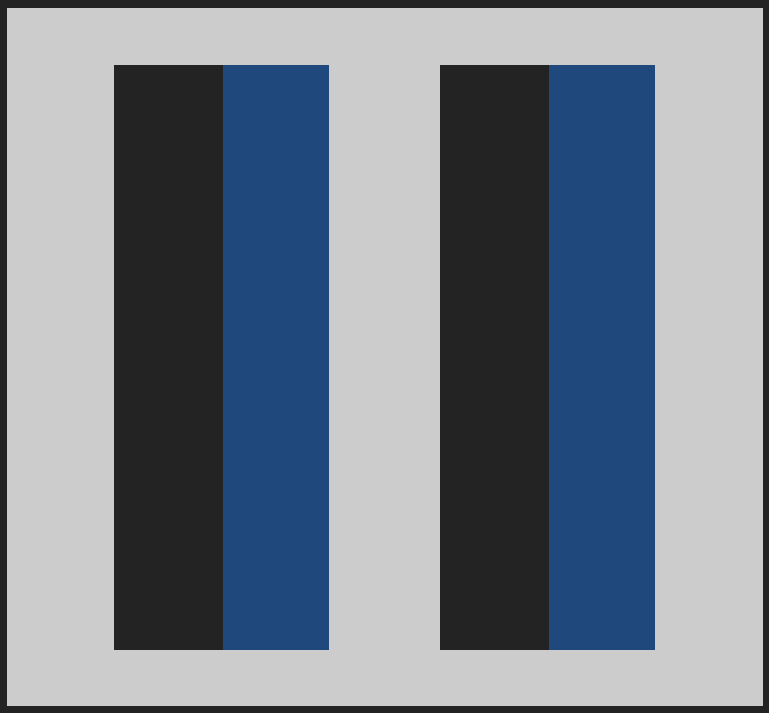
ECONOMETRÍA

JOSÉ ALBERTO MAURICIO



1

INTRODUCCIÓN

Departamento de Análisis Económico y Economía Cuantitativa
Universidad Complutense de Madrid

EctrGr-JAM-1.pdf

ucm.randomshock.com/ectrgr

Versión 2.0 - Enero 2023

OBSERVACIÓN IMPORTANTE

En varias partes de este Tema 1 se mencionan y se utilizan algunos conceptos y métodos que no se explican detalladamente hasta el Tema 2. En particular, para elaborar los gráficos y obtener los resultados numéricos que se incluyen en este Tema 1 se han seguido algunos procedimientos descritos en la guía *Introducción al Uso de EViews 4.1* que se consideran, a medida que van siendo relevantes, a lo largo del Tema 2.

BIBLIOGRAFÍA PARA EL TEMA 1

Hill, Griffiths, Lim (2018): Capítulo 1.
Wooldridge (2020): Capítulo 1.

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 (2002)
A Concise Introduction to Econometrics (Cambridge University Press)

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, H.K. VAN DIJK (2004)
Econometric Methods with Applications in Business and Economics (Oxford University Press)

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.

F.J. ANSCOMBE (1973)
Graphs in Statistical Analysis (The American Statistician 27-1, pp. 17-21)

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 (1995)
Identification Problems in the Social Sciences (Harvard University Press)

PARTE 1 - CONCEPTO Y METODOLOGÍA

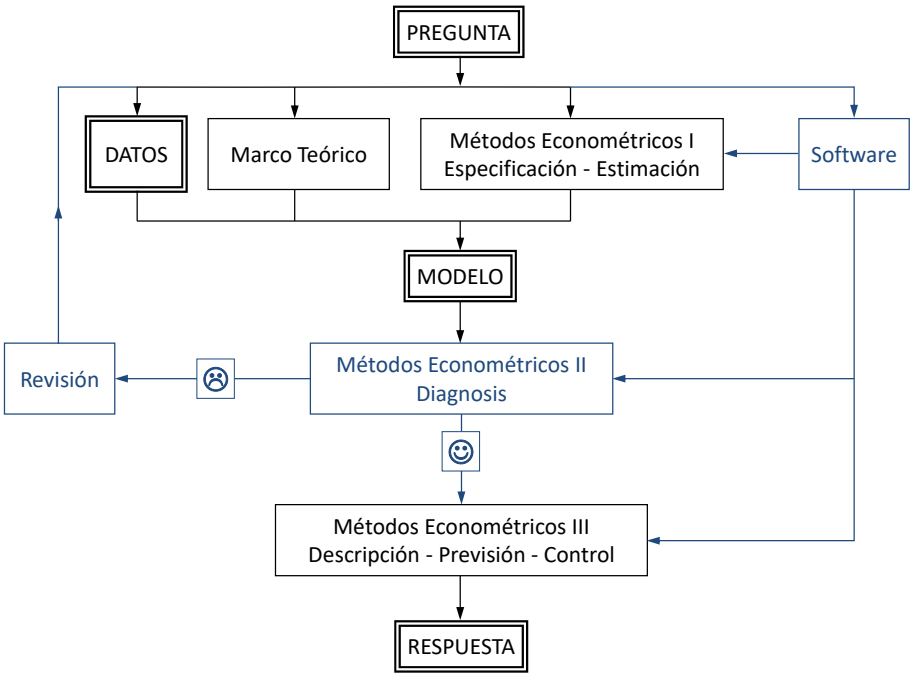
Wanderers in the land of Osten Ard are cautioned not to put blind trust in old rules and forms, and to observe all rituals with a careful eye, for they often mask being with seeming.

