas secciones cruzadas referidas a la misma población en diferentes momentos de su historia, puede resultar útil para evaluar cómo han variado con el paso del tiempo ciertas características de la población considerada (debido, por ejemplo, a la implantación de nuevas políticas o a la ocurrencia de sucesos especiales). Cuando las entidades observadas son exactamente las **mismas** en cada momento considerado, la colección de datos correspondiente se denomina una colección de **datos de panel**.

1.2.5 Definición - Datos de Panel

Una colección de **datos de panel** (o un **panel de datos**) es una secuencia de datos ordenados cronológicamente sobre varias características de las **mismas** entidades observables en diferentes momentos.

TABLA 3

Una Combinación de Dos Secciones Cruzadas de
550 Trabajadores en 1978 y 534 Trabajadores en 1985

Número de Observación	Fecha (año)	Salario Medio Anual (dólares por hora)	Educación (años)	Experiencia (años)	Género
1	1978	3.37	12	8	Hombre
2	1978	5.00	12	30	Mujer
3	1978	8.50	6	38	Hombre
⋮	⋮	⋮	⋮	⋮	⋮
550	1978	12.50	15	12	Hombre
551	1985	9.00	10	27	Hombre
552	1985	5.50	12	20	Hombre
553	1985	3.80	12	4	Mujer
⋮	⋮	⋮	⋮	⋮	⋮
1084	1985	19.47	12	9	Hombre

TABLA 4

Un Panel de Datos de 545 Trabajadores en 1980 y 1981

Número de Observación	Trabajador	Fecha (año)	Salario Medio Anual (dólares por hora)	Educación (años)	Experiencia (años)
1	1	1980	3.31	14	1
2	1	1981	6.38	14	2
3	2	1980	5.34	13	4
4	2	1981	4.56	13	5
⋮	⋮	⋮	⋮	⋮	⋮
1087	544	1980	6.30	11	2
1088	544	1981	8.80	11	3
1089	545	1980	3.10	9	5
1090	545	1981	3.71	9	6

1.2.6 Ejemplo - Un Panel de Datos de Trabajadores

La Tabla 3 contiene algunas observaciones de una combinación de dos secciones cruzadas extraídas de la misma población en dos momentos de su historia. Estos datos podrían utilizarse para investigar cómo ha cambiado de un momento a otro la cuestión a la que se refiere el problema P04 del Ejemplo 1.1.1. Las dos secciones cruzadas completas se encuentran en el archivo PA01-SALARIOS1.WF1.

Como en el caso de una única sección cruzada, el orden en el que están dispuestas las observaciones individuales en la Tabla 3 no es importante. Sin embargo, identificar claramente a qué momento se refiere cada observación (como se hace en la segunda columna) es crucial para evaluar posibles cambios poblacionales de un momento a otro.

Por otro lado, la Tabla 4 de la página anterior contiene parte de un panel de datos que podría utilizarse para intentar resolver un problema como P10 en el Ejemplo 1.1.1. El panel completo se encuentra en el archivo PA02-SALARIOS2.WF1.

La Tabla 4 está organizada de manera que cada dos filas se refieren a los dos años observados para cada trabajador. La Tabla 4 se podría haber organizado igual que la Tabla 3 (con 545 filas primero para todos los trabajadores en el año 1980 y 545 filas debajo para los mismos trabajadores en 1981), lo que sugiere que una colección de datos de panel (como la de la Tabla 4) puede analizarse como si fuese una mera combinación de secciones cruzadas (como la de la Tabla 3). No obstante, el hecho de que en un panel de datos las entidades observadas en cada momento sean exactamente las mismas, permite la utilización de métodos especiales para su análisis con la finalidad de responder a ciertas preguntas que no son tratables con una mera combinación de secciones cruzadas.

■

En general, la dimensión temporal que los datos de panel añaden a los datos de sección cruzada permite responder a preguntas que son difíciles o imposibles de responder utilizando solamente una sección cruzada. Concretamente, el hecho de disponer de múltiples observaciones temporales sobre las mismas entidades observables facilita la evaluación de efectos causales en situaciones en las que sería imposible evaluar dichos efectos de manera fiable con una única sección cruzada.

1.3 Modelos

Una vez escogidos los datos que se van a utilizar para resolver el problema planteado al comienzo de un análisis, el paso siguiente consiste en elaborar con ellos un **modelo econométrico** (o, simplemente, un modelo). Conceptualmente, un modelo es tan sólo un **resumen** de la información esencial contenida en los datos. Para hacer operativa esta idea en la práctica, un modelo suele plantearse a través de un conjunto de **hipótesis** o **supuestos plausibles** sobre algún aspecto del funcionamiento real del sistema que ha generado los

datos. Un modelo debe ser, ante todo, una herramienta **útil** y **fiable** para responder a las preguntas planteadas al comienzo de un análisis. En consecuencia, el tipo de modelo adecuado en cada caso puede depender crucialmente tanto de dichas preguntas como de las propiedades de los datos que se van a emplear para intentar responderlas.

1.3.1 Ejemplo - El Modelo de Regresión Lineal Simple

En este ejemplo se introduce el modelo de **regresión lineal simple** (**RLS**) como una herramienta sencilla para intentar **evaluar efectos causales** entre dos variables y para **calcular previsiones** de una variable en función de (condicionada por) otra.

En las Figuras 2 y 3 de la página siguiente están representadas cuatro colecciones de datos referidas a los problemas P03, P05, P06 y P07, respectivamente, del Ejemplo 1.1.1. Cada gráfico contiene una **nube de puntos** formada por N pares de observaciones (y_1, x_1), (y_2, x_2), ..., (y_N, x_N) sobre dos variables, Y (en el eje vertical) y X (en el eje horizontal). Los datos (y_i, x_i), $i = 1, 2, ..., N$, constituyen en cada caso una muestra de N observaciones, referida a un grupo o a una secuencia de N **puntos muestrales** (entidades observables o momentos) que forman parte de un colectivo (población) o de un desarrollo histórico (proceso estocástico) más amplio. EJERCICIO 6

En un modelo RLS, la información contenida en los datos (y_i, x_i), $i = 1, 2, ..., N$, se emplea con el propósito de cuantificar una hipotética **relación causal unidireccional estocástica** entre la característica Y a la que se refieren los datos $y_1, y_2, ..., y_N$, que se denomina la **variable dependiente** del modelo, y la característica X a la que se refieren los datos $x_1, x_2, ..., x_N$, que se denomina la **variable explicativa**.

Con esta finalidad, en el modelo RLS todas las influencias posibles que recibe Y se resumen (comparar con [1]) mediante una expresión matemática del tipo

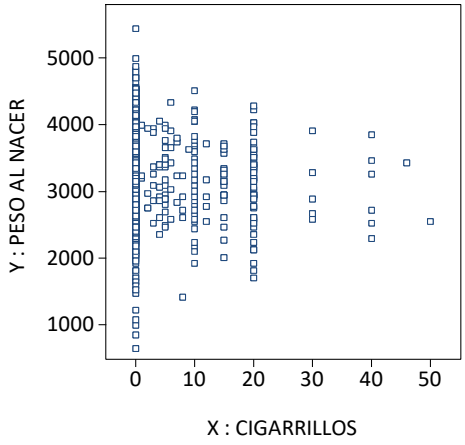
$$Y = \beta_1 + \beta_2 X + U\,, \qquad [4]$$

donde β_1 (el **término constante**) y β_2 (la **pendiente**) son dos **parámetros** (números) cuyos valores desconocidos se pretende inferir a partir de los datos disponibles; por su parte, U (un **término de error**, o una **perturbación**) representa todas las influencias observables y no observables sobre Y que no están recogidas en $\beta_1 + \beta_2 X$.

En línea con la Definición 1.1.2, la expresión [4] implica que la respuesta total de Y frente a una variación de cuantía ΔX en X es

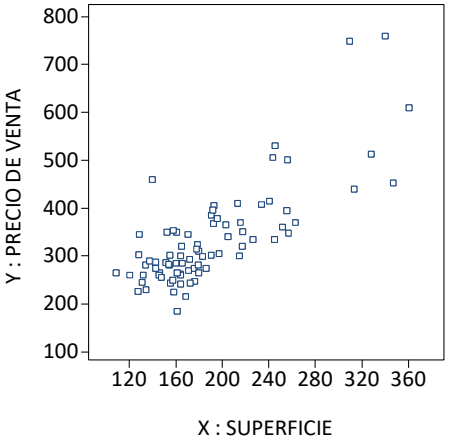
FIGURA 2

Datos de Sección Cruzada Referidos a las Preguntas P03 y P05 del Ejemplo 1.1.1



Peso (gramos) de 1387 recién nacidos en 1988 y consumo de cigarrillos durante el embarazo (número medio al día) por parte de sus madres.

SC02-RECIENNACIDOS.WF1

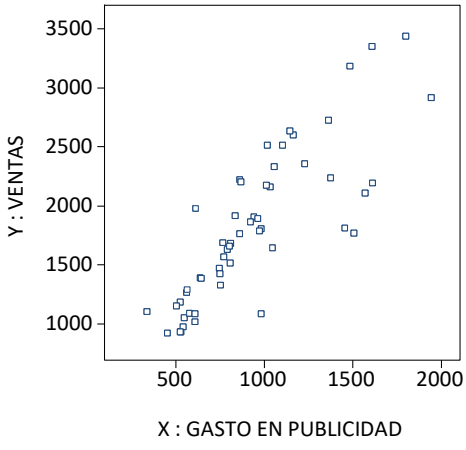


Precio (miles de dólares) y superficie (metros cuadrados) de 88 viviendas unifamiliares vendidas en el área metropolitana de Boston en 1990.

SC01-VIVIENDAS.WF1

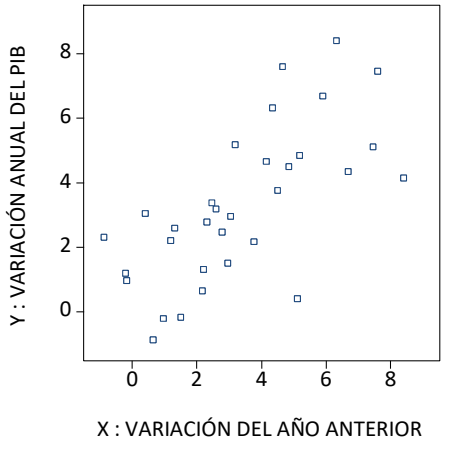
FIGURA 3

Datos de Series Temporales Referidos a las Preguntas P06 y P07 del Ejemplo 1.1.1



Ventas y gasto en publicidad anuales (miles de dólares) de la empresa Lydia E. Pinkham's Medicine Co. desde 1907 hasta 1960.

ST01-PINKHAM.WF1



Variación anual del PIB real en España (porcentaje) desde 1966 hasta 1997, junto con la misma serie retardada un año.

ST02-PIB.WF1

$$\Delta Y_X = \overbrace{\underbrace{\beta_2 \times \Delta X}_{\text{Respuesta Directa}} + \underbrace{\Delta U_X}_{\text{Respuesta Indirecta}}}^{\text{Respuesta Total}} = \left[\overbrace{\underbrace{\beta_2}_{\text{Efecto Directo}} + \underbrace{\frac{\Delta U_X}{\Delta X}}_{\text{Efecto Indirecto}}}^{\text{Efecto Total}} \right] \Delta X \qquad [5]$$

(comparar con [3]), donde $\beta_2 = \partial Y / \partial X$ es el efecto causal o "ceteris paribus" de X sobre Y, y ΔU_X representa posibles variaciones en U asociadas con X.

De [5] se deduce que el grado de asociación relativa entre Y y X observado en los datos (y_i, x_i), $i = 1, 2, ..., N$, (ver Figuras 2 y 3) representará el efecto causal de X sobre Y cuando las influencias sobre Y recogidas en U no están relacionadas con X (es decir, cuando $\Delta U_X = 0$). Por el contrario, si existe algún tipo de relación sistemática entre U y X, entonces el grado de asociación observado entre Y y X puede deberse, al menos en parte, a la existencia de dicha relación (es decir, al efecto indirecto en [5]), y no sólo a la existencia de un efecto causal o directo legítimo de X sobre Y.

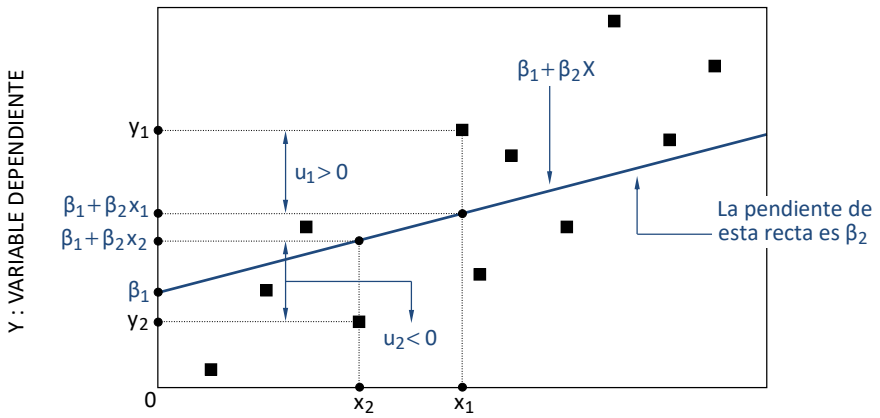
Por otro lado, [4] implica que cualquier valor observado (dato) de la variable dependiente del modelo puede descomponerse como la suma de dos términos:

$$\underbrace{Y}_{\substack{\text{Valor} \\ \text{Observado}}} = \underbrace{\beta_1 + \beta_2 X}_{\substack{\text{Parte Sistemática} \\ \text{o Previsible}}} + \underbrace{U,}_{\substack{\text{Error o} \\ \text{Parte Imprevisible}}} \qquad [6]$$

donde tanto $\beta_1 + \beta_2 X$ (la parte **sistemática** o **previsible** de Y, que se supone depende de X) como U (el **error** o la parte **imprevisible** de Y, que se supone **independiente** de X) son cantidades que **no** se conocen. La Figura 4 de la página siguiente contiene una representación gráfica del modelo RLS descrito en [6], junto con su significado referido a dos pares de datos concretos sobre Y y X $[(y_1, x_1), (y_2, x_2)$; los $N = 11$ pares de datos observados sobre Y y X está representados en la Figura 4 con un color más oscuro]. Al igual que [4] ó [6], la Figura 4 representa tan sólo (salvo por la nube de puntos) una situación **hipotética** o **supuesta**, en el sentido de que en cualquier aplicación práctica del modelo RLS **no** se sabe cuánto valen ni los parámetros β_1 y β_2, ni, por lo tanto, los errores $u_1, u_2, ..., u_N$ asociados con cada par de datos observados sobre Y y X.

FIGURA 4

Elementos Básicos del Modelo de Regresión Lineal Simple



Una vez planteado un modelo RLS como instrumento para intentar resumir el contenido informativo de una colección de datos (y_i, x_i), $i = 1, 2, ..., N$, se puede recurrir al empleo de algún **método de estimación** para asignar a β_1 y β_2 unos valores numéricos $\hat{\beta}_1$ y $\hat{\beta}_2$ concretos, que, a su vez, pueden utilizarse para calcular **estimaciones** de los dos términos del lado derecho de [6]: $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ (**valor ajustado**), $\hat{U} = Y - \hat{Y} = Y - \hat{\beta}_1 - \hat{\beta}_2 X$ (**residuo**), de manera que $Y = \hat{Y} + \hat{U}$. La Figura 5 contiene una representación gráfica del **modelo estimado**, en el que

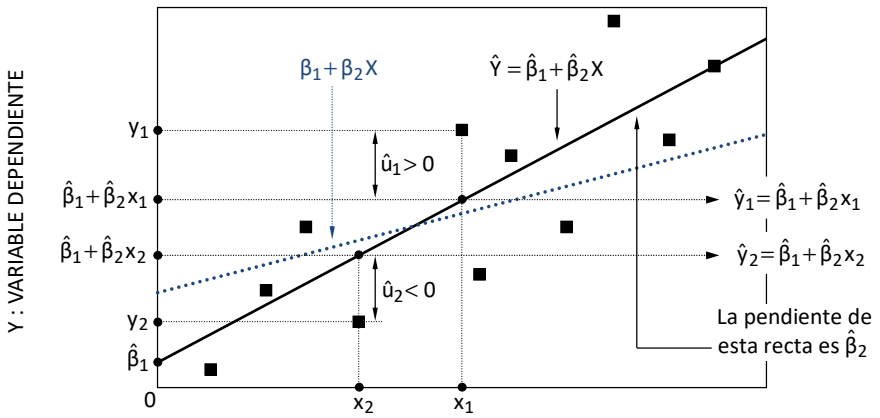
$$\underbrace{Y}_{\text{Valor Observado}} = \underbrace{\hat{\beta}_1 + \hat{\beta}_2 X}_{\text{Valor Ajustado}} + \underbrace{\hat{U}.}_{\text{Residuo}} \qquad [7]$$

A diferencia de lo que ocurre en la Figura 4, todos los elementos de la Figura 5 (excepto la línea punteada, que es la recta $\beta_1 + \beta_2 X$ de la Figura 4) pueden conocerse (calcularse) en cualquier aplicación práctica utilizando los datos disponibles (y_i, x_i), $i = 1, 2, ..., N$. El modelo RLS estimado [7] (representado en la Figura 5) puede utilizarse en la práctica con alguno de los fines que se describen en el Apartado 1.4.