T. WILLIAMS (1988)
The Dragonbone Chair - Memory, Sorrow and Thorn: Book 1 (DAW Books)

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, en algún sentido, teorías y relaciones económicas. Ante todo, la econometría moderna es una herramienta de propósito bastante general, que puede contribuir a dar **respuestas** a ciertos tipos de **preguntas** prácticas en contextos muy variados, utilizando la información contenida en una colección de **datos** y resumida a través de uno o varios **modelos**.

Durante un primer curso de econometría es importante mantener presente en todo momento la idea central del párrafo anterior (Figura 1), con la doble finalidad de (1) situar adecuadamente cada elemento del programa de la asignatura dentro de un panorama amplio que resulte progresivamente familiar a medida que se avanza en el curso, y (2) entender la utilidad del instrumental técnico (**matemáticas**, **estadística**, **informática**)

FIGURA 1
La Metodología de la Econometría Aplicada



asociado con la **econometría teórica** y **aplicada** a medida que se va haciendo uso de él.

PARTE 2 - EJEMPLO

2.1 PREGUNTA: Evaluar la influencia de la asistencia a clase sobre las notas finales en cierta asignatura impartida en un curso académico determinado.

2.2 DATOS: Notas finales $(nf_1, nf_2, ..., nf_N)$ y horas de asistencia a clase $(ha_1, ha_2, ..., ha_N)$ de un conjunto de N estudiantes matriculados en la asignatura y el curso considerados:

$$\begin{bmatrix} nf_1 & ha_1 \\ nf_2 & ha_2 \\ \vdots & \vdots \\ nf_N & ha_N \end{bmatrix}. \qquad [1]$$

Los datos (Figura 2) pueden interpretarse como una **muestra** de N observaciones (puntos muestrales, o filas),

$$[\, nf_i, ha_i \,], \quad i = 1, 2, ..., N, \qquad [2]$$

referida a N estudiantes de una **población** determinada.

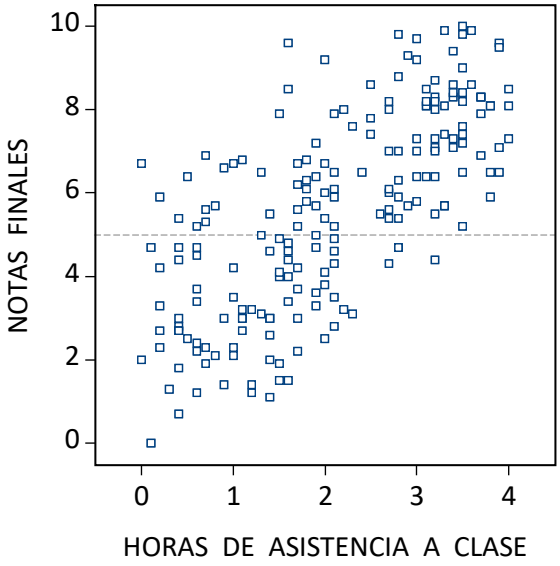
2.3 MODELO: Un modelo de **regresión lineal simple** (RLS) que quizás resulte útil para dar algún tipo de respuesta a la pregunta planteada en 2.1, se puede **especificar** de manera inicial (provisional, o tentativa) como

$$NF = \beta_1 + \beta_2 HA + U, \qquad [3]$$

donde:

FIGURA 2
Datos sobre Notas Finales y Horas de Asistencia a Clase

Notas finales (de 0 a 10) y horas de asistencia a clase (número medio por semana) de 200 estudiantes de cierta asignatura en la convocatoria ordinaria del curso 2021-22.
SC00-NOTAS.WF1



- NF (la **variable dependiente** del modelo) y HA (la **variable explicativa**) representan la nota final y la asistencia a clase de cualquier estudiante en la muestra considerada.
- β_1 (el **término constante** del modelo) y β_2 (la **pendiente** del modelo) son **parámetros** (números cuyos valores no se conocen).
- El término $\beta_1 + \beta_2 HA$ representa una parte de NF (la parte considerada, recogida o incluida, de forma explícita en el modelo); el símbolo U representa la parte restante (el **término de error**, o la **perturbación**, del modelo).
- En particular, U recoge todo aquello que realmente determina NF más allá de lo recogido en $\beta_1 + \beta_2 HA$, como (entre otras cosas) las horas de estudio fuera de clase (observables) y la calidad académica referida a la preparación, la disciplina de estudio, el interés y la responsabilidad (difíciles o imposibles de observar).