El modelo RLS del Ejemplo anterior constituye una de las herramientas más sencillas que se emplean en la econometría aplicada para intentar resolver problemas como los del

FIGURA 5

El Modelo de Regresión Lineal Simple Estimado



Ejemplo 1.1.1. Actualmente, hay una gran variedad de modelos disponibles para hacer econometría aplicada en la práctica. No obstante, la variedad de opciones disponibles no debe ocultar el hecho de que, en última instancia, la **utilidad práctica** de un **modelo** sencillo o de uno muy sofisticado reside simplemente en su capacidad para proporcionar **respuestas fiables** y **entendibles** a las preguntas planteadas al comienzo de un análisis.

Una posibilidad para **calibrar** el grado de **fiabilidad** de las respuestas proporcionadas por un modelo (especialmente en entornos no experimentales), consiste en plantear ciertas **hipótesis** sobre los elementos que intervienen en él, y derivar las **conclusiones** a las que lleva el cumplimiento (o el incumplimiento) de dichas hipótesis. La validez de esas conclusiones depende de si las hipótesis utilizadas son **razonables** en cada caso concreto, en especial al compararlas con las **características muestrales** de los **datos** empleados.

En la econometría aplicada, la hipótesis central de un modelo consiste en suponer que cada **dato** disponible para el análisis es (salvo en casos excepcionales) una **realización particular** de una **variable aleatoria**.

1.3.2 Definición - Variable Aleatoria

Una **variable aleatoria** es una característica concreta de una entidad observable tal que:

1. Es susceptible de tomar distintos valores numéricos.
2. Todos los valores que puede tomar son perfectamente conocidos.
3. El valor que realmente toma no se puede prever o anticipar con absoluta certeza.

Cada uno de los valores que puede tomar una variable aleatoria se denomina un **valor posible** o una **realización posible**. Un valor que realmente toma (o que se supone que toma) una variable aleatoria es un **valor observado** o una **realización particular**.

Desde el punto de vista de un análisis aplicado, puede decirse que una variable es aleatoria cuando el analista no puede decidir o conocer de antemano su valor observado. En relación con el Ejemplo 1.1.1, el rendimiento de una zona de cultivo, la nota final y las horas de asistencia a clase de un alumno, las ventas y el gasto en publicidad de una empresa, la tasa de variación del PIB real, o el salario y los años de educación de un trabajador, son todas ellas variables aleatorias. En algunos casos excepcionales, un analista puede decidir de antemano cuáles son los valores de ciertas características observables relevantes para su análisis, porque dichas características se encuentran bajo su control, o bien porque ha elegido su colección de datos seleccionando una muestra con valores predeterminados para dichas características. No obstante, esta posibilidad es muy remota en entornos no experimentales. Aunque a veces es posible concebir ciertos experimentos (como los del Ejemplo 1.1.3) en los que se fijan de antemano los valores de algunas características relevantes para un análisis, en muchos casos es imposible, prohibitivamente costoso, o moralmente inaceptable ejecutar dichos experimentos. En este sentido, la econometría no debe explicarse ni entenderse como un conjunto variado de fórmulas y técnicas dirigidas cada una por su cuenta al análisis de distintos tipos de datos económicos o financieros, sino como un intento unificador y riguroso de analizar la **información no experimental** que se puede obtener mediante la **mera observación de sistemas** (entornos, o fenómenos) muy diversos tanto sociales como naturales.

1.3.3 Ejemplo - Algunas Hipótesis en el Modelo RLS

Para calibrar el grado de fiabilidad de las respuestas que puede proporcionar un modelo RLS estimado (como el de la ecuación [7], representado en la Figura 5), es fundamental conocer las **propiedades estadísticas** del método empleado para calcular las estimaciones $\hat{\beta}_1$ y $\hat{\beta}_2$. Bajo ciertas hipótesis sobre algunos de los elementos del modelo RLS, es posible demostrar que el método de estimación denominado M**ínimos Cuadrados Ordinarios**

(MCO) posee buenas propiedades, en el sentido de que cuando dichas hipótesis son razonables, entonces la probabilidad de que el método MCO proporcione estimaciones $\hat{\beta}_1$ y $\hat{\beta}_2$ próximas a lo que realmente valen β_1 y β_2 es relativamente elevada.

La hipótesis central en todo esto consiste en suponer que cada par de datos (y_i, x_i), $i = 1, 2, ..., N$, disponible para el análisis es una realización particular de un par de variables aleatorias (Y_i, X_i) tales que (ver 3.1.4, en especial [38])

$$Y_i = \beta_1 + \beta_2 X_i + U_i, \text{ con } \mathrm{E}[U_i \mid X_i] = \mathrm{E}[U_i] \text{ para todo } i = 1, 2, ..., N, \qquad [8]$$

de manera que el **valor esperado** o **medio** de U_i es **independiente** de X_i en cada punto muestral ($\Rightarrow U_i$ y X_i están **incorrelacionadas**). Esta hipótesis constituye un enunciado formal de la idea de independencia entre influencias omitidas e influencias incluidas en un análisis aplicado, mencionada repetidamente en el Apartado 1.1 y en el Ejemplo 1.3.1.

Adicionalmente, suele suponerse que el valor esperado de cada U_i es el mismo para todo $i = 1, 2, ..., N$ (de manera que las perturbaciones son, en cierto sentido, homogéneas a lo largo de toda la muestra), y que dicho valor esperado es cero (lo que puede garantizarse por la presencia del término constante β_1 en el modelo). EJERCICIO 7

En consecuencia, [8] puede escribirse finalmente como

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + U_i, \text{ con } \mathrm{E}[U_i \mid X_i] = 0, \text{ o bien} \\ \mathrm{E}[Y_i \mid X_i] &= \beta_1 + \beta_2 X_i, \text{ para todo } i = 1, 2, ..., N. \end{aligned} \qquad [9]$$

Esta expresión contiene las hipótesis estadísticas básicas del modelo RLS, que garantizan en cierto sentido unas buenas propiedades para el método de estimación MCO.

Observación: Cuando se plantea un modelo RLS para una colección de datos de series temporales, el índice i y la letra N que se han utilizado hasta ahora para identificar cada punto muestral y el número total de ellos, suelen cambiarse por t y T, respectivamente. Por ejemplo, una colección de datos de series temporales sobre dos variables Y y X suele representarse como (y_t, x_t), $t = 1, 2, ..., T$, mientras que [9] queda

$$Y_t = \beta_1 + \beta_2 X_t + U_t, \text{ con } \mathrm{E}[U_t \mid X_t] = 0 \text{ para todo } t = 1, 2, ..., T.$$

Con el empleo de t y T en lugar de i y N, se pretende hacer explícito que las variables y los datos considerados siguen un **orden temporal** o **cronológico** (ver Ejemplo 1.2.4) que es fundamental para su análisis. Con esta notación también se pretende poner de manifiesto que en el análisis de datos de series temporales deben tenerse muy en cuenta ciertas posibilidades (**efectos dinámicos**, **autocorrelación**, **no estacionariedad**, **tendencias**, ...) que no suelen ser relevantes en el análisis de datos de sección cruzada (para los que, en general, el orden de las observaciones no importa; ver Ejemplo 1.2.2).

1.4 Respuestas

Las respuestas que proporciona un modelo estimado a partir de unos datos, pueden ir desde una mera **descripción cuantitativa** hasta una serie de **previsiones** muy valoradas a la hora de tomar decisiones importantes.

El tipo de respuesta que es adecuado en cada caso depende, lógicamente, de cuál haya sido la pregunta planteada al comienzo del análisis. En consecuencia, el planteamiento inicial preciso sobre el objetivo y el interés de un análisis es fundamental no sólo para decidir adecuadamente qué datos utilizar y qué modelo emplear para resumir su contenido informativo, sino también para decidir, en última instancia, qué hacer con dicho modelo.

En el Ejemplo 1.4.1 se describen algunos usos posibles del modelo RLS estimado de la ecuación [7] (representado gráficamente en la Figura 5). Dichos usos se ilustran en el Ejemplo 1.4.2 a través de cuatro modelos RLS estimados con las colecciones de datos representadas en las Figuras 2 y 3 del Ejemplo 1.3.1 del Apartado 1.3 anterior.

1.4.1 Ejemplo - Algunos Usos de la Regresión Lineal Simple I

En las Figuras 6 y 7 de la página siguiente están representados gráficamente cuatro modelos RLS estimados con los datos de las Figuras 2 y 3 del Ejemplo 1.3.1. Como en la Figura 5, la representación de cada modelo estimado tiene la forma de una línea recta del tipo $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$, cuya pendiente es igual a $\hat{\beta}_2$ (una estimación numérica del parámetro β_2). En todos los casos, $\hat{\beta}_1$ y $\hat{\beta}_2$ se han calculado utilizando el método MCO mencionado en el Ejemplo 1.3.3, lo que implica, en particular, que

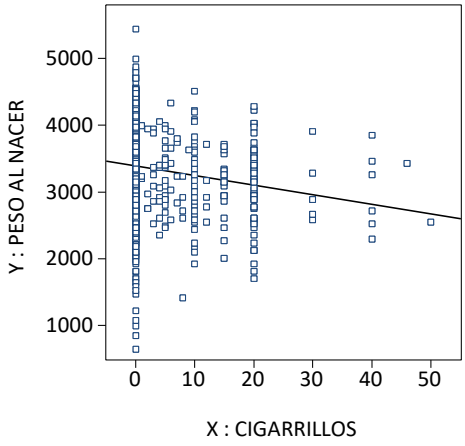
$$\hat{\beta}_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\text{Covarianza Muestral entre } Y, X}{\text{Varianza Muestral de } X}. \qquad [10]$$

Por lo tanto, $\hat{\beta}_2$ mide simplemente el grado de **asociación lineal relativa** entre Y y X observado en los datos disponibles sobre dichas variables. En este sentido, puede resultar útil comparar [10] con la ecuación [5] del Ejemplo 1.3.1 (y la discusión posterior), así como recordar toda la discusión del Ejemplo 1.3.3. EJERCICIO 8

Un modelo RLS estimado como el de la ecuación [7] (representado en la Figura 5), puede utilizarse en la práctica (quizás entre otros fines) para lo siguiente:

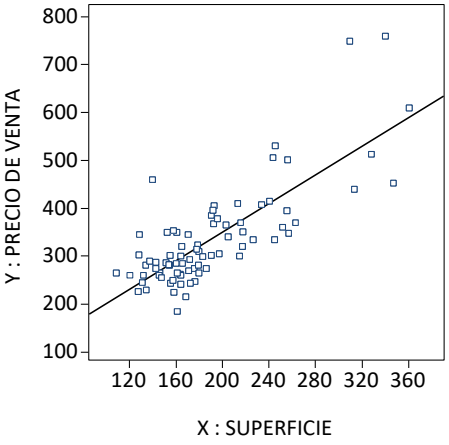
FIGURA 6

Modelos RLS Estimados por MCO con los Datos de la Figura 2



$$\widehat{\text{PESO}} = 3392 - 14.4 \times \text{CIGM}$$

N = 1387 recién nacidos

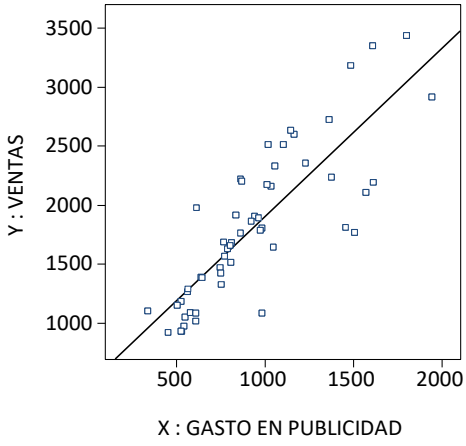


$$\widehat{\text{PRECIO}} = 51.551 + 1.494 \times \text{SUPM2}$$

N = 80 viviendas

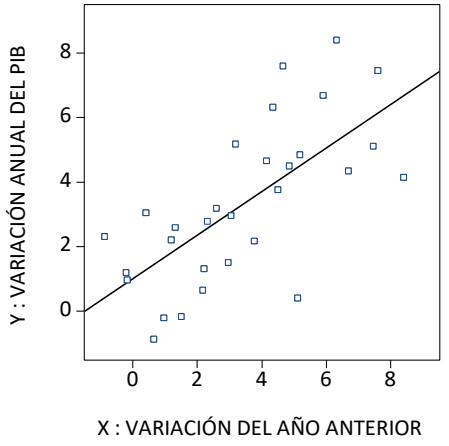
FIGURA 7

Modelos RLS Estimados por MCO con los Datos de la Figura 3



$$\widehat{\text{VENTAS}} = 477.500 - 1.429 \times \text{GPUB}$$

N = 54 años (1907-1960)



$$\widehat{\text{TVPIB}} = 1.01 + 0.67 \times \text{TVPIB}(-1)$$

N = 32 años (1966-1997)

I. Evaluar el efecto causal o "ceteris paribus" de X sobre Y

Del modelo RLS [4] se deduce que $\beta_2 = \partial Y/\partial X$, por lo que

$$\Delta Y = \beta_2 \times \Delta X \quad \text{cuando} \quad \Delta U = 0\,.$$

En consecuencia, β_2 representa la variación absoluta en Y provocada por una variación absoluta unitaria en X ($\Delta X = 1$), cuando todos los factores contenidos en U se mantienen constantes ($\Delta U = 0$). Aunque esta interpretación del parámetro β_2 en [4] es correcta sin ambigüedades, también es cierto que la fiabilidad de $\hat{\beta}_2$ en [10] como estimación de β_2 depende crucialmente de que $\mathrm{E}[U_i \mid X_i] = 0$ para todo $i = 1, 2, \ldots, N$ (como se supone en la ecuación [9] del Ejemplo 1.3.3). Dado que $\hat{\beta}_2$ mide el grado de asociación lineal relativa entre Y y X observado en los datos, $\hat{\beta}_2$ será una estimación fiable del efecto causal de X sobre Y cuando las influencias sobre Y recogidas en U no estén relacionadas con X; en caso contrario, el grado de asociación entre Y y X captado por $\hat{\beta}_2$ puede deberse, al menos en parte, a la existencia de una respuesta indirecta de Y frente a X a través de U (como indica la ecuación [5]), y no sólo a la existencia de un efecto causal o directo legítimo de X sobre Y. Cuando la estimación por MCO de un modelo RLS no es fiable porque $\mathrm{E}[U_i \mid X_i] \neq 0$ (una posibilidad que siempre debería considerarse en la práctica), es muy importante (i) investigar las causas posibles de esa situación para intentar corregirlas, o bien (si no es posible corregir las causas) (ii) emplear otro método alternativo a MCO para estimar el modelo de manera más fiable. EJERCICIO 9

II. Prever Y en función de X

Una hipótesis implícita en muchos análisis econométricos aplicados es que los datos empleados son suficientemente representativos de la población o del proceso estocástico del que han sido extraídos. Bajo esta hipótesis, se espera que las conclusiones del análisis sean aplicables más allá de la muestra empleada. En particular, si p hace referencia a una entidad observable o a un momento fuera de dicha muestra, un modelo estimado como [7] puede resultar útil para estimar una esperanza del tipo $\mathrm{E}[Y_p \mid X_p]$ siempre que la ecuación [9] sea aplicable en $i = p$, es decir, siempre que $\mathrm{E}[Y_p \mid X_p] = \beta_1 + \beta_2 X_p$. Si x_p es un valor dado (conocido o supuesto) para X_p, entonces $\mathrm{E}[Y_p \mid X_p = x_p] = \beta_1 + \beta_2 x_p$ se puede estimar, de acuerdo con [7], como $\hat{y}_p = \hat{\beta}_1 + \hat{\beta}_2 x_p$, que se denomina una **previsión puntual**. La fiabilidad de una previsión de este tipo depende crucialmente de que $\hat{\beta}_2$ sea una estimación fiable del efecto causal de X sobre Y, y de que dicho efecto sea

aplicable a la entidad observable o al momento p considerado.

III. Controlar Y a través de X

En ciertas ocasiones, la variable explicativa de un modelo RLS (como el gasto en publicidad de una empresa en un año determinado) puede manipularse o controlarse en algún sentido práctico para conseguir o para simular un objetivo (como un volumen de ventas esperado al final del año). En estos casos, [7] puede resultar útil para estimar el valor de X asociado con un valor esperado para Y fijado de antemano. En particular, si c hace referencia a una entidad observable o a un momento tal que [9] es aplicable en $i = c$, entonces $\mathrm{E}[Y_c \,|\, X_c] = \beta_1 + \beta_2 X_c$. Por lo tanto, si se desea que $\mathrm{E}[Y_c \,|\, X_c]$ sea igual a un valor y_c^* dado, se debe asignar a X_c un valor x_c^* tal que $\beta_1 + \beta_2 x_c^* = y_c^*$, lo que implica que x_c^* debe ser igual a $(y_c^* - \beta_1) / \beta_2$. Esta cantidad puede estimarse como $\hat{x}_c^* = (y_c^* - \hat{\beta}_1) / \hat{\beta}_2$, lo que garantiza, de acuerdo con [7]-[9], que $\hat{\mathrm{E}}[Y_c \mid X_c = \hat{x}_c^*] = y_c^*$. Como en el caso del punto II anterior, la fiabilidad de $\hat{x}_c^*$ depende crucialmente de que $\hat{\beta}_2$ sea una estimación fiable del efecto causal de X sobre Y, y de que dicho efecto sea aplicable a la entidad observable o al momento c considerado.

■

1.4.2 Ejemplo - Algunos Usos de la Regresión Lineal Simple II

A continuación se emplean las posibilidades descritas en el Ejemplo 1.4.1 para responder a las preguntas P03, P05, P06 y P07 planteadas en el Ejemplo 1.1.1. EJERCICIO 10

Respuesta a la pregunta P03 del Ejemplo 1.1.1:

Del modelo estimado en el panel izquierdo de la Figura 6 se deduce que:

➪ $\hat{\beta}_1 = 3392$ gramos es una estimación del peso esperado cuando $CIGM = 0$.

➪ $\hat{\beta}_2 = -14.4 \Rightarrow \hat{\Delta} PESO = -14.4$ gramos cuando $\Delta CIGM = 1$ cigarrillo, $\Delta U = 0$.

Respuesta a la pregunta P05 del Ejemplo 1.1.1:

Del modelo estimado en el panel derecho de la Figura 6 se deduce que:

➪ $\hat{\beta}_1 = 51551$ dólares es la parte estimada del precio esperado que no depende de *SUPM2*. En este caso, interpretar $\hat{\beta}_1 = 51551$ como el precio previsto de una vivienda con $SUPM2 = 0$ no tiene sentido práctico porque no existen "viviendas sin superficie".

➪ $\hat{\beta}_2 = 1.494 \Rightarrow \hat{\Delta} PRECIO = 1494$ dólares cuando $\Delta SUPM2 = 1\ \mathrm{m}^2$, $\Delta U = 0$.

➪ El precio previsto para una vivienda con, por ejemplo, 200 metros cuadrados de superficie, es $\hat{\mathrm{E}}[PRECIO \,|\, SUPM2 = 200] = 51551 + 1494 \times 200 = 350351$ dólares.

Respuesta a la pregunta P06 del Ejemplo 1.1.1:

Del modelo estimado en el panel izquierdo de la Figura 7 se deduce que:

➪ $\hat{\beta}_1 = 477500$ dólares es el volumen de ventas previsto cuando $GPUB = 0$.

➪ $\hat{\beta}_2 = 1.429 \Rightarrow \hat{\Delta} VENTAS = 1429$ dólares cuando $\Delta PUB =$ mil dólares, $\Delta U = 0$.

➪ El volumen de ventas previsto para un gasto en publicidad de un millón de dólares en 1961, es $\hat{\mathrm{E}}[VENTAS \,|\, GPUB = 10^6] = 477500 + 1.429 \times 10^6 = 1906500$ dólares.

➪ Si se pretende un volumen de ventas de, por ejemplo, dos millones de dólares para el año 1961, entonces $2000000 = 477500 + 1.429 \times \hat{GPUB}^* \Rightarrow \hat{GPUB}^* \cong 1065000$ dólares.

Respuesta a la pregunta P07 del Ejemplo 1.1.1:

Del modelo estimado en el panel derecho de la Figura 7 se deduce que:

➪ $\hat{\beta}_1 = 1.01\%$ es la variación anual prevista si la última variación anual (es decir, la del año anterior) fue igual a cero.

➪ Si la última variación anual (el dato disponible sobre 1997) fue 3.38%, entonces:

Previsión para 1998: $\hat{\mathrm{E}}[TVPIB \,|\, TVPIB(-1) = 3.38] = 1.01 + 0.67 \times 3.38 = 3.27\%$.

Previsión para 1999: $\hat{\mathrm{E}}[TVPIB \,|\, TVPIB(-1) = 3.27] = 1.01 + 0.67 \times 3.27 = 3.20\%$.

La serie de previsiones converge a largo plazo a una variación anual del 3.06%.

■

Observación: De acuerdo con el modelo RLS [4] (ver el punto I del Ejemplo 1.4.1), el efecto sobre Y de una variación absoluta unitaria en X no depende de los valores de partida de X ni de Y. Por ejemplo, un gasto adicional de mil dólares en publicidad tendría un efecto absoluto sobre las ventas que sería independiente del gasto en publicidad ya efectuado y del volumen de ventas de partida. Esta implicación del modelo RLS no es adecuada en muchas situaciones prácticas. No obstante, el modelo RLS puede adaptarse a dichas situaciones definiendo su variable dependiente y su variable explicativa como alguna **transformación** de cada variable original en el análisis. Por ejemplo, si $Y = \ln(F)$ y $X = \ln(P)$ en [4], entonces β_2 es la **elasticidad** de F con respecto a P, es decir, la respuesta porcentual (aproximada) de F provocada por una variación de un 1% en P cuando $\Delta U = 0$. Un modelo como $\ln(F) = \beta_1 + \beta_2 \ln(P) + U$ es no lineal con respecto a F y P [porque

$\ln(\cdot)$ es una función no lineal], aunque es un modelo que sí es lineal con respecto a los parámetros β_1 y β_2. El modelo RLS es **lineal** justamente por este motivo, es decir, porque β_1 y β_2 aparecen en [4] de forma lineal. Al mismo tiempo, Y y X pueden representar en [4] cualesquiera transformaciones de las variables originales de interés en el análisis, lo que proporciona al modelo RLS bastante más flexibilidad que la que sugieren a simple vista la ecuación [4] o la Figura 4. Ver 3.1.1-3.1.3, en especial [23]-[30].

La **fiabilidad** de las **respuestas** proporcionadas por un modelo elaborado a partir de unos datos depende en la práctica de varios factores. Entre ellos, el factor crítico quizás sea (sobre todo en entornos no experimentales) el grado de **compatibilidad** entre las **hipótesis** que conforman el modelo y las **pautas muestrales** que están realmente presentes en los **datos**. Sin una evaluación adecuada de ese grado de compatibilidad, es imposible juzgar la fiabilidad de las conclusiones obtenidas al final de un análisis. La cita siguiente resume con bastante precisión este aspecto crucial de cualquier análisis de datos:

> *Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong, we ought to be able to perceive in what ways they are wrong. […] Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.* ANSCOMBE (1973) PÁGINA 17.

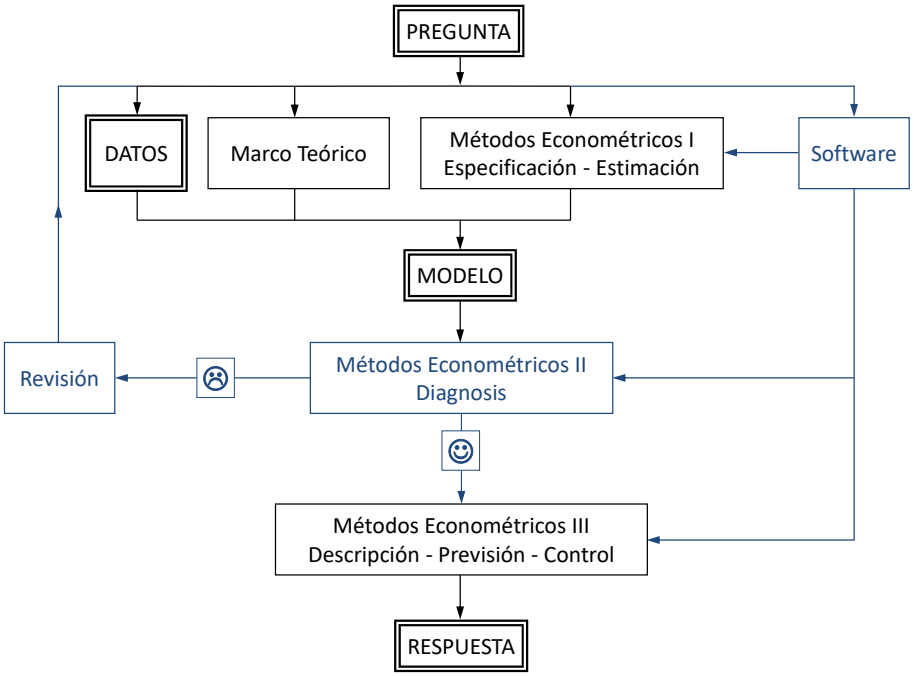
En la Sección 2 siguiente se describe una estrategia posible para abordar en la práctica el problema central de la econometría aplicada: cómo elaborar modelos que resuman de manera adecuada el contenido informativo de una colección de datos. Las operaciones más importantes en dicha estrategia son, precisamente, las que tienen que ver con la cita anterior: la **diagnosis** y la **revisión** de un modelo. De la ejecución adecuada de estas operaciones depende, en gran medida, la fiabilidad de las respuestas que proporciona un análisis econométrico aplicado.

2 La Metodología de la Econometría Aplicada

En general, las operaciones fundamentales en la elaboración de un modelo a partir de unos datos no tienen mucho que ver con plantear una especificación inicial lo más completa y rigurosa posible. Por el contrario, dichas operaciones tienen que ver más bien con la

FIGURA 8

La Metodología de la Econometría Aplicada



revisión crítica y sistemática de una especificación inicial razonable, hasta obtener la que mejor describa la información contenida en los datos.

El diagrama de la Figura 8 resume los pasos que suelen seguirse actualmente en la econometría aplicada para obtener una **respuesta** a una **pregunta** concreta, sobre la base de la información contenida en unos **datos** y resumida en un **modelo**. Las líneas y los recuadros de trazo más claro representan procedimientos y recursos especialmente críticos en la práctica, que, no obstante, reciben a veces poca atención (o ninguna en absoluto) tanto en muchos manuales de econometría como en muchas publicaciones y estudios académicos y profesionales.

La forma de un modelo y los métodos econométricos adecuados para su elaboración y su uso en la práctica, pueden diferir según la naturaleza de los datos empleados (ver Ejemplos 1.2.2, 1.2.4 y 1.2.6) y el tipo de pregunta que se pretende responder (es decir, el uso que se pretende dar al modelo; ver Ejemplo 1.4.1). No obstante, cualquier análisis econométrico

aplicado sigue (o, más bien, debería seguir) un desarrollo semejante al que está descrito en la Figura 8.

En esta Sección 2 se ofrece una descripción breve de algunos de los elementos de la Figura 8 que no han sido considerados en la Sección 1 anterior. En particular, en esta Sección 2 se discuten algunas cuestiones relacionadas con la especificación, la estimación y la diagnosis de modelos econométricos.

2.1 Especificación

La **especificación** de un modelo se refiere al planteamiento de un conjunto de hipótesis plausibles (aunque no incuestionables ni definitivas) sobre las características esenciales de los datos recogidos para el análisis. Para contribuir a la formulación inicial de dichas hipótesis, a veces es útil recurrir a un **marco teórico**. En el contexto de la Figura 8, un marco (o un modelo) teórico es tan sólo una construcción matemática, de origen más o menos formal (como una teoría económica, financiera, sociológica, biológica o física, o bien simplemente la **intuición**, el **criterio subjetivo** y la **experiencia** del analista), en la que intervienen las características o variables a las que refieren los datos.

La posible utilidad de un marco teórico, sobre todo en relación con la especificación inicial de un modelo econométrico, siempre debe tenerse en cuenta. No obstante, la visión tradicional de la econometría como una herramienta para evaluar empíricamente teorías económicas es probablemente una visión anticuada e innecesariamente limitada. Nada obliga a que un modelo teórico constituya el elemento central de un análisis econométrico aplicado. En cualquier caso, un examen detallado de los datos al comienzo de un análisis suele resultar tan revelador o más que cualquier modelo teórico por muy sofisticado que sea. La cita siguiente resume este punto de vista sobre la econometría aplicada:

> *In this book we describe econometric modelling from an applied point of view where we start from the data. We consider models as constructs that we can change in the light of data information. By incorporating more of the relevant data characteristics in the model, we may improve our understanding of the underlying economic processes. […] This view of econometric modelling differs from a more traditional one that has more confidence in the theory and the postulated model and less in the observed data. In this view econometrics is concerned with the measurement of theoretical relations as suggested by economic theory. In our approach, on the other hand, we are not primarily interested in testing a particular theory but in using data to*

> *get a better understanding of an observed phenomenon of interest.* HEIJ, DE BOER, FRANSES, KLOEK Y VAN DIJK (2004) PÁGINAS 274-275.

En la práctica, este punto de vista es aplicable no sólo a cuestiones o problemas de tipo económico o financiero, sino, en general, a cualquier cuestión de interés tanto social como natural que se pretenda analizar utilizando la información cuantitativa contenida en una colección de datos.

En los dos ejemplos siguientes se introduce el modelo de **regresión lineal múltiple** (**RLM**) como una herramienta sencilla para mitigar las dificultades asociadas con el modelo RLS a la hora de evaluar efectos causales. También se ilustra la posibilidad de que una mera inspección visual de los datos proporcione básicamente la misma información que un marco teórico formal a la hora de especificar inicialmente un modelo RLM.

2.1.1 Ejemplo - El Modelo de Regresión Lineal Múltiple

Como se vio en la Definición 1.1.2 a partir de las expresiones [1]-[3], una dificultad asociada con la evaluación del efecto causal de X sobre Y utilizando solamente datos sobre Y y X, tiene que ver con que Y puede depender de otros factores observables **relacionados** con X que **no** se consideran explícitamente en el análisis. Si esos factores pudieran incluirse de forma explícita en la investigación de una posible relación causal entre Y y X, entonces la dificultad mencionada no estaría presente. Para verlo, consideramos, en línea con la Definición 1.1.2, la respuesta total de Y ante una variación de cuantía ΔX en X,

$$\Delta Y_X \cong F_X \times \Delta X + F_W \times \Delta W_X + F_V \times \Delta V_X ,$$

junto con la respuesta total de Y ante una variación de cuantía ΔW en W,

$$\Delta Y_W \cong F_W \times \Delta W + F_X \times \Delta X_W + F_V \times \Delta V_W .$$

En la Definición 1.1.2 se vieron las implicaciones de considerar individualmente tan sólo la primera de las dos expresiones anteriores. Sin embargo, las dos expresiones se pueden escribir **conjuntamente** como

$$\underbrace{\begin{bmatrix} \Delta Y_X \\ \Delta Y_W \end{bmatrix}}_{[\,\mathbf{\Delta Y_{XW}}\,]} \cong \underbrace{\begin{bmatrix} \Delta X & \Delta W_X \\ \Delta X_W & \Delta W \end{bmatrix}}_{[\,\mathbf{\Delta XW}\,]} \underbrace{\begin{bmatrix} F_X \\ F_W \end{bmatrix}}_{[\,\mathbf{F_{XW}}\,]} + F_V \underbrace{\begin{bmatrix} \Delta V_X \\ \Delta V_W \end{bmatrix}}_{[\,\mathbf{\Delta V_{XW}}\,]}, \qquad [11]$$

o bien, de manera más compacta y explícita, como

$$[\,\mathbf{\Delta Y_{XW}}\,] \cong [\,\mathbf{\Delta XW}\,][\,\mathbf{F_{XW}}\,] + F_V\,[\,\mathbf{\Delta V_{XW}}\,] =$$

$$= [\,\mathbf{\Delta XW}\,]\left[\overbrace{\underbrace{[\,\mathbf{F_{XW}}\,]}_{\text{Efectos Directos}} + \underbrace{F_V\,[\,\mathbf{\Delta XW}\,]^{-1}\,[\,\mathbf{\Delta V_{XW}}\,]}_{\text{Efectos Indirectos}}}^{\text{Efectos Totales}}\right]. \qquad [12]$$

Esta expresión (que no es más que una extensión de [3] en la Definición 1.1.2) pone de manifiesto que $[\,\mathbf{\Delta XW}\,]^{-1}\,[\,\mathbf{\Delta Y_{XW}}\,] \cong [\,\mathbf{F_{XW}}\,]$ siempre que (i) $F_V = 0$, de manera que V no influye directamente sobre Y (una posibilidad, en general, más bien remota), o bien que (ii) $[\,\mathbf{\Delta V_{XW}}\,] = \mathbf{0}$, de manera que V no varía sistemáticamente **ni** con X **ni** con W. Por lo tanto, si las influencias no observadas recogidas en V **no** están **relacionadas ni** con X **ni** con W, entonces el efecto causal de X sobre Y (al igual que el de W sobre Y, que también podría tener cierto interés) se puede evaluar de manera fiable a partir de [11]-[12] utilizando datos sobre Y, X y W, a pesar de que X y W estén relacionadas entre sí. Por el contrario, como se vio en la Definición 1.1.2, el efecto causal de X sobre Y no podría evaluarse fiablemente a partir tan sólo de la primera ecuación de [11] (incluso suponiendo que $\Delta V_X = 0$), a menos que o bien $F_W = 0$, o bien $\Delta W_X = 0$. La ventaja de incluir a W en un análisis de causalidad de X sobre Y, reside justamente en que la inclusión de W elimina la necesidad de asumir ciertas hipótesis (como $F_W = 0$, o $\Delta W_X = 0$) que pueden resultar difíciles de justificar en la práctica.