De acuerdo con [3],

$$\beta_1 = \beta_1 + \beta_2 HA \text{ cuando } HA = 0, \qquad [4]$$

$$\beta_2 = \partial NF \,/\, \partial HA \ (= \Delta NF \text{ cuando } \Delta HA = 1 \text{ y } \Delta U = 0), \qquad [5]$$

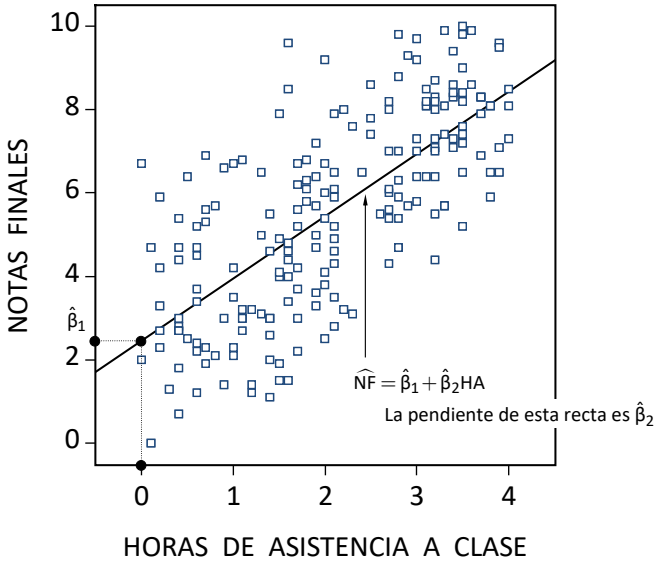
por lo que, en particular, β_2 es el **efecto** "**ceteris paribus**" (**causal**, **directo**, o **parcial**) de la asistencia a clase sobre la nota final.

Observación 1: Si NF_0 y NF_1 representan dos valores cualesquiera para NF, entonces [3] $\Rightarrow$ $\Delta NF = NF_1 - NF_0 = (\beta_1 + \beta_2 HA_1 + U_1) - (\beta_1 + \beta_2 HA_0 + U_0) = \beta_2 (HA_1 - HA_0) + (U_1 - U_0) = \beta_2 \Delta HA + \Delta U$, por lo que, como se indica en [5], $\beta_2 = \Delta NF$ cuando $\Delta HA = 1$ ($HA_1 = HA_0 + 1$) y $\Delta U = 0$ ($U_1 = U_0$).

Si se pudiese obtener alguna información **fiable** sobre lo que realmente vale β_2, entonces sería posible dar algún tipo de respuesta a la pregunta planteada en 2.1.

FIGURA 3
Modelo RLS Estimado con los Datos de la Figura 2

Estimaciones de β_1 y β_2 calculadas a través del método MCO (mínimos cuadrados ordinarios): $\hat{\beta}_1 = 2.47$ (término constante) y $\hat{\beta}_2 = 1.49$ (pendiente). Ver [A.1]-[A.3].



Para empezar, esa información debería incluir alguna asignación numérica concreta para el valor desconocido de β_2, es decir, alguna **estimación puntual** de β_2 (también de β_1, aunque su papel y su relevancia son otros). Si $\hat{\beta}_1, \hat{\beta}_2$ son estimaciones concretas de β_1, β_2, calculadas a través de algún **método** o **criterio** de **estimación**, entonces la expresión

$$\widehat{NF} = \hat{\beta}_1 + \hat{\beta}_2 HA \qquad [6]$$

representa la parte numérica concreta de NF considerada (recogida, incluida) de forma explícita en el modelo (Figura 3). Una expresión como [6] se denomina un **modelo** RLS **estimado** (correspondiente a un **modelo** RLS **especificado** como [3]), que no es otra cosa que un mero **resumen** de la información que contienen los datos.

Observación 2: Un método (criterio) de estimación es un procedimiento diseñado para asignar unos valores numéricos concretos a los parámetros de un modelo, sobre la base de (1) la información contenida en los datos y (2) la forma del modelo considerado para organizar (ordenar, resumir) dicha información.

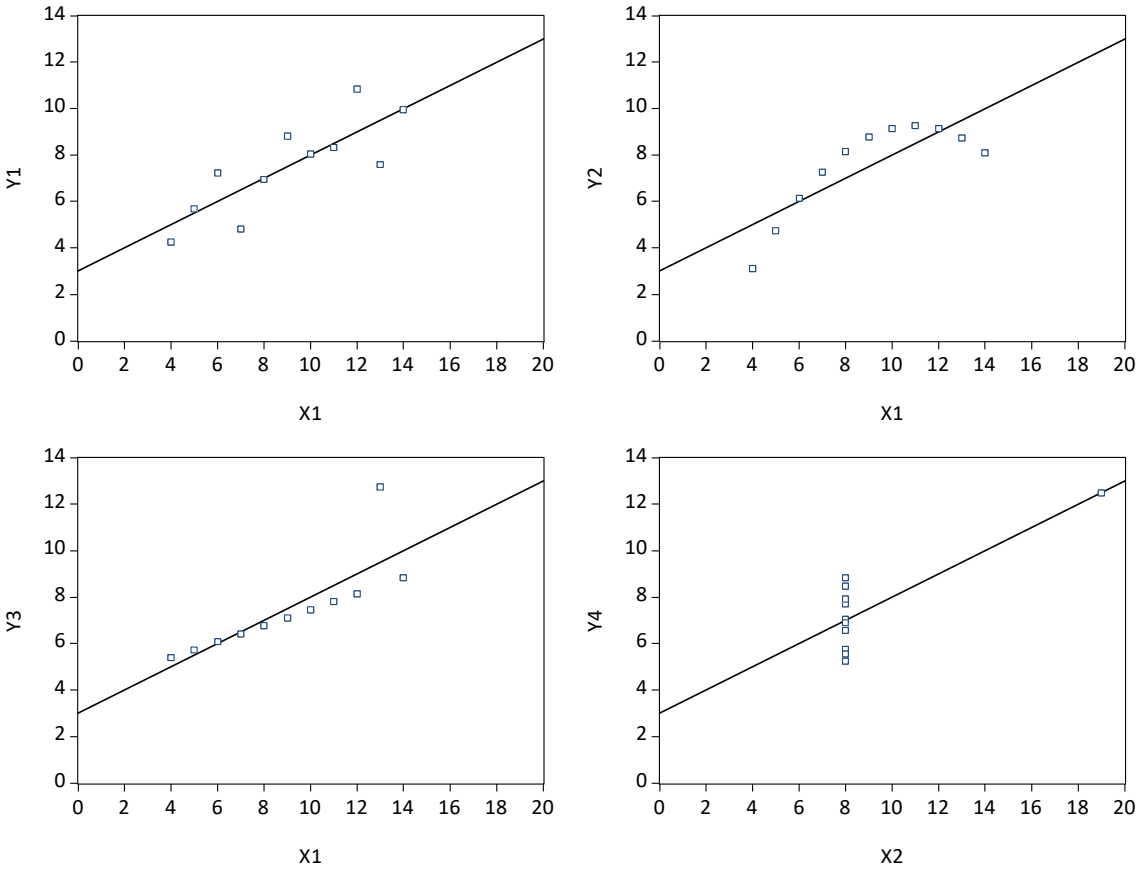
De acuerdo con [5], $\hat{\beta}_2$ es una estimación del efecto causal de HA sobre NF, es decir, una asignación numérica concreta, basada en algún criterio, para el valor (desconocido) de la influencia directa, parcial, o "ceteris paribus" de HA sobre NF. Una estimación puntual $\hat{\beta}_2$ de β_2 proporciona alguna información sobre lo que realmente puede valer β_2, pero aún quedan cuestiones que considerar al respecto, la más importante de las cuales tiene que ver con la **diagnosis** del modelo RLS especificado en [3] y estimado en [6].

Un modelo RLS estimado no siempre resume adecuadamente la información contenida en una colección de datos sobre dos variables (Figura 4). En esos casos, las estimaciones calculadas para β_1, β_2 pueden no ser fiables y llevar a conclusiones (respuestas) erróneas, por lo que, antes de utilizarlo con cualquier finalidad, un modelo estimado debe ser diagnosticado cuidadosamente y, en su caso, revisado en la dirección adecuada.

FIGURA 4
Modelos RLS Estimados por MCO con Diferentes Colecciones de Datos

En los cuatro casos $\hat{\beta}_1 = 3.0$ (término constante) y $\hat{\beta}_2 = 0.5$ (pendiente).

NUM02-ANSCOMBE.WF1



2.4 RESPUESTAS: Más allá de la información que ofrece una mera estimación puntual de un parámetro (más bien poca, y a veces casi ninguna), los resultados de diferentes **contrastes de hipótesis** y del cálculo de **intervalos de confianza** proporcionan, en general, la información más completa que se puede obtener sobre lo que realmente vale el parámetro (o los parámetros) de interés:

➪ Un contraste de hipótesis es un procedimiento diseñado para decidir si rechazar o no, según el caso, alguna conjetura sobre lo que realmente valen los parámetros de un modelo (la **hipótesis nula**, H_0) en favor de otra conjetura diferente (la **hipótesis alternativa**, H_1), utilizando para ello la información contenida en los datos y resumida en el modelo estimado correspondiente. Por ejemplo, un contraste importante en el modelo [3] sería el de

$$H_0\colon \beta_2 = 0 \text{ frente a } H_1\colon \beta_2 \neq 0, \text{ o frente a } H_1\colon \beta_2 > 0. \qquad [7]$$

➪ Un intervalo de confianza para un parámetro de un modelo es un conjunto de **valores posibles** para dicho parámetro que son **relativamente compatibles** con la información contenida en los datos y resumida en el modelo estimado correspondiente.

Adicionalmente a la descripción cuantitativa del posible efecto causal de *HA* sobre *NF*, un modelo estimado como [6] se puede utilizar también para lo siguiente:

- **Previsión:** Para calcular una estimación $\hat{nf}_*$ de la nota final de un estudiante que asista a clase durante este curso un número de horas ha_* dado, el modelo estimado [6] sugiere que

$$\hat{nf}_* = \hat{\beta}_1 + \hat{\beta}_2 ha_* . \qquad [8]$$

  Si fuera posible, también tendría cierto interés estimar la probabilidad de que el estudiante considerado aprobase la asignatura (es decir, obtuviese una nota final mayor o igual que cinco) asistiendo a clase ese número de horas ha_*.