Esta conclusión sugiere que en un análisis de causalidad sobre Y es recomendable considerar explícitamente tantas influencias observables como sea posible y razonable, a pesar de que en última instancia quizás interese evaluar el efecto directo o causal de tan sólo una de dichas influencias. En este sentido, un modelo RLS como el de los Ejemplos 1.3.1 y 1.3.3 puede ampliarse para recoger explícitamente otras influencias directas sobre Y adicionales a la influencia de X. Esta ampliación da lugar a un modelo de **regresión lineal múltiple** (**RLM**), en el que todas las influencias posibles que recibe Y se resumen mediante una expresión matemática del tipo

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + U\,, \qquad [13]$$

donde X_2, X_3, ..., X_K son $K-1$ variables explicativas observables, posiblemente relacionadas entre sí; por su parte, U representa todas las influencias (observables y no observables) sobre Y que no están recogidas en $\beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K$.

A partir de [13], las respuestas totales de Y ante variaciones de cuantía ΔX_j en X_j $(j = 2, 3, ..., K)$ pueden escribirse conjuntamente (al estilo de [11]) como

$$\underbrace{\begin{bmatrix} \Delta Y_{X_2} \\ \Delta Y_{X_3} \\ \vdots \\ \Delta Y_{X_K} \end{bmatrix}}_{[\,\boldsymbol{\Delta Y_X}\,]} = \underbrace{\begin{bmatrix} \Delta X_2 & \Delta X_{3.2} & \cdots & \Delta X_{K.2} \\ \Delta X_{2.3} & \Delta X_3 & \cdots & \Delta X_{K.3} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta X_{2.K} & \Delta X_{3.K} & \cdots & \Delta X_K \end{bmatrix}}_{[\,\boldsymbol{\Delta XX}\,]} \underbrace{\begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix}}_{\boldsymbol{\beta}_2} + \underbrace{\begin{bmatrix} \Delta U_{X_2} \\ \Delta U_{X_3} \\ \vdots \\ \Delta U_{X_K} \end{bmatrix}}_{[\,\boldsymbol{\Delta U_X}\,]}, \qquad [14]$$

o bien, de manera más compacta y explícita (al estilo de [12]), como

$$[\,\boldsymbol{\Delta Y_X}\,] = [\,\boldsymbol{\Delta XX}\,] \times \boldsymbol{\beta}_2 + [\,\boldsymbol{\Delta U_X}\,] =$$

$$= [\,\boldsymbol{\Delta XX}\,] \left[\overbrace{\underbrace{\boldsymbol{\beta}_2}_{\text{Efectos Directos}} + \underbrace{[\,\boldsymbol{\Delta XX}\,]^{-1}\,[\,\boldsymbol{\Delta U_X}\,]}_{\text{Efectos Indirectos}}}^{\text{Efectos Totales}} \right]. \qquad [15]$$

Esta expresión (que es simplemente una extensión de [5] en el Ejemplo 1.3.1) indica que $[\,\boldsymbol{\Delta XX}\,]^{-1}\,[\,\boldsymbol{\Delta Y_X}\,] = \boldsymbol{\beta}_2$ siempre que $[\,\boldsymbol{\Delta U_X}\,] = \mathbf{0}$, es decir, siempre que U en [13] no varíe sistemáticamente con **ninguna** variable explicativa. Por lo tanto, si las influencias recogidas en U son independientes de **todas** las variables explicativas, entonces tanto β_2 como $\beta_3, ..., \beta_K$ pueden evaluarse de manera fiable a partir de [15] utilizando datos sobre Y, X_2, X_3, ..., X_K. Por el contrario, como se vio en la Definición 1.1.2, β_2 no podría evaluarse fiablemente a partir tan sólo de la primera ecuación de [14] (incluso suponiendo que $\Delta U_{X_2} = 0$), a menos que $\beta_j \times \Delta X_{j.2} = 0$ para todo $j = 3, ..., K$. La ventaja de incluir a $X_3, ..., X_K$ junto con X_2 en [13] en relación con un análisis de causalidad de X_2 sobre Y, reside justamente en que la inclusión de $X_3, ..., X_K$ elimina la necesidad de asumir ciertas hipótesis (como que o bien $\beta_j = 0$, o bien $\Delta X_{j.2} = 0$ para todo $j = 3, ..., K$) que pueden resultar difíciles de justificar en la práctica.

Según esto y como extensión del Ejemplo 1.3.3, en un modelo RLM se supone que cada observación o vector de datos $(y_i, x_{i2}, x_{i3}, ..., x_{iK})$, $i = 1, 2, \ldots, N$, sobre todas las variables incluidas en el análisis, es una realización particular de un vector de variables aleatorias $(Y_i, X_{i2}, X_{i3}, ..., X_{iK})$ tal que

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_K X_{iK} + U_i \qquad [16]$$

con

$$\mathrm{E}[U_i \mid X_{i2}, X_{i3}, ..., X_{iK}] = \mathrm{E}[U_i] = 0 \qquad [17]$$

para todo $i = 1, 2, ..., N$, de manera que el **valor esperado** o **medio** de U_i es **independiente** de $X_{i2}, X_{i3}, ..., X_{iK}$ en cada punto muestral ($\Rightarrow U_i$ y X_{ij} están **incorrelacionadas** para todo $j = 2, ..., K$). El análisis teórico y el uso en la práctica de modelos RLM constituyen uno de los elementos centrales de toda la econometría.

Las expresiones [16]-[17] contienen las hipótesis estadísticas básicas del modelo RLM, que, como en el caso del modelo RLS del Ejemplo 1.3.3, garantizan en cierto sentido unas buenas propiedades para el método de estimación MCO.

■

2.1.2 Ejemplo - Especificación Inicial de un Modelo RLM

Con el fin de ilustrar cómo especificar inicialmente un modelo RLM en la práctica, en la Figura 9 de la página siguiente se ha representado parte de una sección cruzada que podría utilizarse en relación con el problema P10 del Ejemplo 1.1.1. El modelo más sencillo que puede considerarse para evaluar el efecto causal de la educación sobre el salario es un modelo RLS del tipo

$$SLRPH = \beta_1 + \beta_2 EDUC + U\,, \qquad [18]$$

aunque el panel izquierdo de la Figura 9 sugiere que un modelo RLS quizás más adecuado que [18] podría ser en este caso EJERCICIO 11

$$\ln SLRPH = \beta_1 + \beta_2 EDUC + U\,. \qquad [19]$$

El problema fundamental de [19] reside en que si U contiene alguna influencia sobre el salario de un trabajador que esté relacionada con su educación (como, por ejemplo, su experiencia, o su habilidad o su capacidad para desempeñar adecuadamente su trabajo), entonces la posible independencia entre U y $EDUC$ no es justificable en [19]. En este sentido, un modelo RLM que incluye la experiencia como variable explicativa es

$$\ln SLRPH = \beta_1 + \beta_2 EDUC + \beta_3 EXLP + U\,, \qquad [20]$$

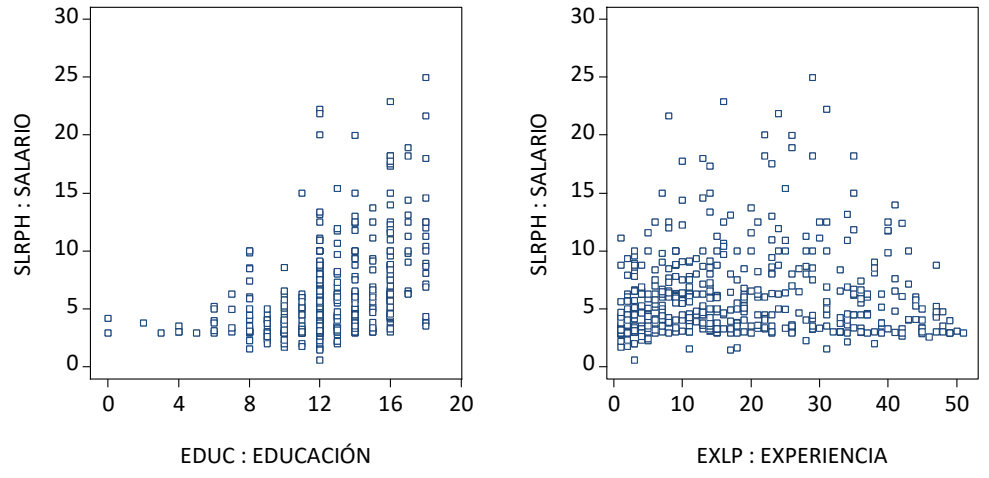
aunque el panel derecho de la Figura 9 sugiere que un modelo RLM quizás más adecuado que [20] podría ser en este caso EJERCICIO 11

$$\ln SLRPH = \beta_1 + \beta_2 EDUC + \beta_3 EXLP + \beta_4 EXLP^2 + U\,. \qquad [21]$$

FIGURA 9

Datos de Sección Cruzada referidos a la Pregunta P10 del Ejemplo 1.1.1

Salario medio anual (dólares por hora), educación (años) y experiencia laboral (años) de 526 personas trabajadoras entrevistadas en 1976 (*Current Population Survey - U.S. Census Bureau*).
SC03-SALARIOS3.WF1



La ventaja fundamental de [21] frente a [19] reside en que en [21] es seguro que U no contiene ni a $EXLP$ ni a $EXLP^2$, por lo que la posible relación entre U y $EDUC$ es seguramente notablemente menor en [21] que en [19]. En todo caso, para poder evaluar fiablemente tanto β_2 como β_3 y β_4 en el modelo RLM [21], U debe ser independiente no sólo de la educación, sino también de la experiencia. Aunque la presencia en U de factores no observables (como la habilidad para desempeñar un trabajo) siempre es problemática en este sentido, también es cierto que a [21] se le pueden añadir otras influencias observables relacionadas con la educación, con la experiencia o con ambas. En esta línea, una extensión razonable de [21] podría ser la siguiente: EJERCICIO 11

$$\overline{\ln SLRPH} = \beta_1 + \beta_2 \overline{EDUC} + \beta_3 \overline{EXLP} + \\ + \beta_4 \overline{EXLP^2} + \beta_5 \overline{EDUC \times EXLP} + U \, . \qquad [22]$$

Si β_5 en [22] es distinto de cero, entonces U en [21] no puede ser independiente ni de la educación ni de la experiencia. Quizás U en [22] tampoco lo es, aunque seguro que el

motivo no sería en ese caso la ausencia de un término asociado con el producto de la educación por la experiencia.

Un modelo como [22] puede justificarse a partir de un modelo teórico como el llamado Modelo del Capital Humano (en inglés, *Human Capital Model*; ver, por ejemplo, Berndt 1991, Capítulo 5). No obstante, toda la discusión que precede a [22] sugiere que un modelo RLM como [22] puede surgir de manera bastante natural sin más que combinar algo de intuición y buen juicio con lo que puede observarse acerca de las posibles relaciones (parciales) entre los datos sobre *SLRPH* y *EDUC*, por un lado, y sobre *SLRPH* y *EXLP*, por otro, representados en la Figura 9.

Observación: Cuando se dice que un modelo como [22] (o como cualquiera de los de las ecuaciones [19]-[21]) es **lineal**, se está haciendo referencia a cómo figuran los parámetros β_1 (término constante), β_2, β_3, β_4 y β_5 (pendientes) en el lado derecho de [22], **no** a cómo figuran las variables *SLRPH*, *EDUC* y *EXLP* en [22]. (Ver la Observación que sigue al Ejemplo 1.4.2 en relación con el modelo RLS.) En general, las variables Y, X_2, X_3, ..., X_K pueden representar en [13] cualesquiera transformaciones de las variables originales de interés en un análisis. Esto proporciona una gran flexibilidad a la hora de representar diferentes tipos de efectos causales mediante modelos RLM. En la práctica, la finalidad esencial de transformar las variables originales de interés en un análisis, consiste en especificar un modelo cuyas perturbaciones tengan unas propiedades que garanticen la validez de los métodos econométricos que vayan a ser utilizados; al mismo tiempo, es muy importante que las transformaciones empleadas proporcionen un modelo que represente efectos causales razonablemente interpretables, como en [22].

Del ejemplo anterior puede concluirse (igual que se indica al final del Ejemplo 1.1.3) que la evaluación fiable en la práctica de efectos causales descansa, en cierta medida, en la confianza que se tenga en que las influencias que no se consideran explícitamente en un análisis sean razonablemente independientes de las influencias que sí se consideran de manera explícita. La cita siguiente resume con precisión este aspecto fundamental de la econometría aplicada:

> *The key question in most empirical studies is: Have enough other factors been held fixed to make a case for causality? Rarely is an econometric study evaluated without raising this issue. In most serious applications, the number of factors that can affect the variable of interest — such as [...] wages— is immense, and the isolation of any particular variable may seem like a hopeless effort. However, we will eventually see that, when carefully applied, econometric methods can simulate a ceteris paribus experiment. [...] As we will see throughout this text, accounting for other observed factors, such as experience, when estimating the ceteris paribus*

> *effect of another variable, such as education, is relatively straightforward. We will also find that accounting for inherently unobservable factors, such as ability, is much more problematic. It is fair to say that many of the advances in econometric methods have tried to deal with unobserved factors in econometric models.* WOOLDRIDGE (2020) PÁGINAS 10-12.

Los Ejemplos 2.1.1-2.1.2 sugieren que la especificación inicial de un modelo de regresión requiere considerar tres cuestiones: (i) la elección de las **variables explicativas**, (ii) la elección de la **forma funcional** de la relación entre la variable dependiente y las variables explicativas, y (iii) el planteamiento de las **hipótesis** que garantizan unas buenas propiedades para los métodos econométricos que serán utilizados posteriormente. Las cuestiones (i)-(ii) suelen resolverse mediante algún tipo de razonamiento lógico basado (quizás) en algún **modelo teórico** y (fundamentalmente) en un análisis inicial detallado de las **características muestrales** de los **datos** disponibles sobre todas las variables. También son importantes en (i)-(ii) la **intuición**, la **experiencia** y el **buen juicio** del analista. La cuestión (iii) suele resolverse planteando de forma tentativa sobre el resultado de (i)-(ii) las hipótesis que justifican el empleo de métodos econométricos óptimos.

2.2 Estimación

Una vez especificado inicialmente un modelo econométrico, el paso siguiente en un análisis aplicado consiste en estimar los parámetros que figuran en el modelo. La **estimación** de un modelo econométrico se refiere a la asignación de valores numéricos concretos a sus parámetros (es decir, al cálculo de **estimaciones** de dichos parámetros) empleando la información contenida en los datos. Buena parte de la econometría teórica está dedicada al diseño y al análisis de las propiedades de diferentes **métodos de estimación** para diferentes modelos, así como a la comparación entre las propiedades de métodos alternativos para cada tipo de modelo. Actualmente, en la econometría aplicada puede utilizarse una gran variedad tanto de modelos como de métodos para estimarlos. No obstante, la utilidad práctica de las estimaciones proporcionadas por cualquier método de estimación siempre puede evaluarse respondiendo a la misma pregunta en todos los casos: ¿Qué grado de **confianza** tenemos en que las estimaciones obtenidas estén **próximas** a lo que realmente valen los parámetros de nuestro modelo?

En la práctica, ese grado de confianza puede evaluarse en función de las **propiedades estadísticas** (como la insesgadez, la eficiencia o la consistencia) que posea el método de

estimación empleado, que dependen crucialmente, a su vez, de cuáles sean las **hipótesis** que conforman el modelo que se pretende estimar y de cuál sea el **grado de adecuación** de éstas a las **pautas muestrales** que están presentes en los datos.