- **Control:** Para calcular una estimación $\hat{ha}_*$ del número de horas de asistencia a clase que implicaría la obtención de una nota final nf_* dada, el modelo estimado [6] sugiere en este caso (de manera recíproca al caso de la previsión) que

$$nf_* = \hat{\beta}_1 + \hat{\beta}_2 \hat{ha}_* \Rightarrow \hat{ha}_* = (nf_* - \hat{\beta}_1) / \hat{\beta}_2 . \qquad [9]$$

Como en el caso de las estimaciones puntuales $\hat{\beta}_1$ y $\hat{\beta}_2$ consideradas en 2.3, la fiabilidad de las respuestas obtenidas en este Apartado 2.4 dependerá del grado en que el modelo estimado [6] resuma adecuadamente la información contenida en los datos.

PARTE 3 - ALGUNOS TIPOS DE DATOS

Los datos de la Figura 2 son un ejemplo de lo que se conoce como datos de **sección cruzada**, o transversales. Otros tipos de datos son los datos de **series temporales** y los datos de **panel**, o longitudinales. Cada uno de estos tipos de datos resulta adecuado para intentar 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.

DATOS DE SECCIÓN CRUZADA

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

La Tabla 1 de la página siguiente 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 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 se puede esperar que las conclusiones obtenidas del análisis de dicha sección cruzada (muestra) sean aplicables a todo el colectivo (población).

TABLA 1
Datos sobre Algunas Características de 80 Viviendas Unifamiliares Vendidas en el Área Metropolitana de Boston en 1990
SC01-VIVIENDAS.WF1

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
⋮	⋮	⋮	⋮
79	281.0	179.1	4
80	260.0	120.2	3

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.

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.

TABLA 2
Datos sobre Ventas y Gasto en Publicidad Anuales de la Empresa Lydia E. Pinkham's Medicine Co. desde 1907 hasta 1960
ST01-PINKHAM.WF1

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

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.

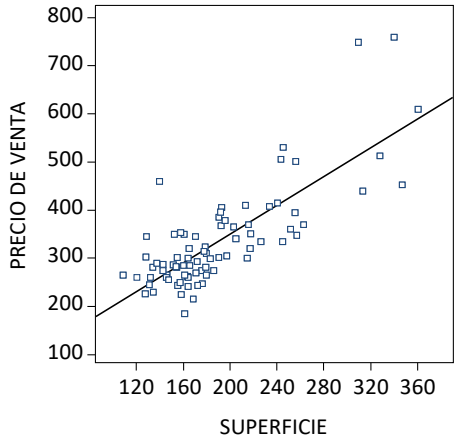
En ocasiones, puede resultar útil analizar información que combine datos de sección cruzada con datos de series temporales. En particular, el análisis de una combinación de distintas secciones cruzadas referidas a la misma población en diferentes momentos de su historia, puede resultar útil para evaluar 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 uno de los momentos considerados, la colección de datos correspondiente se denomina una colección de **datos de panel**.

FIGURA 5
Modelos RLS Estimados por MCO con Datos de Sección Cruzada

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

SC01-VIVIENDAS.WF1

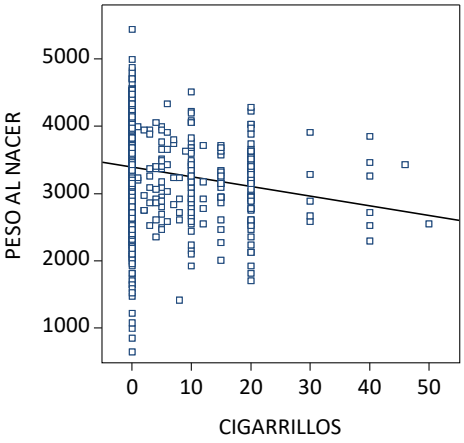


$$\widehat{\text{PRECIO}} = 51.551 + 1.494 \times \text{SUPM2}$$

$N = 80$ viviendas

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



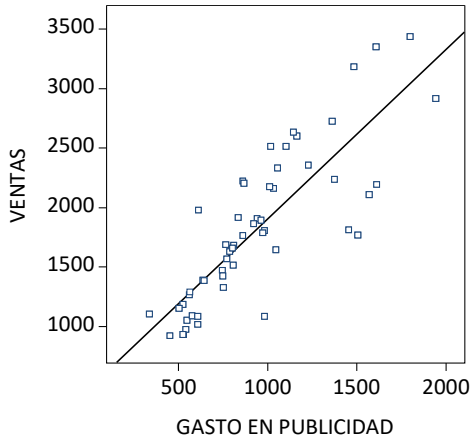
$$\widehat{\text{PESO}} = 3392 - 14.4 \times \text{CIGM}$$

$N = 1387$ recién nacidos

FIGURA 6
Modelos RLS Estimados por MCO con Datos de Series Temporales

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

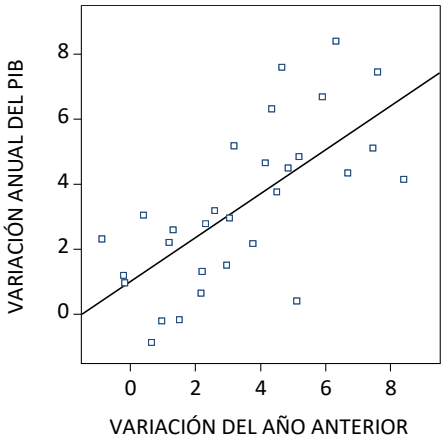


$$\widehat{\text{VENTAS}} = 477.500 + 1.429 \times \text{GPUB}$$

$N = 54$ años (1907-1960)

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

ST02-PIB.WF1



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

$N = 32$ años (1966-1997)

PARTE 4 - LIMITACIÓN FUNDAMENTAL DE LA REGRESIÓN LINEAL SIMPLE

De la información contenida en la Figura 3 se puede concluir razonablemente que, en la muestra de 200 estudiantes utilizada, existe una relación positiva entre las notas finales y las horas de asistencia a clase: los estudiantes que asisten más a clase tienen, en general, notas finales más altas. Además, la estimación $\hat{\beta}_2 = 1.49$ del parámetro β_2 (el efecto causal de la asistencia a clase sobre las notas finales) es moderadamente grande desde un punto de vista práctico; por ejemplo, un estudiante que asiste a clase dos horas por semana más que otro tiene, por término medio, una nota final del orden de tres puntos más alta.

Observación 3: Por otro lado, el intervalo de confianza del 99% para β_2 es [1.22, 1.76], por lo que, de acuerdo con el modelo RLS estimado, β_2 es estadísticamente significativo incluso al 1% (Tema 2: Sección 2.3).

A pesar de estos resultados aparentemente interesantes, la clave de todo el análisis está en cómo decidir si la relación encontrada es una **relación causal** legítima entre las notas finales y las horas de asistencia a clase, con un efecto causal asociado relativamente próximo a la estimación $\hat{\beta}_2$. Los estudiantes que más asisten a clase tienen, por término medio, mejores notas finales, pero eso no significa necesariamente que el mero hecho de asistir más a clase implique por sí sólo ("ceteris paribus") obtener una nota final mejor. Puede ocurrir, en particular, que la relación encontrada entre las notas finales y la asistencia a clase se deba tan sólo a que las horas de asistencia a clase estén relacionadas con otros factores, y que, de hecho, sean esos otros factores los que realmente determinen las notas finales. Uno de ellos puede ser el número de horas de estudio fuera de clase, que (al igual que otros muchos determinantes de las notas finales) no se ha incluido explícitamente en el análisis. En consecuencia, sería importante considerar las dos posibilidades siguientes:

[1] Quizás los estudiantes que más asisten a clase son también, en general, los que más tiempo le dedican a la asignatura fuera de clase.

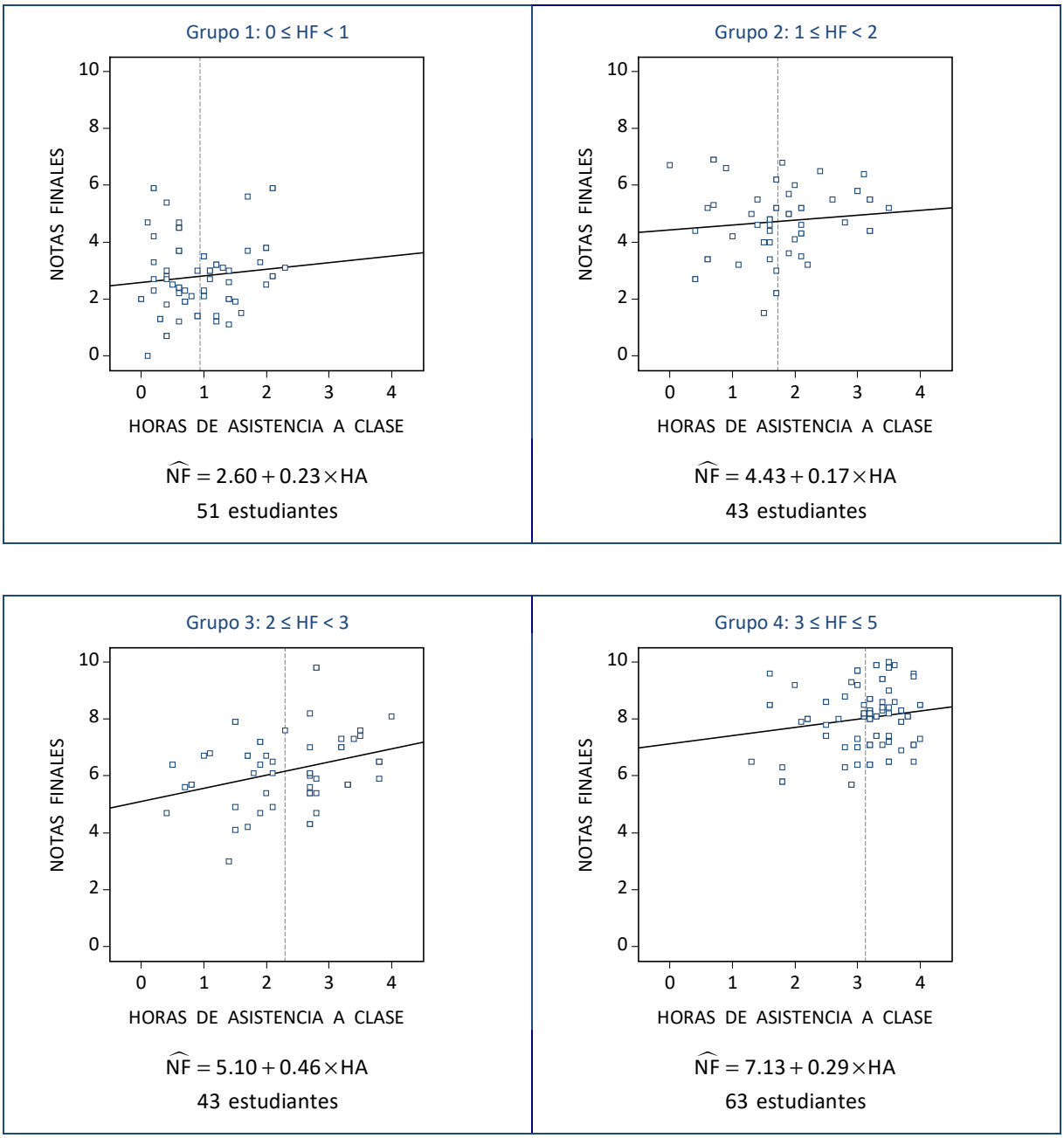
[2] Seguramente los estudiantes que le dedican más tiempo a la asignatura fuera de clase tienen, en general, notas más altas que los que le dedican menos tiempo.