2.2.1 Ejemplo - Propiedades de un Estimador

En este ejemplo se introducen los conceptos de un **estimador** y una **estimación** de un vector de parámetros en relación con un modelo RLM. También se presentan algunas de las propiedades de un estimador asociadas con su posible utilidad en la práctica.

El modelo RLM resumido en [16]-[17] puede interpretarse como una proposición concreta sobre ciertos aspectos de la **distribución de probabilidad** de una colección $\mathbf{M}$ de variables aleatorias,

$$\mathbf{M} = \begin{bmatrix} Y_1 & X_{12} & X_{13} & \cdots & X_{1K} \\ Y_2 & X_{22} & X_{23} & \cdots & X_{2K} \\ \vdots & \vdots & \vdots & & \vdots \\ Y_N & X_{N2} & X_{N3} & \cdots & X_{NK} \end{bmatrix},$$

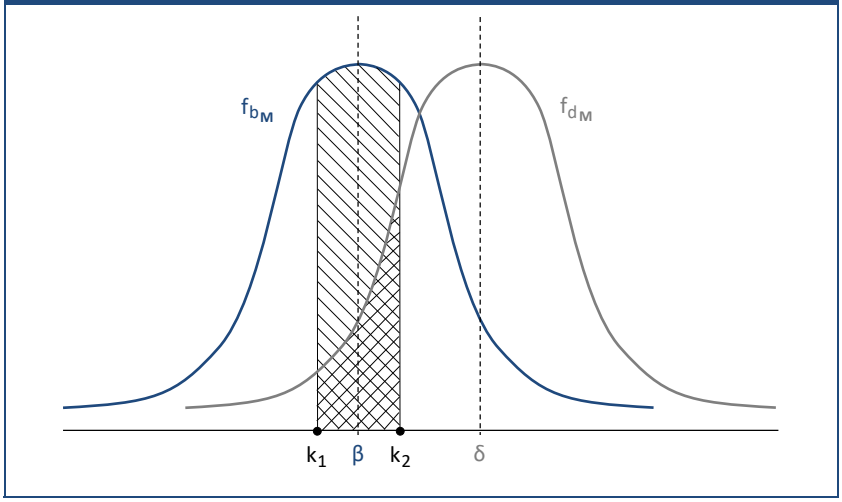
de la que los datos

$$\mathbf{m} = \begin{bmatrix} y_1 & x_{12} & x_{13} & \cdots & x_{1K} \\ y_2 & x_{22} & x_{23} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & & \vdots \\ y_N & x_{N2} & x_{N3} & \cdots & x_{NK} \end{bmatrix}$$

constituyen una realización particular. Para obtener información sobre los parámetros que figuran en [16], a veces es posible diseñar una función $\mathbf{b}(\cdot) = [b_1(\cdot), b_2(\cdot), ..., b_K(\cdot)]'$ tal que la distribución de probabilidad de un **estimador** del tipo $\mathbf{b_M} = \mathbf{b}(\mathbf{M})$ sea conocida (exacta o aproximadamente), e implique una **probabilidad** relativamente elevada de que una **estimación** puntual $\mathbf{b} = \mathbf{b}(\mathbf{m})$ calculada con cualquier realización particular $\mathbf{m}$ de $\mathbf{M}$ esté próxima al verdadero valor de $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_K]'$, independientemente de cuál sea ese valor. Como la distribución de $\mathbf{M}$ depende de $\boldsymbol{\beta}$ (porque cada Y_i en $\mathbf{M}$ depende de $\boldsymbol{\beta}$ a través de [16]) y $\mathbf{b_M}$ es una función de $\mathbf{M}$, la distribución de probabilidad del estimador $\mathbf{b_M}$ también dependerá de $\boldsymbol{\beta}$. Esto implica, en particular, que la esperanza y la varianza de $\mathbf{b_M}$ pueden ser distintas para distintos valores de $\boldsymbol{\beta}$. Por lo tanto, $\mathbf{b_M}$ será un

FIGURA 10

Insesgadez de un Estimador



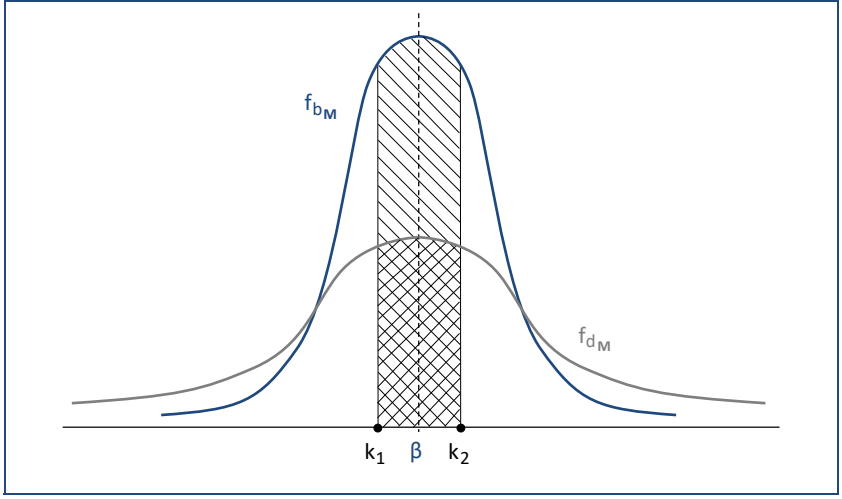
estimador útil de $\boldsymbol{\beta}$ en la práctica cuando su distribución de probabilidad varíe con $\boldsymbol{\beta}$ de tal manera que la probabilidad de que $\mathbf{b_M}$ proporcione estimaciones próximas a $\boldsymbol{\beta}$, sea relativamente elevada con independencia de cuál sea el verdadero valor de $\boldsymbol{\beta}$. La utilidad del estimador $\mathbf{b_M}$ en este sentido puede establecerse cuando $\mathbf{b_M}$ tiene ciertas **propiedades estadísticas** derivadas de su distribución de probabilidad, tales como la **insesgadez**, la **eficiencia** o la **consistencia**.

En la Figura 10 están representadas dos funciones de densidad Normales, $f_{b_\mathbf{M}}$ y $f_{d_\mathbf{M}}$, asociadas con las distribuciones de probabilidad de dos estimadores alternativos, $b_\mathbf{M}$ y $d_\mathbf{M}$, de un parámetro β. En este caso, los dos estimadores tienen las mismas propiedades estadísticas excepto por sus valores esperados: $\mathrm{E}[b_\mathbf{M}] = \beta$, mientras que $\mathrm{E}[d_\mathbf{M}] = \delta$. Esta diferencia implica que la $\Pr[k_1 \leq b_\mathbf{M} \leq k_2]$ es mayor que la $\Pr[k_1 \leq d_\mathbf{M} \leq k_2]$, lo que significa que la probabilidad de obtener una estimación próxima a β es mayor si se emplea el estimador $b_\mathbf{M}$ que si emplea el estimador $d_\mathbf{M}$. Un estimador como $b_\mathbf{M}$, cuyo valor esperado coincide con el verdadero valor de lo que se pretende estimar, se denomina un estimador **insesgado**. Como se ilustra en la Figura 10, a igualdad de otras propiedades, un estimador insesgado es preferible a otro que no lo sea.

En la Figura 11, los dos estimadores considerados tienen las mismas propiedades (en particular, ambos son insesgados) excepto porque $\mathrm{Var}[b_\mathbf{M}] < \mathrm{Var}[d_\mathbf{M}]$. Como en el caso

FIGURA 11

Eficiencia Relativa de un Estimador



anterior, esta diferencia implica que $\Pr[k_1 \leq b_{\mathbf{M}} \leq k_2] > \Pr[k_1 \leq d_{\mathbf{M}} \leq k_2]$. En este caso, se dice que $b_{\mathbf{M}}$ es **relativamente más eficiente** que $d_{\mathbf{M}}$. Como se ilustra en la Figura 11, entre varios estimadores insesgados es preferible el que tenga menor varianza. Cuando un estimador insesgado tiene **varianza mínima** con respecto a **todos** los estimadores insesgados que pueden considerarse, se dice que dicho estimador es un estimador (absolutamente) **eficiente**.

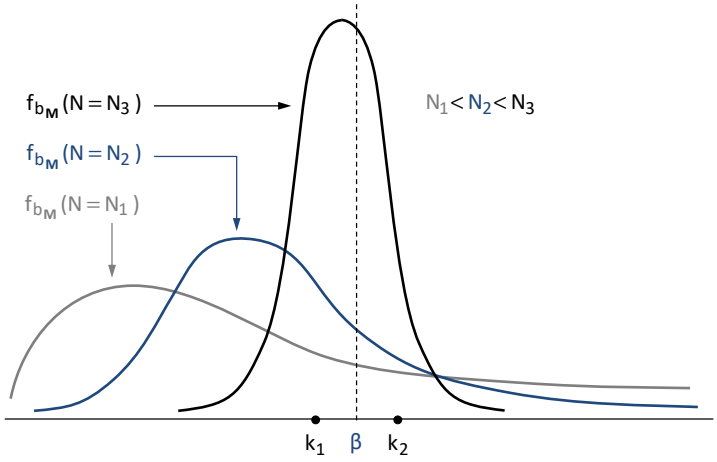
Por último, la Figura 12 de la página siguiente representa el caso de un estimador $b_{\mathbf{M}}$ que, aunque no es insesgado, es **consistente**: aunque $\mathrm{E}[b_{\mathbf{M}}] \neq \beta$ para cualquier tamaño muestral N (finito), la distribución de probabilidad de $b_{\mathbf{M}}$ varía con N de tal manera que la $\Pr[k_1 \leq b_{\mathbf{M}} \leq k_2] \rightarrow 1$ a medida que N aumenta. La consistencia es lo mínimo que debería satisfacer cualquier estimador para poder resultar de utilidad en la práctica.

En resumen, la insesgadez, la eficiencia (relativa o absoluta) y la consistencia son propiedades deseables porque el empleo en la práctica de un estimador con dichas propiedades otorga cierto grado de fiabilidad a las estimaciones que proporciona.

En la práctica, la fiabilidad de un modelo estimado siempre debe juzgarse contrastando el grado de validez de las hipótesis que garantizan, en cada caso, unas buenas propiedades

FIGURA 12

Consistencia de un Estimador



estadísticas para el método de estimación empleado. No obstante, el hecho de que un estimador tenga buenas propiedades estadísticas no debe tomarse como una garantía absoluta de obtener estimaciones que estén próximas a lo que se pretende estimar (por ejemplo, el estimador $b_{\mathbf{M}}$ de las Figuras 10-12 podría proporcionar, en teoría, cualquier estimación de β que estuviera comprendida entre $-\infty$ y $+\infty$). Lo único que realmente permite el empleo de un estimador con buenas propiedades es tener más confianza en las estimaciones obtenidas que si se hubiera empleado un estimador carente de esas propiedades. En el contexto de un análisis econométrico aplicado, en el que la incertidumbre juega un papel fundamental en muchos sentidos, éste es el único tipo de razonamiento que puede justificar el tomar o no tomar determinadas decisiones con el fin de hacer algún progreso en el análisis de los datos.

2.3 Diagnosis y Revisión

La **diagnosis** de un modelo estimado a partir de unos datos se refiere a la verificación o a la refutación, según el caso, de que las hipótesis que conforman el modelo describen adecuadamente las características de los datos empleados para estimarlo. Dado que la finalidad principal de un modelo consiste en resumir adecuadamente la información

contenida en los datos, un modelo puede considerarse una herramienta útil solamente cuando cumple con dicha finalidad. En caso contrario, la utilidad práctica de un modelo es probablemente nula. En particular, si un método de estimación determinado sólo tiene buenas propiedades bajo ciertas condiciones, entonces las estimaciones derivadas de su uso sólo serán fiables si los datos satisfacen razonablemente dichas condiciones.

Al igual que ocurre con la inspección visual de los datos en relación con la especificación inicial de un modelo econométrico (como se ilustra en el Ejemplo 2.1.2), el **análisis gráfico** de los resultados de la estimación de un modelo suele ser bastante revelador en relación con el grado de validez de algunas de las hipótesis que conforman el modelo. La cita siguiente hace referencia a esta cuestión:

> *It's almost always a good idea to begin forecasting projects with graphical data analysis. When compared to the modern array of statistical modelling methods, graphical analysis might seem trivially simple, perhaps even so simple as to be incapable of delivering serious insights into the series to be forecast. Such is not the case: In many respects the human eye is a far more sophisticated tool for data analysis and modelling than even the most sophisticated modern modelling techniques. That's certainly not to say that graphical analysis alone will get the job done —certainly, graphical analysis has its limitations— but it's usually the best place to start. […] Graphics help us summarize and reveal patterns in data […] Graphics help us identify anomalies in data […] Graphics facilitate and encourage comparison of different pieces of data […] Graphics enable us to present a huge amount of data in a small space, and to make huge data sets coherent.* DIEBOLD (2007) PÁGINAS 51 Y 53-54.

En general, este punto de vista es aplicable no sólo a cuestiones relacionadas con la previsión, sino a cualquier cuestión que se pretenda analizar utilizando la información contenida en una colección de datos. El ejemplo siguiente trata sobre el empleo de instrumentos gráficos muy sencillos para de diagnosticar modelos estimados.

2.3.1 Ejemplo - Gráficos como Instrumentos de Diagnosis

En este ejemplo se ilustra la posibilidad de que un modelo RLS estimado no resuma adecuadamente el contenido informativo de una colección de datos. Esta posibilidad no es fácil de detectar si sólo se mira a la información numérica asociada con un modelo estimado, aunque es muy fácilmente detectable mediante un simple examen de alguna representación gráfica de dicha información. La Tabla 5 de la página siguiente contiene datos simulados (artificiales) sobre seis variables, tomados (al igual que la cita del final de

TABLA 5
Datos de Anscombe (1973)

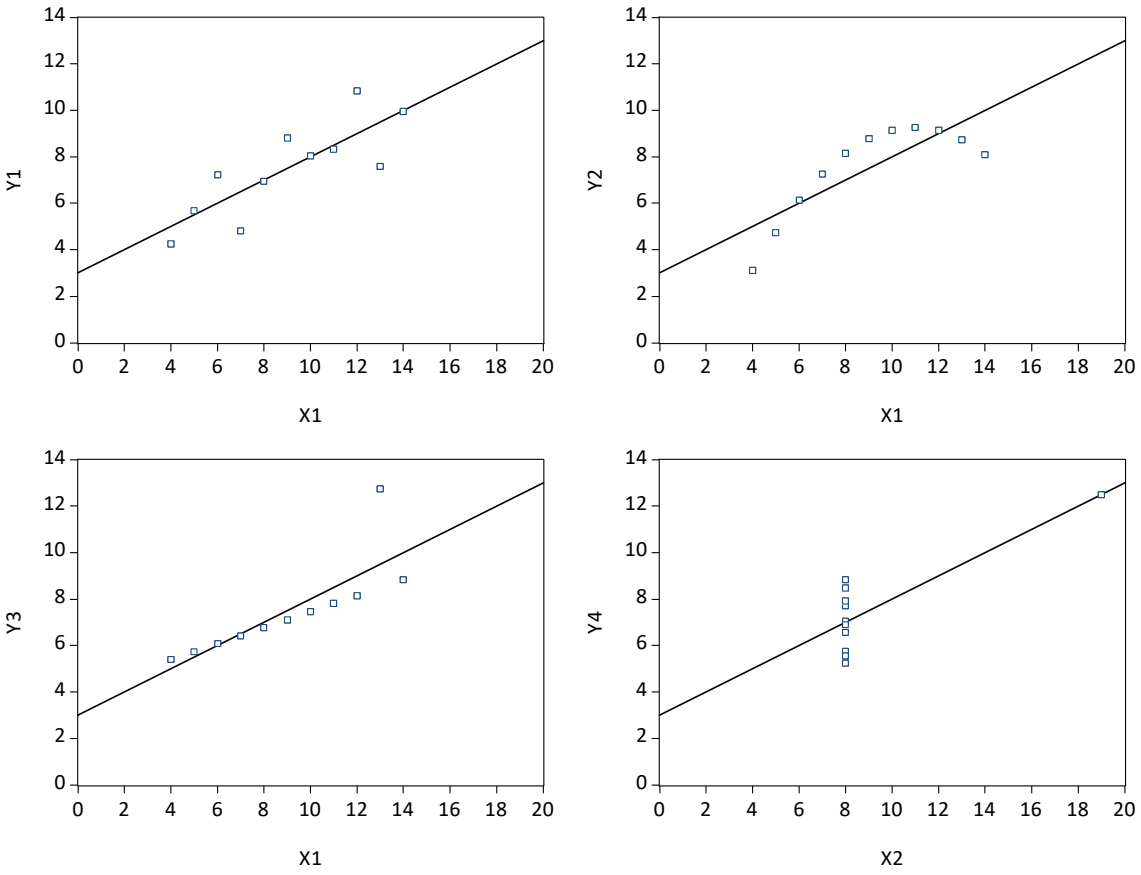
i	Y1	Y2	Y3	X1	Y4	X2
1	8.04	9.14	7.46	10.00	6.58	8.00
2	6.95	8.14	6.77	8.00	5.76	8.00
3	7.58	8.74	12.74	13.00	7.71	8.00
4	8.81	8.77	7.11	9.00	8.84	8.00
5	8.33	9.26	7.81	11.00	8.47	8.00
6	9.96	8.10	8.84	14.00	7.04	8.00
7	7.24	6.13	6.08	6.00	5.25	8.00
8	4.26	3.10	5.39	4.00	12.50	19.00
9	10.84	9.13	8.15	12.00	5.56	8.00
10	4.82	7.26	6.42	7.00	7.91	8.00
11	5.68	4.74	5.73	5.00	6.89	8.00

la Sección 1) de Anscombe (1973) (archivo NUM02-ANSCOMBE.WF1). Con los datos sobre cada uno de los pares de variables (Y_1, X_1), (Y_2, X_1), (Y_3, X_1), (Y_4, X_2), se han estimado por MCO cuatro modelos RLS, lo que ha proporcionado en todos los casos las mismas estimaciones MCO de β_1 (término constante) y β_2 (pendiente): $\hat{\beta}_1 = 3.0$ y $\hat{\beta}_2 = 0.5$. Adicionalmente, en los cuatro casos se han obtenido los mismos **errores estándar** (desviaciones típicas estimadas) para los estimadores MCO de β_1 y β_2 (1.12 y 0.12, respectivamente, que miden el grado de precisión o de fiabilidad del estimador correspondiente; ver la Figura 11 del Ejemplo 2.2.1), y el mismo valor para el **coeficiente de determinación** (un 67%, que representa el porcentaje de la variación total observada en los datos sobre la variable dependiente que el modelo estimado consigue explicar). Estos resultados sugieren que la relación entre cada par de variables podría ser la misma en los cuatro casos considerados.