Si [1]-[2] ocurren de hecho a la vez, entonces la relación encontrada entre las notas finales y la asistencia a clase quizás se debe, sólo o en parte, a dicha ocurrencia y no a que exista un efecto causal (directo) auténtico o legítimo de la asistencia a clase sobre las notas finales. En tal caso, β_2 puede ser un número bastante más pequeño de lo que sugiere la estimación $\hat{\beta}_2$ en [6] (Figura 3), incluyendo la posibilidad de que β_2 sea igual a cero.

Una forma de evaluar estas posibilidades consiste en dividir la muestra de 200 estudiantes en varios grupos de manera que, dentro de cada grupo, las horas de estudio fuera de clase sean aproximadamente las mismas, y estimar un modelo RLS para cada grupo. Esta estrategia permite estimar β_2 en cada grupo con la seguridad de que las horas de estudio fuera de clase son aproximadamente **constantes**, y por lo tanto **independientes** (porque no varían) de las horas de asistencia a clase, **dentro de cada grupo**. Si las estimaciones de β_2 en los diferentes grupos son semejantes entre ellas y parecidas a la estimación $\hat{\beta}_2$ en la muestra completa, entonces $\hat{\beta}_2$ puede ser una estimación fiable del efecto causal buscado; pero si las estimaciones en los grupos son apreciablemente distintas de $\hat{\beta}_2$ (comparar la Figura 7 con la Figura 3), entonces $\hat{\beta}_2$ no es fiable porque su valor viene dado en alguna medida por la ocurrencia conjunta de las posibilidades [1]-[2] mencionadas anteriormente.

FIGURA 7
Modelos RLS Estimados por MCO con los Datos de la Figura 2 por Grupos

HF = Horas de estudio fuera de clase (número medio por semana durante el curso)



La estimación $\hat{\beta}_2$ en la muestra completa ($\hat{\beta}_2 = 1.49$) es del orden de entre tres y casi nueve veces mayor que las estimaciones obtenidas por grupos, lo que sugiere que $\hat{\beta}_2$ es una estimación muy poco fiable (porque está muy **sesgada** al alza) del efecto causal de la asistencia a clase sobre las notas finales en la muestra completa. Dicho de otro modo, el modelo RLS estimado de la Figura 3 no resume adecuadamente la información contenida en los datos, por lo que prácticamente cualquier conclusión (respuesta) basada en ese modelo será poco fiable. En particular, el modelo RLS estimado de la Figura 3 sugiere la existencia de una relación causal significativa entre las notas finales y la asistencia a clase cuando, en realidad, esa relación es muy limitada o quizás ni siquiera existe (es decir, la relación que sugiere el modelo RLS es una **relación espuria**, carente de autenticidad o de legitimidad).

PARTE 5 - REGRESIÓN LINEAL: DE SIMPLE A MÚLTIPLE

La dificultad fundamental en un análisis de causalidad basado en un único modelo RLS, reside en que no es posible considerar explícitamente (porque no tienen cabida en el modelo empleado) otras influencias sobre la variable dependiente que pueden estar relacionadas con la única influencia (la única variable explicativa) que sí se considera de manera explícita. Si fuera posible tener explícitamente en cuenta esas otras influencias, dentro de un modelo en el que sí tuviesen cabida, entonces la dificultad mencionada ya no estaría presente. En este sentido, para evaluar el efecto causal de la asistencia a clase sobre las notas finales de manera más fiable que con un modelo RLS como [3], se puede plantear un modelo alternativo como

$$NF = \beta_1 + \beta_2 HA + \beta_3 HF + U, \qquad [10]$$

donde NF, HA y HF son la nota final, las horas de asistencia a clase y las horas de estudio fuera de clase, respectivamente, de cualquier estudiante en la población considerada.

Observación 4: Las estimaciones MCO de $\beta_1, \beta_2, \beta_3$ en [10], calculadas con los datos del archivo SCOO-NOTAS.WF1, son $\hat{\beta}_1 = 2.27$, $\hat{\beta}_2 = 0.14$, $\hat{\beta}_3 = 1.41$; ver [A.4]-[A.6]. Además, β_2 no es estadísticamente significativo ni siquiera al 20%, mientras que tanto β_1 como β_3 sí lo son incluso al 1% (Tema 2: Sección 2.3). Estos resultados están bastante de acuerdo con las conclusiones obtenidas en la PARTE 4.

La expresión [10] es un ejemplo de un modelo de **regresión lineal múltiple** (RLM) del tipo

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + U, \qquad [11]$$

donde $X_2, X_3, ..., X_K$ son $K-1$ variables explicativas observables que pueden estar relacionadas entre sí, y U recoge todas las influencias (observables y no observables) sobre Y que no están incluidas en $\beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K$. Un modelo como [11] permite incluir explícitamente en un análisis de causalidad tantas influencias observables como se considere oportuno, a pesar de que en última instancia quizás interese evaluar el efecto directo o causal de tan sólo una de ellas. En particular, de acuerdo con [11],

$$\beta_2 = \frac{\partial Y}{\partial X_2} \Rightarrow \beta_2 = \Delta Y \text{ cuando } \Delta X_2 = 1 \text{ y } \Delta X_3 = ... = \Delta X_K = \Delta U = 0, \qquad [12]$$

por lo que, en general, β_j en [11] es el efecto causal (directo, parcial, o "ceteris paribus") de X_j $(2 \le j \le K)$ sobre Y.

En todo caso, en un análisis de causalidad basado en un modelo RLM como [11] también cabe la posibilidad de que no se consideren explícitamente otros factores (especialmente, factores no observables) relacionados con alguna/s variable/s explicativa/s del modelo. Aunque esta posibilidad es seguramente menos probable que en el caso de un análisis basado en un modelo RLS (porque en un modelo RLS el número de factores omitidos es mayor que en un modelo RLM), lo cierto es que, en cualquier caso, la evaluación fiable en la práctica de efectos causales mediante modelos de regresión descansa, en buena medida, en la confianza que se tenga en que las influencias que no se consideran explícitamente (por error, o por imposibilidad) sean razonablemente independientes de las influencias que sí se consideran de manera explícita. Por este motivo, dos **elementos centrales** en cualquier análisis de regresión son los que tienen que ver con qué **variables explicativas** se **incluyen**

explícitamente en el análisis y cuál es su posible **relación** con otras variables que se **omiten**. Aunque en general no es posible "demostrar" que las variables omitidas (especialmente cuando no son observables) son independientes de las incluidas, como mínimo se debe razonar cuidadosamente al respecto para intentar concluir si esa independencia es asumible (razonable, factible) o no lo es.

Observación 5: Las consideraciones hechas acerca de la evaluación del efecto causal de la asistencia a clase sobre las notas finales son también aplicables a muchos otros casos, incluyendo, por ejemplo, la evaluación del efecto causal de la educación sobre los salarios, la del uso de fertilizantes sobre el rendimiento de zonas de cultivo, la del gasto en publicidad sobre las ventas empresariales, o la de la inflación sobre el crecimiento del producto interior bruto. En particular, algunas dificultades importantes asociadas con esos cuatro casos (o con cualesquiera otros del mismo estilo) son completamente análogas a las mencionadas en el ejemplo de la asistencia a clase y las notas finales.

APÉNDICE

A continuación se presentan algunas fórmulas para las estimaciones por mínimos cuadrados ordinarios (MCO) de los parámetros en un modelo de regresión lineal simple (RLS) como [3] y en un modelo de regresión lineal múltiple (RLM) como [10]. El origen y la obtención de estas fórmulas se consideran en la Sección 2.2 del Tema 2.

Modelo RLS:
$$Y = \beta_1 + \beta_2 X_2 + U. \qquad [A.1]$$

Datos:
$$\begin{bmatrix} y_1 & x_{12} \\ y_2 & x_{22} \\ \vdots & \vdots \\ y_N & x_{N2} \end{bmatrix}. \qquad [A.2]$$

Estimaciones MCO:
$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^N x_{i2} \\ \sum_{i=1}^N x_{i2} & \sum_{i=1}^N x_{i2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_{i2} y_i \end{bmatrix}. \qquad [A.3]$$

Modelo RLM:
$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U. \qquad [A.4]$$

Datos:
$$\begin{bmatrix} y_1 & x_{12} & x_{13} \\ y_2 & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ y_N & x_{N2} & x_{N3} \end{bmatrix}. \qquad [A.5]$$

Estimaciones MCO:
$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^N x_{i2} & \sum_{i=1}^N x_{i3} \\ \sum_{i=1}^N x_{i2} & \sum_{i=1}^N x_{i2}^2 & \sum_{i=1}^N x_{i2}x_{i3} \\ \sum_{i=1}^N x_{i3} & \sum_{i=1}^N x_{i3}x_{i2} & \sum_{i=1}^N x_{i3}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_{i2} y_i \\ \sum_{i=1}^N x_{i3} y_i \end{bmatrix}. \qquad [A.6]$$