La Figura 13 de la página siguiente contiene las representaciones gráficas (nubes de puntos) de los datos correspondientes a los cuatro pares de variables considerados, junto con la RLS estimada por MCO en cada caso (que es exactamente la misma en todos los casos). Un simple vistazo a los cuatro gráficos de la Figura 13 indica claramente, a pesar de lo que los resultados numéricos sugieren, que la relación entre cada par de variables **no** es la misma en los cuatro casos considerados. EJERCICIO 12

FIGURA 13

Modelos RLS Estimados por MCO con los Datos de la Tabla 5.
En Todos los Casos $\hat{\beta}_1 = 3.0$ (Término Constante) y $\hat{\beta}_2 = 0.5$ (Pendiente).



En el caso de las variables (Y_1, X_1), el modelo estimado parece razonable: los datos sugieren una relación positiva entre Y_1 y X_1 que está bastante bien recogida por la línea recta $\hat{Y} = 3.0 + 0.5X$. En particular, el nivel medio de los residuos parece independiente de los valores que toma X_1.

En el caso de las variables (Y_2, X_1), la situación es muy diferente: los datos sugieren una relación muy clara entre Y_2 y X_1, aunque dicha relación no es lineal. En este caso, el nivel medio de los residuos depende claramente de los valores que toma X_1.

En el caso de las variables (Y_3, X_1), los datos sugieren una relación lineal positiva entre Y_3 y X_1. Sin embargo, hay un par de datos (el tercero) que no parece formar parte de dicha relación y que ejerce una influencia notable sobre el modelo estimado. También en este caso, el nivel medio de los residuos depende de los valores que toma X_1.

Por último, en el caso de las variables (Y_4, X_2), todos los pares de datos están dispuestos en vertical con la excepción de uno de ellos (el octavo), que ejerce una influencia muy notable sobre el modelo estimado. Con independencia de otras consideraciones, el gráfico correspondiente de la Figura 13 revela inmediatamente el carácter anómalo de este caso.

■

En general, la diagnosis de un modelo de regresión tiene que ver con los dos elementos centrales del modelo estimado: las estimaciones de sus parámetros y sus residuos. Por ejemplo, un signo o una magnitud chocante (que contrasta con lo esperado inicialmente) en un parámetro estimado, debe examinarse cuidadosamente para intentar detectar sus posibles causas y, en su caso, revisarlas y corregirlas. Por su lado, los residuos contienen la parte de la información total disponible sobre la variable dependiente del modelo que queda fuera del modelo estimado. En consecuencia, los residuos de un modelo adecuado no deben contener ningún tipo de información susceptible de poder ser utilizada para revisar y mejorar el modelo (como suele decirse, los residuos deben ser "puramente aleatorios" o "ruido blanco"). De manera equivalente, las propiedades muestrales de los residuos deben ser análogas a las propiedades teóricas supuestas sobre las perturbaciones del modelo. La cita siguiente tiene que ver con la importancia del análisis de residuos:

> *Complicated phenomena, in which several causes concurring, opposing, or quite independent of each other, operate at once, so as to produce a compound effect, may be simplified by subducting the effect of all the known causes, as well as the nature of the case permits, either by deductive reasoning or by appeal to experience, and thus leaving, as it were, a* residual phenomenon *to be explained. It is by this process, in fact, that science, in its present advanced state, is chiefly promoted.* JOHN F.W. HERSCHEL (1830) "A PRELIMINARY DISCOURSE ON THE STUDY OF NATURAL PHILOSOPHY" CITADO EN COOK Y WEISBERG (1982) PÁGINA 1.

En última instancia, la diagnosis de un modelo estimado es útil no sólo para descubrir posibles deficiencias en el modelo, sino también para intentar descubrir pistas que puedan indicar cómo revisarlo y mejorarlo. Actualmente, existe una gran variedad de métodos gráficos, numéricos y estadísticos para diagnosticar y revisar modelos econométricos en la

práctica. Sólo una vez concluidas satisfactoriamente estas operaciones, podrá servir un modelo para lo que en última instancia debe servir: proporcionar respuestas fiables y entendibles a las preguntas planteadas al comienzo de un análisis.

2.3.2 Una Última Cita ...

> *El hombre trata de penetrar los secretos del universo como si lo observara desde el exterior, como si dominara la perspectiva. Piensa que hace el recuento de los mecanismos de la vida cuando él es un agente, una manifestación de esta vida. El hombre es limitado. No tiene todos los datos del problema en la mano. ¡No se puede explicar un fenómeno cuando se es una expresión de este fenómeno! Es absurdo. En todo caso, su visión está condenada a permanecer fragmentaria, incompleta.*
>
> R. SARDOU
> *El Peregrino del Tiempo (L'Eclat de Dieu)*

3 Los Recursos Instrumentales de la Econometría

La incertidumbre juega un papel fundamental en cualquier análisis de datos con el que se pretenda describir el funcionamiento de un sistema que no puede ser caracterizado con total exactitud o previsto con absoluta certeza. Por este motivo, muchos de los métodos que conforman la econometría teórica están tomados directamente o adaptados de la estadística. Otros son métodos desarrollados específicamente para analizar ciertos tipos de datos que son especiales en algún sentido. Todos estos métodos se fundamentan, en mayor o menor medida, en la Teoría de la Probabilidad y en la Inferencia Estadística, que constituyen dos de los recursos instrumentales básicos de la Teoría Econométrica. En consecuencia, un buen conocimiento de estos dos **recursos estadísticos** es imprescindible para entender tanto el desarrollo y el funcionamiento de los métodos econométricos disponibles actualmente, como la manera de aplicarlos en la práctica.

Al mismo tiempo, el desarrollo ordenado y formal de la Teoría de la Probabilidad, de la Inferencia Estadística y, en consecuencia, de la Teoría Econométrica, hace uso de multitud de **recursos matemáticos** sobre lógica, geometría, conjuntos, números reales, vectores, matrices y funciones (por mencionar unos pocos), cuyo conocimiento es absolutamente esencial para entender bien dicho desarrollo.

Por último, la aplicación práctica de la Teoría Econométrica para analizar datos reales requiere inevitablemente (salvo en casos muy simples) el uso de **recursos informáticos** adecuados, que sean capaces de procesar eficientemente grandes colecciones de datos. Dichos recursos están disponibles actualmente en una gran variedad de formatos, aunque, esencialmente, un ordenador personal modesto y un programa de cálculo econométrico sencillo suelen ser recursos suficientes para poder llevar a cabo una gran variedad de análisis econométricos aplicados.

En esta Sección 3 se hacen algunas consideraciones muy breves sobre los recursos citados en los tres párrafos anteriores (estadística, matemáticas e informática). También se mencionan varias referencias bibliográficas donde pueden encontrarse exposiciones completas y detalladas sobre el contenido de dichos recursos.

3.1 Matemáticas y Estadística

Las matemáticas y la estadística utilizadas regularmente en la econometría teórica y aplicada pueden consultarse en multitud de manuales de econometría. Dos de ellos, que también se citan en la Sección 1, son Heij, de Boer, Franses, Kloek y van Dijk (2004, Capítulos 1-2 y Apéndice A), y Wooldridge (2020, Apéndices A-D). Otras referencias más completas y avanzadas son Dhrymes (2013), Johnston (1984, Capítulos 4-5 y Apéndice A) [todo un "clásico" ...], Mittelhammer (2013), y Greene (2018, Apéndices A-E).

A continuación se mencionan brevemente tres cuestiones matemáticas elementales (que son, no obstante, esenciales para dotar de significado práctico a los parámetros de muchos modelos econométricos), y algunas cuestiones sobre esperanzas condicionales (que son muy útiles a la hora de interpretar hipótesis como [8], [9] o [17] en modelos de regresión) y sobre vectores y matrices de variables aleatorias (indispensables en la Teoría Econométrica).

3.1.1 Variaciones

VP.1 Variación absoluta: $\Delta Y = Y_1 - Y_0$.

VP.2 Variación relativa o proporcional: $\frac{\Delta Y}{Y_0} = \frac{Y_1 - Y_0}{Y_0}$.

VP.3 Variación logarítmica: $\Delta \ln Y = \ln Y_1 - \ln Y_0 = \ln \frac{Y_1}{Y_0} = \ln \left[1 + \frac{\Delta Y}{Y_0} \right] \cong \frac{\Delta Y}{Y_0}$.

Observación: La última relación de la línea anterior proporciona una "buena" aproximación solamente cuando

$\Delta Y/Y_0$ es una cantidad "pequeña", en el sentido de que una aproximación de Taylor de primer orden (lineal) al valor de la función $\ln(1+x)$ para valores de x próximos a x_0 proporciona lo siguiente:

$$\ln(1+x) \cong \ln(1+x_0) + \frac{1}{1+x_0} \times [(1+x) - (1+x_0)] = \ln(1+x_0) + \frac{x-x_0}{1+x_0} = x \text{ cuando } x_0 = 0.$$

VP.4 Variación porcentual: $\%\Delta Y = 100\,\frac{\Delta Y}{Y_0} \cong 100\Delta \ln Y$.

3.1.2 Elasticidad - Rendimientos

ER.1 Si $Y = g(X)$, la **elasticidad** de Y con respecto a X es $l_{YX} = \frac{\%\Delta Y}{\%\Delta X} \cong \frac{\Delta \ln Y}{\Delta \ln X}$, de manera que l_{YX} representa la **variación porcentual** que tiene lugar en Y cuando X varía un 1% ($\%\Delta Y = l_{YX} \times \%\Delta X$).

ER.2 Si $Y = g(X_1, X_2) = aX_1^b X_2^c$ (una **función** de tipo **Cobb-Douglas**), entonces ocurre que $\ln Y = \ln a + b \ln X_1 + c \ln X_2$, por lo que $l_{YX_1} = b$ y $l_{YX_2} = c$. Al mismo tiempo, $g(\mu X_1, \mu X_2) = a(\mu X_1)^b (\mu X_2)^c = \mu^{b+c}(aX_1^b X_2^c)$, por lo que en este caso $g(X_1, X_2)$ es una función **homogénea** de grado $b + c$, es decir, $g(X_1, X_2)$ presenta rendimientos **constantes** si $b + c = 1$, **crecientes** si $b + c > 1$, y **decrecientes** si $b + c < 1$.

3.1.3 Interpretación de Parámetros

IP.1 En la función
$$F = \beta_1 + \beta_2 P\,, \qquad [23]$$
$\frac{dF}{dP} = \beta_2 \Rightarrow \beta_2 = \frac{\Delta F}{\Delta P}$, o bien
$$\Delta F = \beta_2 \Delta P\,. \qquad [24]$$

Por lo tanto, β_2 en [23] es (exactamente) igual a la respuesta (variación) absoluta de F ante una variación absoluta unitaria en P.

IP.2 En la función
$$\ln F = \beta_1 + \beta_2 \ln P\,, \qquad [25]$$
$F = e^{\ln F} = e^{\beta_1 + \beta_2 \ln P} \Rightarrow \frac{dF}{dP} = \frac{\beta_2}{P} e^{\beta_1 + \beta_2 \ln P} = \beta_2 \frac{F}{P} \Rightarrow \beta_2 \cong \frac{\Delta F}{\Delta P}\frac{P}{F} \Rightarrow \frac{\Delta F}{F} \cong \beta_2 \frac{\Delta P}{P}$; en este caso, multiplicando por 100,

$$\%\Delta F \cong \beta_2 \%\Delta P\,. \qquad [26]$$

Por lo tanto, β_2 en [25] es la **elasticidad** (aproximada) de F con respecto a P, es decir, la respuesta (aproximada) en % de F ante una variación de un 1% en P.

IP.3 En la función $$\ln F = \beta_1 + \beta_2 P\,, \qquad [27]$$

$F = e^{\ln F} = e^{\beta_1+\beta_2 P} \Rightarrow \frac{dF}{dP} = \beta_2 e^{\beta_1+\beta_2 P} = \beta_2 F \Rightarrow \beta_2 \cong \frac{\Delta F}{\Delta P}\frac{1}{F} \Rightarrow \frac{\Delta F}{F} \cong \beta_2 \Delta P$; por lo tanto, multiplicando por 100,

$$\%\Delta F \cong 100\,\beta_2 \Delta P\,. \qquad [28]$$

En relación con [27], $100\beta_2$ es la **semielasticidad** (aproximada) de F con respecto a P, es decir, la respuesta (aproximada) en % de F ante una variación absoluta unitaria en P.

IP.4 En la función $$F = \beta_1 + \beta_2 \ln P\,, \qquad [29]$$

$\frac{dF}{dP} = \beta_2 \frac{1}{P} \Rightarrow \beta_2 \cong \frac{\Delta F}{\Delta P} P \Rightarrow \Delta F \cong \beta_2 \frac{\Delta P}{P}$; por lo tanto, multiplicando y dividiendo el lado derecho por 100,

$$\Delta F \cong \frac{\beta_2}{100}\,\%\Delta P\,. \qquad [30]$$

En relación con [29], $\frac{\beta_2}{100}$ es la respuesta absoluta (aproximada) de F provocada por una variación de un 1% en P.

IP.5 En la función $$F = \beta_1 + \beta_2 P + \beta_3 P^2\,, \qquad [31]$$

$\frac{dF}{dP} = \beta_2 + 2\beta_3 P$, $\frac{d^2F}{dP^2} = 2\beta_3$, $\frac{dF}{dP} = 0 \Leftrightarrow P^* = -\frac{\beta_2}{2\beta_3}$; por lo tanto, P^* es un máximo (mínimo) de F con respecto a P si $\beta_3 < 0$ ($\beta_3 > 0$). Adicionalmente, el valor P^* que minimiza (maximiza) F es el mismo que minimiza (maximiza) $\ln F$ [porque $\ln(\cdot)$ es una función monótona estrictamente creciente].

3.1.4 Esperanzas Condicionales

EC.1 Si U y X son dos variables aleatorias continuas con función de densidad (pdf, del inglés *probability density function*) conjunta $f_{UX}(u,x)$, y $g(\cdot)$ es una función real continua, entonces:

Esperanza de g(·): $$\mathrm{E}[g(U,X)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(u,x) f_{UX}(u,x)\,du\,dx. \qquad [32]$$

Pdfs Marginales: $$f_U(u) = \int_{-\infty}^{+\infty} f_{UX}(u,x)\,dx, \qquad f_X(x) = \int_{-\infty}^{+\infty} f_{UX}(u,x)\,du. \qquad [33]$$

Esperanzas Marginales:

$$\begin{aligned} \mathrm{E}[U] &= \textstyle\int_{-\infty}^{+\infty} u f_U(u) du, \\ \mathrm{E}[X] &= \textstyle\int_{-\infty}^{+\infty} x f_X(x) dx. \end{aligned} \qquad [34]$$

Pdfs Condicionales:

$$\begin{aligned} f_{U|X}(u,x) &= \frac{1}{f_X(x)} \times f_{UX}(u,x), \\ f_{X|U}(u,x) &= \frac{1}{f_U(u)} \times f_{UX}(u,x). \end{aligned} \qquad [35]$$

Esperanzas Condicionales:

$$\begin{aligned} \mathrm{E}[U \,|\, X] &= \textstyle\int_{-\infty}^{+\infty} u f_{U|X}(u,X) du, \\ \mathrm{E}[X \,|\, U] &= \textstyle\int_{-\infty}^{+\infty} x f_{X|U}(U,x) dx. \end{aligned} \qquad [36]$$

EC.2 Si $g(\cdot)$ es una función real continua, y U y X son dos variables aleatorias continuas con pdfs y esperanzas como en [33]-[36], entonces

Law of Iterated Expectations: $\mathrm{E}[g(U,X)] = \mathrm{E}[\mathrm{E}[g(U,X) \,|\, X]]$. [37]

Demostración:

$$\begin{aligned} \mathrm{E}[g(U,X) \,|\, X] &= \textstyle\int_{-\infty}^{+\infty} g(u,X) f_{U|X}(u,X) du = h(X), \text{ por lo que} \\ \mathrm{E}[\mathrm{E}[g(U,X) \,|\, X]] &= \mathrm{E}[h(X)] = \textstyle\int_{-\infty}^{+\infty} h(x) f_X(x) dx \\ &= \textstyle\int_{-\infty}^{+\infty} \Big[\int_{-\infty}^{+\infty} g(u,x) f_{U|X}(u,x) du \Big] f_X(x) dx \\ &= \textstyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(u,x) f_{UX}(u,x) du dx = \mathrm{E}[g(U,X)]. \end{aligned}$$

■

EC.3 Si U y X son dos variables aleatorias continuas con pdfs y esperanzas como en [33]-[36], entonces

$$\mathrm{E}[U \,|\, X] = \mathrm{E}[U] \Rightarrow \mathrm{Cov}[U,X] = 0 \Rightarrow \mathrm{Corr}[U,X] = 0. \qquad [38]$$

Demostración:

$$\begin{aligned} \mathrm{Cov}[U,X] &= \mathrm{E}[(U - \mathrm{E}[U])(X - \mathrm{E}[X])] = \mathrm{E}[UX] - \mathrm{E}[U] \times \mathrm{E}[X] \\ &= \mathrm{E}[\mathrm{E}[UX \,|\, X]] - \mathrm{E}[U] \times \mathrm{E}[X] \\ &= \mathrm{E}[\underbrace{\mathrm{E}[U \,|\, X]}_{\mathrm{E}[U]} \times X] - \mathrm{E}[U] \times \mathrm{E}[X] = \mathrm{E}[\mathrm{E}[U] \times X] - \mathrm{E}[U] \times \mathrm{E}[X] \\ &= \mathrm{E}[U] \times \mathrm{E}[X] - \mathrm{E}[U] \times \mathrm{E}[X] = 0. \end{aligned}$$

■

El resultado [38] permite, por ejemplo, interpretar la hipótesis en la segunda parte de [8] en términos de **ausencia de correlación** entre la perturbación y la variable explicativa de un modelo RLS en cada observación o punto muestral. Análogamente, la hipótesis [17] implica la ausencia de correlación entre la perturbación y **todas** las variables explicativas de un modelo RLM en cada observación o punto muestral.

3.1.5 Vectores de Esperanzas y Matrices de Varianzas-Covarianzas

Si $\mathbf{W}$ y $\mathbf{X}$ son matrices $N \times M$ y $N \times K$, respectivamente, de variables aleatorias, entonces los símbolos $\mathrm{E}[\mathbf{W}]$ y $\mathrm{E}[\mathbf{W} \,|\, \mathbf{X}]$ representan matrices de orden $N \times M$ cuyos elementos en las posiciones ij $(i = 1, 2, ..., N; j = 1, 2, ..., M)$ son $\mathrm{E}[W_{ij}]$ y $\mathrm{E}[W_{ij} \,|\, \mathbf{X}]$, respectivamente. En particular, si $\mathbf{V} = [V_1, V_2, ..., V_N]'$ es un vector $N \times 1$ (columna) de variables aleatorias (el símbolo $'$ representa la operación de **trasposición**), entonces

$$\mathrm{E}[\mathbf{V}] = \begin{bmatrix} \mathrm{E}[V_1] \\ \mathrm{E}[V_2] \\ \vdots \\ \mathrm{E}[V_N] \end{bmatrix}, \; \mathrm{E}[\mathbf{V} \,|\, \mathbf{X}] = \begin{bmatrix} \mathrm{E}[V_1 \,|\, \mathbf{X}] \\ \mathrm{E}[V_2 \,|\, \mathbf{X}] \\ \vdots \\ \mathrm{E}[V_N \,|\, \mathbf{X}] \end{bmatrix} \qquad [39]$$

son los **vectores de esperanzas** (medias) incondicionales y condicionales, respectivamente, de $\mathbf{V}$. Por otro lado,

$$(\mathbf{V} - \mathrm{E}[\mathbf{V}])(\mathbf{V} - \mathrm{E}[\mathbf{V}])' = \begin{pmatrix} V_1 - \mathrm{E}[V_1] \\ V_2 - \mathrm{E}[V_2] \\ \vdots \\ V_N - \mathrm{E}[V_N] \end{pmatrix} (V_1 - \mathrm{E}[V_1], V_2 - \mathrm{E}[V_2], ..., V_N - \mathrm{E}[V_N]) =$$

$$= \begin{bmatrix} (V_1 - \mathrm{E}[V_1])^2 & (V_1 - \mathrm{E}[V_1])(V_2 - \mathrm{E}[V_2]) & \cdots & (V_1 - \mathrm{E}[V_1])(V_N - \mathrm{E}[V_N]) \\ (V_2 - \mathrm{E}[V_2])(V_1 - \mathrm{E}[V_1]) & (V_2 - \mathrm{E}[V_2])^2 & \cdots & (V_2 - \mathrm{E}[V_2])(V_N - \mathrm{E}[V_N]) \\ \vdots & \vdots & \ddots & \vdots \\ (V_N - \mathrm{E}[V_N])(V_1 - \mathrm{E}[V_1]) & (V_N - \mathrm{E}[V_N])(V_2 - \mathrm{E}[V_2]) & \cdots & (V_N - \mathrm{E}[V_N])^2 \end{bmatrix}.$$

Por lo tanto, teniendo en cuenta que $\mathrm{E}[(V_i - \mathrm{E}[V_i])^2] = \mathrm{Var}[V_i]$ $(i = 1, 2, ..., N)$, y que $\mathrm{E}[(V_i - \mathrm{E}[V_i])(V_j - \mathrm{E}[V_j])] = \mathrm{Cov}[V_i, V_j]$ $(i, j = 1, 2, ..., N; i \neq j)$, aplicando el operador

E[·] a la matriz $(\mathbf{V}-\mathrm{E}[\mathbf{V}])(\mathbf{V}-\mathrm{E}[\mathbf{V}])'$ anterior resulta que

$$\mathrm{E}[(\mathbf{V}-\mathrm{E}[\mathbf{V}])(\mathbf{V}-\mathrm{E}[\mathbf{V}])'] = \begin{bmatrix} \mathrm{Var}[V_1] & \mathrm{Cov}[V_1,V_2] & \cdots & \mathrm{Cov}[V_1,V_N] \\ \mathrm{Cov}[V_2,V_1] & \mathrm{Var}[V_2] & \cdots & \mathrm{Cov}[V_2,V_N] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[V_N,V_1] & \mathrm{Cov}[V_N,V_2] & \cdots & \mathrm{Var}[V_N] \end{bmatrix},$$

que se denomina la **matriz de varianzas-covarianzas** (o la matriz de varianzas, o la matriz de covarianzas) incondicionales de $\mathbf{V}$:

$$\mathrm{Var}[\mathbf{V}] = \mathrm{E}[(\mathbf{V}-\mathrm{E}[\mathbf{V}])(\mathbf{V}-\mathrm{E}[\mathbf{V}])']. \qquad [40]$$

La matriz $\mathrm{Var}[\mathbf{V}]$ es una matriz cuadrada $(N\times N)$ y simétrica. Los elementos en su diagonal principal son las varianzas de los componentes individuales de $\mathbf{V}$, y los elementos fuera de su diagonal principal son las covarianzas entre cada par de componentes de $\mathbf{V}$:

$$\mathrm{Var}[\mathbf{V}]_{ii} = \mathrm{Var}[V_i] \text{ para } i = 1,2,...,N,$$
$$\mathrm{Var}[\mathbf{V}]_{ij} = \mathrm{Cov}[V_i,V_j] \text{ para } i,j = 1,2,...,N \text{ con } i \neq j.$$

Análogamente, la matriz de varianzas-covarianzas condicionales de $\mathbf{V}$ es

$$\mathrm{Var}[\mathbf{V}\,|\,\mathbf{X}] = \mathrm{E}[(\mathbf{V}-\mathrm{E}[\mathbf{V}\,|\,\mathbf{X}])(\mathbf{V}-\mathrm{E}[\mathbf{V}\,|\,\mathbf{X}])'\,|\,\mathbf{X}]. \qquad [41]$$

Como añadido importante a [39]-[40], un vector $\mathbf{V} = [V_1,V_2,...,V_N]'$ de variables aleatorias sigue una distribución Normal multivariante con vector de medias $\mathrm{E}[\mathbf{V}] = \boldsymbol{\mu}$ y matriz de varianzas-covarianzas $\mathrm{Var}[\mathbf{V}] = \boldsymbol{\Sigma}$ sí y sólo sí su función de densidad es

$$f(\mathbf{v}) = (2\pi)^{-N/2}\,|\boldsymbol{\Sigma}|^{-1/2}\,\exp\left[-\tfrac{1}{2}(\mathbf{v}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right], \qquad [42]$$

lo que suele representarse como $\mathbf{V} \sim \mathrm{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

3.2 Recursos Informáticos

Aunque actualmente existen tanto ordenadores como programas de análisis de datos muy potentes y sofisticados, lo cierto es que (salvo en casos especiales) un ordenador personal

modesto y un software sencillo son recursos suficientes para poder llevar a cabo una gran variedad de análisis econométricos en la práctica, cómoda y eficientemente.

En la guía *Introducción al Uso de EViews 4.1* disponible en ucm.randomshock.com se explica cómo utilizar el programa EViews (versión 4.1 para Windows) en relación con varios de los temas que suelen impartirse en los estudios de grado de muchas universidades en dos cuatrimestres de econometría. Existen, desde luego, muchas alternativas (como versiones más recientes de EViews, programas libres y de pago, y documentación gratuita en Internet), por lo que cualquier usuario de la econometría aplicada (estudiante, analista, investigador) puede elegir su software a la medida de sus gustos y sus necesidades.

Resumen

La toma de decisiones en muchos contextos tanto sociales como naturales se basa con frecuencia en el **análisis de datos**. Los datos reflejan el funcionamiento real de algún sistema cuyo entendimiento es importante para tomar decisiones razonadamente.

La **econometría aplicada** moderna trata de cómo analizar datos para responder a preguntas diversas referidas a sistemas cuyo funcionamiento es imposible de caracterizar con total exactitud o de prever con absoluta certeza.

Los métodos estadísticos y matemáticos que se utilizan en la econometría aplicada para analizar datos conforman, en conjunto, lo que suele denominarse **econometría teórica** o **métodos econométricos**.

El punto de partida de un análisis econométrico aplicado consiste en el **planteamiento preciso** de una **pregunta concreta** sobre algún aspecto de un sistema dado, cuya **respuesta** se pretende obtener usando la **evidencia empírica** contenida en una colección de datos.

Muchas preguntas que se plantean en la econometría aplicada tratan sobre la **evaluación de efectos causales** entre variables, o sobre la **previsión** de cantidades desconocidas.

Evaluar de manera fiable el efecto causal o directo de una variable sobre otra utilizando **datos no experimentales** es una tarea difícil, especialmente cuando la influencia que se pretende aislar está relacionada con otras que no se consideran de forma explícita en el análisis. La presencia de **efectos indirectos** asociados con esta posibilidad puede llevar a la

estimación de **relaciones espurias**, carentes de legitimidad o de autenticidad. Por este motivo, dos elementos centrales en la elaboración de un análisis econométrico aplicado son los que tienen que ver con qué **variables** se **incluyen** explícitamente en el análisis y cuál es su posible **relación** con otras variables que se **omiten**.

Cada dato que se emplea en un análisis econométrico aplicado es un valor numérico de cierta característica de una **entidad observable** que forma parte del sistema social o natural al que se refiere la pregunta planteada al comienzo del análisis. En función de cómo se haya llevado a cabo la observación de dicho sistema, los datos resultantes pueden ser de **sección cruzada** (transversales), de **series temporales**, o de **panel** (longitudinales).

Una vez escogidos los datos que se van a utilizar para resolver el problema planteado al comienzo de un análisis, el paso siguiente consiste en resumir la información cuantitativa contenida en los datos mediante un **modelo econométrico**, como un modelo de **regresión lineal simple** (RLS), o un modelo de **regresión lineal múltiple** (RLM).

Un modelo se plantea como un conjunto de **hipótesis plausibles** que pretenden resumir algún aspecto del funcionamiento real de un sistema reflejado en unos datos. Para que un modelo pueda dar respuestas fiables a las preguntas formuladas al comienzo de un análisis, las hipótesis que dan forma al modelo deben adecuarse tanto a dichas preguntas como a las propiedades de los datos que se emplearán para intentar responderlas.

Las respuestas que proporciona un modelo elaborado a partir de unos datos, pueden ir desde una mera **descripción cuantitativa** de algún aspecto de un sistema, hasta una serie de **previsiones** muy valoradas a la hora de tomar decisiones importantes.

La elaboración de un modelo requiere inicialmente la **especificación** de las hipótesis que le dan forma y la **estimación** de los **parámetros** que figuran en dicha especificación. Tanto la especificación inicial de un modelo como la estimación de sus parámetros son pasos importantes en un análisis aplicado, pero no proporcionan por sí solos ninguna garantía para la obtención de respuestas fiables.

La **fiabilidad** de las **respuestas** proporcionadas por un modelo elaborado a partir de unos datos depende fundamentalmente (sobre todo en entornos no experimentales) del grado de **compatibilidad** entre las **hipótesis** que conforman el modelo y las **pautas muestrales** que están realmente presentes en los **datos**. Sin una evaluación adecuada de ese grado de compatibilidad, es imposible juzgar la fiabilidad de las conclusiones obtenidas al final de un

análisis. Por este motivo, la **diagnosis** y, en su caso, la **revisión** de un modelo constituyen probablemente la etapa esencial de cualquier análisis econométrico aplicado.

Aunque existen muchos métodos útiles y más o menos sofisticados tanto para especificar como para diagnosticar y, en su caso, revisar un modelo, el **análisis gráfico** de los datos y de otras cantidades asociadas con la elaboración de un modelo suele ser el mejor punto de partida para su especificación y su diagnosis.

Bibliografía

Las citas en las páginas 1 y 43 están tomadas de dos obras literarias bien conocidas, que han sido publicadas por diversas editoriales en fechas y ediciones variadas. Por su parte, las referencias siguientes han sido utilizadas como fuentes de citas, ejemplos y datos, o han sido mencionadas por algún otro motivo, a lo largo del texto:

- Anscombe, F.J. (1973), *Graphs in Statistical Analysis*, The American Statistician 27 (1) pp. 17-21.
- Baltagi, B.H. (2008), *Econometrics (Fourth Edition)*, Springer.
- Berndt, E.R. (1991), *The Practice of Econometrics*, Addison-Wesley.
- Cook, D., Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman & Hall.
- Dhrymes, P.J. (2013), *Mathematics for Econometrics (Fourth Edition)*, Springer.
- Diebold, F.X. (2007), *Elements of Forecasting (Fourth Edition)*, Thomson South-Western.
- Franses, P.H. (2002), *A Concise Introduction to Econometrics*, Cambridge University Press.
- Greene, W.H. (2018), *Econometric Analysis (Eighth Edition)*, Pearson.
- Heij, C., de Boer, P., Franses, P.H., Kloek, T., van Dijk, H.K. (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press.
- Johnston, J. (1984), *Econometric Methods (Third Edition)*, McGraw-Hill.
- Mittelhammer, R.C. (2013), *Mathematical Statistics for Economics and Business (Second Edition)*, Springer.
- Wooldridge, J.M. (2020), *Introductory Econometrics - A Modern Approach (Seventh Edition)*, Cengage.

El Capítulo 1 de Baltagi (2008), el Capítulo 1 de Berndt (1991), el libro de Franses (2002) y el Capítulo 1 de Wooldridge (2020) son lecturas introductorias muy recomendables. También lo son el libro de Franses (2018), el Capítulo 1 de Hill, Griffiths y Lim (2018), los Capítulos 1 y 22 de Kennedy (2008), y los Capítulos 1-3 de Stock y Watson (2019), cuyas referencias completas son las siguientes:

- Franses, P.H. (2018), *Enjoyable Econometrics*, Cambridge University Press.
- Hill, R.C., Griffiths, W.E., Lim, G.C. (2018), *Principles of Econometrics (Fifth Edition)*, Wiley.
- Kennedy, P. (2008), *A Guide to Econometrics (Sixth Edition)*, Blackwell.
- Stock, J.H., Watson, M.W. (2019), *Introduction to Econometrics (Fourth Edition)*, Pearson.

Ejercicios

EJERCICIO 1

Explicar a quién y para qué podría interesar resolver cada uno de los problemas planteados en el Ejemplo 1.1.1. Para ello, replantear cada problema de la forma más precisa posible, refiriéndolo a una situación o a un entorno concreto.

EJERCICIO 2

En relación con la ecuación [1] de la Definición 1.1.2, explicar qué podrían representar Y, X, W y V en cada uno de los problemas planteados en el Ejemplo 1.1.1. Discutir la posible independencia de W y de V con respecto a X en cada caso.

EJERCICIO 3

Diseñar varios experimentos del tipo de alguno de los dos descritos en el Ejemplo 1.1.3 que permitan obtener datos referidos a los problemas P01, P02, P03, P06 y P10 del Ejemplo 1.1.1. Explicar en cada caso las dificultades prácticas asociadas con la ejecución de dichos experimentos.

EJERCICIO 4

Indicar para cuáles de los problemas planteados en el Ejemplo 1.1.1 sería razonable utilizar

datos de sección cruzada.

EJERCICIO 5

Indicar para cuáles de los problemas planteados en el Ejemplo 1.1.1 sería razonable utilizar datos de series temporales.

EJERCICIO 6 APÉNDICE A.2

Elaborar los gráficos de las Figuras 2-3. [Requiere EViews o similar.]

EJERCICIO 7 APÉNDICE A.2

Explicar por qué la presencia del término constante en un modelo RLS garantiza que la hipótesis de que $\mathrm{E}[U_i] = 0$ sea perfectamente asumible.

EJERCICIO 8 APÉNDICE A.2

[A] Elaborar los gráficos de las Figuras 6-7 y estimar por MCO las cuatro regresiones correspondientes. [Requiere EViews o similar.] [B] Comprobar que la estimación MCO [10] de β_2 es la misma que proporciona el **método de momentos** aplicado al modelo RLS [9].

EJERCICIO 9

Discutir la posible independencia entre U y X (y, por lo tanto, la fiabilidad de las estimaciones MCO) en los cuatro modelos RLS estimados en el Ejemplo 1.4.1.

EJERCICIO 10

Indicar cuáles de las respuestas obtenidas en el Ejemplo 1.4.2 encajan en alguno de los tres tipos de respuestas considerados en el Ejemplo 1.4.1.

EJERCICIO 11

Justificar e interpretar cada uno de los modelos [19], [21] y [22] del Ejemplo 2.1.2.

EJERCICIO 12 APÉNDICE A.2

Comprobar todos los resultados numéricos y gráficos del Ejemplo 2.3.1. [Requiere EViews o similar.]

EJERCICIO DE REPASO

Al comienzo del primer trimestre del año 2021 (día 1 de enero de 2021), pretendemos decidir cómo invertir a tres meses (hasta el día 31 de marzo) cierta cantidad de dinero en el mercado secundario de la deuda pública. Consideramos dos inversiones alternativas:

(1) Comprar un valor a 3 meses el día 1 de enero y esperar a su vencimiento el día 31 de marzo para hacerlo efectivo.

(2) Comprar un valor a 6 meses el día 1 de enero y, en vez de esperar a su vencimiento el 30 de junio, venderlo tres meses después de comprarlo (el día 31 de marzo) por el precio que tenga entonces en el mercado un valor a tres meses.

Los dos valores considerados se compran al descuento, lo que significa que:

(A) El día 1 de enero se conoce exactamente la rentabilidad asociada con la operación (1), ya que tanto el precio como el valor de reembolso del valor a 3 meses se contratan en el momento de su compra (el día 1 de enero).

(B) Para la operación (2), el día 1 de enero se puede contratar el precio de compra (así como el valor de reembolso para 6 meses después) del valor a 6 meses, pero no se sabe a qué precio se podrá vender dicho valor tres meses después (el día 31 de marzo), por lo que la rentabilidad asociada con la operación (2) no se conoce en el momento de invertir (el día 1 de enero).

Conocida la rentabilidad de la operación (1) en el momento de invertir, el problema consiste en cómo prever la rentabilidad de la operación (2) en dicho momento, para decidir en cuál de las dos operaciones invertir nuestro dinero.

DATOS

Para resolver este problema, disponemos de datos trimestrales sobre las rentabilidades observadas de las operaciones (1) y (2) desde el primer trimestre del año 2011 hasta el último trimestre del año 2020, ambos inclusive (40 trimestres consecutivos). Al comienzo del primer trimestre de 2021 (el día 1 de enero de 2021), también sabemos cuál será la rentabilidad de la operación (1) para dicho trimestre, pero no sabemos cuál será la rentabilidad de la operación (2). Si representamos las rentabilidades de ambas operaciones como X e Y, respectivamente, entonces el día 1 de enero de 2021 contamos con la siguiente

información:

Número de observación	Fecha (trimestre)	Rentabilidad operación (1): X	Rentabilidad operación (2): Y
1	2011:1	x_1	y_1
2	2011:2	x_2	y_2
⋮	⋮	⋮	⋮
39	2020:3	x_{39}	y_{39}
40	2020:4	x_{40}	y_{40}
	2021:1	Conocida (x_*)	Desconocida

MODELO

$$Y = \beta_1 + \beta_2 X + U \quad \rightarrow \quad \text{Estimación} \quad \rightarrow \quad Y = \hat{Y} + \hat{U} = \hat{\beta}_1 + \hat{\beta}_2 X + \hat{U} \, .$$

Pregunta 1: Indique de qué tipo son los datos que figuran en la tabla anterior.

Pregunta 2: En el modelo $Y = \beta_1 + \beta_2 X + U$, indique cuáles son y qué representan las variables, cuáles son y qué representan los parámetros, y qué representa el término U.

Pregunta 3: Explique qué diferencias hay entre lo que representa el símbolo β_2 (sin sombrero) y lo que representa el símbolo $\hat{\beta}_2$ (con sombrero). Haga lo mismo con lo que representan los símbolos U (sin sombrero) y $\hat{U}$ (con sombrero).

Pregunta 4: Explique cómo calcularía, a partir del modelo estimado, una previsión de la rentabilidad de la operación (2) para el primer trimestre de 2021. Si representamos dicha previsión como $\hat{y}_*$, indique cuál es la expresión analítica de $\hat{y}_*$ en términos de x_*. Explique cómo utilizaría las cantidades x_* (un dato) e $\hat{y}_*$ (una previsión) para decidir en cuál de las dos operaciones invertir.

Pregunta 5: Si el símbolo Y_* (en mayúscula y sin sombrero) representa la rentabilidad de la operación (2) en el primer trimestre de 2021 (desconocida), indique cuándo se conocerá el valor real (observado) de Y_*. Por último, explique cómo utilizaría el valor estimado de una probabilidad del tipo $\Pr[Y_* \geq x_*]$ para decidir en cuál de las dos operaciones invertir.

Las respuestas a estas cinco preguntas se encuentran en la página siguiente ...

Respuesta 1: Son datos de series temporales.

Respuesta 2: Las variables son la variable dependiente Y [rentabilidad de la operación (2)], y la variable explicativa X [rentabilidad de la operación (1)]. Los parámetros son el término constante $\beta_1 = \mathrm{E}[Y \,|\, X = 0]$ (siempre que U y X sean independientes en el sentido de que $\mathrm{E}[U \,|\, X] = 0$), y la pendiente $\beta_2 = \partial Y / \partial X$. El término U (error o perturbación) es la parte de Y que difiere de (no considerada en) $\beta_1 + \beta_2 X$.

Respuesta 3: El símbolo β_2 representa lo que realmente vale el efecto causal de X sobre Y. Es una cantidad desconocida. El símbolo $\hat{\beta}_2$ representa una estimación numérica (calculable fácilmente) de dicha cantidad. El símbolo U representa todas las influencias que realmente recibe Y no consideradas explícitamente en el modelo a través del término $\beta_1 + \beta_2 X$. Muchas de esas influencias suelen ser difíciles o imposibles de observar. El símbolo $\hat{U}$ representa la diferencia entre cualquier valor observado (dato) de la variable dependiente y el valor ajustado o previsto correspondiente $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ de acuerdo con el modelo estimado. En ocasiones, $\hat{U}$ (residuo) se interpreta como una estimación numérica del contenido real (desconocido) de U (error o perturbación).

Respuesta 4: La previsión se calcularía como $\hat{y}_* = \hat{\beta}_1 + \hat{\beta}_2 x_* = \hat{\mathrm{E}}[Y \,|\, X = x_*]$. Si $\hat{y}_*$ fuera significativamente mayor [menor] que x_*, sería recomendable invertir en la operación (2) [(1)].

Respuesta 5: El valor de Y_* se conocerá al finalizar el primer trimestre de 2021 (el día 31 de marzo de 2021). Si el valor estimado de $\Pr[Y_* \geq x_*]$ fuera grande [pequeño], sería razonable invertir en la operación (2) [(1)].

APÉNDICE

A.1 Archivos de Datos

En la lista siguiente figuran los nombres de los archivos de datos (WF1) para EViews utilizados a lo largo del texto. También se indican los apartados donde se menciona cada uno de ellos. Los archivos se encuentran en ucm.randomshock.com/data/Intro-Ectr.zip.

SC01-VIVIENDAS.WF1	Apartados 1.2, 1.3.
ST01-PINKHAM.WF1	Apartados 1.2, 1.3.
PA01-SALARIOS1.WF1	Apartado 1.2.
PA02-SALARIOS2.WF1	Apartado 1.2.
SC02-RECIENNACIDOS.WF1	Apartado 1.3.
ST02-PIB.WF1	Apartado 1.3.
SC03-SALARIOS3.WF1	Apartado 2.1.
NUM02-ANSCOMBE.WF1	Apartado 2.3.

A.2 Indicaciones sobre Algunos Ejercicios

La resolución de los Ejercicios 6, 8[A] y 12, requiere utilizar un programa como EViews o similar. En las Secciones 1-10 de la guía *Introducción al Uso de EViews 4.1* disponible en ucm.randomshock.com se explica con todo detalle cómo llevar a cabo las operaciones correspondientes.

A.2.1 Solución del Ejercicio 7

Un modelo RLS como

$$Y_i = \beta_1' + \beta_2 X_i + U_i', \text{ con } \mathrm{E}[U_i'] = \theta \neq 0,$$

es equivalente a

$$Y_i = (\beta_1' + \theta) + \beta_2 X_i + (U_i' - \theta) = \beta_1 + \beta_2 X_i + U_i, \text{ con}$$
$$\beta_1 = (\beta_1' + \theta),\ U_i = (U_i' - \theta),\ \mathrm{E}[U_i] = \mathrm{E}[U_i'] - \mathrm{E}[\theta] = \theta - \theta = 0,$$

lo que implica que el término constante de un modelo de regresión lineal siempre se puede "redefinir" de tal manera que la esperanza de las perturbaciones sea igual a cero.

A.2.2 Solución del Ejercicio 8[B]

La hipótesis de que $\mathrm{E}[U_i \mid X_i] = 0$ en [9] implica, de acuerdo con [37]-[38], que

$$\mathrm{E}[U_i] = \mathrm{E}[\mathrm{E}[U_i \mid X_i]] = \mathrm{E}[0] = 0, \qquad [A1]$$

$$\mathrm{Cov}[U_i, X_i] = 0. \qquad [A2]$$

A su vez, [A2] implica, teniendo en cuenta [A1], que

$$\mathrm{E}[(U_i - \mathrm{E}[U_i])(X_i - \mathrm{E}[X_i]) = \mathrm{E}[U_i X_i] - \mathrm{E}[U_i] \times \mathrm{E}[X_i] = \mathrm{E}[U_i X_i] = 0.$$

Por lo tanto, $Y_i = \beta_1 + \beta_2 X_i + U_i$ con $\mathrm{E}[U_i \mid X_i] = 0$ en [9] implican en conjunto que

$$\mathrm{E}[Y_i - \beta_1 - \beta_2 X_i] = 0, \qquad [A3]$$

$$\mathrm{E}[(Y_i - \beta_1 - \beta_2 X_i) X_i] = 0, \qquad [A4]$$

para todo $i = 1, 2, ..., N$. Las **condiciones muestrales análogas** a [A3]-[A4] para estimar los parámetros β_1 y β_2 por el **método de momentos** son las siguientes:

$$\tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \qquad [A5]$$

$$\tfrac{1}{N} \textstyle\sum_{i=1}^{N} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = 0, \qquad [A6]$$

que constituyen un sistema de dos ecuaciones con dos incógnitas $(\hat{\beta}_1, \hat{\beta}_2)$, cuya única solución, bajo la premisa de que $\sum_{i=1}^{N} (x_i - \overline{x})^2 > 0$, es

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}, \qquad [A7]$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} = \frac{\hat{\mathrm{cov}}[\mathbf{y}, \mathbf{x}]}{\hat{\mathrm{var}}[\mathbf{x}]} = \frac{\hat{\mathrm{dvt}}[\mathbf{y}]}{\hat{\mathrm{dvt}}[\mathbf{x}]} \times \hat{\mathrm{corr}}[\mathbf{y}, \mathbf{x}]. \qquad [A8]$$

La obtención paso a paso de $\hat{\beta}_1$ y $\hat{\beta}_2$ en [A7]-[A8] a partir de [A5]-[A6] constituye un buen

ejercicio de repaso sobre operaciones con sumatorios y estadísticos muestrales. El significado de algunos de los símbolos que figuran en [A7]-[A8] es el siguiente:

Vectores de datos: $\mathbf{y} = [y_1\,, y_2\,, ..., y_N\,]'$, $\mathbf{x} = [x_1\,, x_2\,, ..., x_N\,]'$.

Medias muestrales: $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i\,,\ \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,.$

Varianza muestral: $\hat{\text{var}}[\mathbf{x}] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\,.$

Desviación típica muestral: $\hat{\text{dvt}}[\mathbf{x}] = \sqrt{\hat{\text{var}}[\mathbf{x}]}.$

Covarianza muestral: $\hat{\text{cov}}[\mathbf{y}, \mathbf{x}] = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x}).$

Correlación simple muestral: $\hat{\text{corr}}[\mathbf{y}, \mathbf{x}] = \dfrac{\hat{\text{cov}}[\mathbf{y}, \mathbf{x}]}{\hat{\text{dvt}}[\mathbf{y}] \times \hat{\text{dvt}}[\mathbf{x}]}\,.$

Para más detalles, ver, por ejemplo, Wooldridge (2020), pp. 25-26.

INTRODUCCIÓN A LA
ECONOMETRÍA
UN PRIMER CONTACTO

En un primer curso de Econometría, quizás lo realmente importante no tiene que ver del todo ni con el contenido específico de la materia ni con la forma de impartirla o de estudiarla, sino más bien con intentar explicar y entender claramente desde el principio del curso qué es la Econometría. La idea central a este respecto es muy sencilla: la Econometría es simplemente una herramienta de propósito general que puede contribuir a resolver muchos problemas prácticos, mediante la interpretación y el uso adecuado de la información contenida en una colección de datos resumida a través de uno o varios modelos.

En esta guía ilustrada se introducen, detenida y ordenadamente, los conceptos y los procedimientos fundamentales del Análisis de Regresión Lineal dirigidos a enfocar la Econometría como lo que realmente es en sus aspectos tanto teóricos como aplicados, incluyendo los siguientes:

» Preguntas, datos, modelos y respuestas.

» Evaluación de efectos causales y cálculo de previsiones.

» Especificación, estimación, diagnosis y revisión de modelos.

» Recursos matemáticos, estadísticos e informáticos.

A lo largo del texto se utilizan numerosos ejemplos y se proponen varios ejercicios con la intención de ilustrar tanto las posibilidades como las dificultades asociadas con la elaboración de cualquier análisis econométrico aplicado en la práctica.

Esta guía pretende contribuir, en particular, a que toda la parafernalia técnica asociada típicamente con un primer curso de Econometría (teórica y aplicada) resulte fácilmente explicable y entendible, dentro de un panorama amplio y progresivamente familiar en el que se pueden ir detallando, poco a poco y con sentido, todas las piezas que suelen conformar ese primer curso.

JOSÉ ALBERTO MAURICIO

Profesor Titular
Departamento de Análisis Económico y Economía Cuantitativa
Universidad Complutense de Madrid