Econometría aplicada con EViews para la investigación

Guía práctica paso a paso para tesis, artículos científicos y análisis de datos

Willian Sebastian Flores Sotelo

Lima, Perú; 2026

Econometría aplicada con EViews para la investigación
Guía práctica paso a paso para tesis, artículos científicos y análisis de datos
Serie Econometría sin miedo, n.º 2
Autor-editor: © Willian Sebastian Flores Sotelo. Todos los derechos reservados.
Jr. Piura 282, piso 2, Urb. La Libertad, Comas – Lima – Perú
Primera edición digital: octubre de 2026.
Hecho el Depósito Legal en la Biblioteca Nacional del Perú N.° 2026-12956
ISBN: 978-612-03-3326-6
Libro digital disponible en: https://willianfloressotelo.com
Ninguna parte de esta obra puede reproducirse sin autorización escrita del autor, salvo citas breves con fines de reseña o docencia.
Los programas .prg del libro se ofrecen para uso educativo. Las bases de datos citadas pertenecen a sus instituciones productoras.
Edición digital. Distribuido por Amazon Kindle Direct Publishing, Apple Books y Google Play Books.
EViews es una marca registrada de S&P Global Inc. o de sus filiales. Este libro es una obra independiente: no está patrocinado, autorizado ni avalado por S&P Global. Stata, Python, R y Excel son marcas de sus respectivos titulares.

Contenido breve

ACERCA DEL AUTOR iv
ÍNDICE DE TABLAS xi
PREFACIO: CÓMO USAR ESTE LIBRO xii
Parte Uno. Antes de estimar 1
Capítulo 1 ¿Por qué tu regresión miente? Correlación y causalidad 2
Capítulo 2 Tus datos listos en EViews en una tarde: de la ENAHO y el Banco Mundial a una base limpia 26
Capítulo 3 Prueba de hipótesis e intervalos de confianza: la estadística que tu investigación necesita 60
Parte Dos. La regresión en EViews, paso a paso 87
Capítulo 4 Regresión lineal simple paso a paso 88
Capítulo 5 Regresión lineal múltiple y el sesgo de variable omitida 114
Capítulo 6 Cómo interpretar una regresión sin pánico 134
Capítulo 7 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación 156
Capítulo 8 Regresión logística: modelos logit y probit 189
Parte Tres. Paneles y series de tiempo 216
Capítulo 9 Datos de panel: efectos fijos vs. aleatorios y test de Hausman 217
Capítulo 10 Series de tiempo y modelo ARIMA paso a paso 246
Capítulo 11 Pronósticos con modelos VAR y cointegración 275
Parte Cuatro. Efectos, no solo asociaciones 307
Capítulo 12 Evaluación de impacto I: diferencias en diferencias 308
Capítulo 13 Evaluación de impacto II: variables instrumentales 335
Capítulo 14 Evaluación de impacto III: regresión discontinua y propensity score matching 363
Parte Cinco. Reportar y dar un paso más 391
Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico 392
Capítulo 16 Modelos ARDL y bounds test 415
APÉNDICE A. EQUIVALENCIAS EVIEWS · STATA · PYTHON · R 441
APÉNDICE B. BASES DE DATOS DEL LIBRO 447

Contenido

ACERCA DEL AUTOR iv

ÍNDICE DE TABLAS xi

PREFACIO: CÓMO USAR ESTE LIBRO xii

Parte Uno. Antes de estimar 1

Capítulo 1 ¿Por qué tu regresión miente? Correlación y causalidad 2

Lo que vas a lograr 2

1.1 La correlación no te dice por qué 3

1.2 La pregunta que de verdad importa: ¿qué habría pasado si…? 3

1.3 Dibuja tu problema antes de estimarlo 4

1.4 La trampa, en una sola línea 4

1.5 El truco del experimento 5

1.6 ¿Y si no puedes hacer un experimento? 5

1.7 Manos a la obra 5

1.8 El error que comete casi todo tesista 18

1.9 El error que cometen los investigadores en la redacción de artículos 18

RESUMEN Y CONCLUSIONES 19

EJERCICIOS RESUELTOS 19

Para tu tesis o artículo 24

Referencias 25

Capítulo 2 Tus datos listos en EViews en una tarde: de la ENAHO y el Banco Mundial a una base limpia 26

Lo que vas a lograr 26

2.1 Dos tipos de datos, dos preguntas distintas 27

2.2 La ENAHO: módulos, llaves y un factor que convierte la muestra en un país 27

2.3 El Banco Mundial: países, agregados y dos formatos 28

2.4 Manos a la obra 28

2.5 El error que arruina más bases de lo que crees 52

2.6 El error que cometen los investigadores en la redacción de artículos 53

RESUMEN Y CONCLUSIONES 53

EJERCICIOS RESUELTOS 53

Para tu tesis o artículo 58

Referencias 59

Capítulo 3 Prueba de hipótesis e intervalos de confianza: la estadística que tu investigación necesita 60

Lo que vas a lograr 60

3.1 La pregunta: ¿el campo paga menos? 61

3.2 Tu muestra es una de cien posibles 61

3.3 El intervalo de confianza: una red, no una diana 61

3.4 La prueba de hipótesis: el detector de humo 62

3.5 Tres cosas que el valor p calla 62

3.6 Dos maneras de fallar: errores tipo I y II 63

3.7 Un resultado significativo puede no importar 63

3.8 Manos a la obra 63

3.9 El error que cometen casi todos los tesistas con el valor p 80

3.10 El error que cometen los investigadores en la redacción de artículos 81

RESUMEN Y CONCLUSIONES 81

EJERCICIOS RESUELTOS 82

Para tu tesis o artículo 85

Referencias 86

Parte Dos. La regresión en EViews, paso a paso 87

Capítulo 4 Regresión lineal simple paso a paso 88

Lo que vas a lograr 88

4.1 De la dummy a la recta 89

4.2 ¿Qué recta elige EViews? 89

4.3 ¿Por qué el logaritmo del ingreso? 89

4.4 Del logaritmo al porcentaje 90
4.5 El R^2: qué parte de la nube captura la recta . 90
4.6 Lo que la recta ve y lo que deja fuera 90
4.7 Manos a la obra 91
4.8 El error que convierte una pendiente en una promesa 107
4.9 El error que cometen los investigadores en la redacción de artículos 108
RESUMEN Y CONCLUSIONES 108
EJERCICIOS RESUELTOS 109
Para tu tesis o artículo 112
Referencias 113
Capítulo 5 Regresión lineal múltiple y el sesgo de variable omitida 114
Lo que vas a lograr 114
5.1 ¿Qué cambia cuando pones dos variables a la vez? 115
5.2 La variable que no está y sí cuenta 115
5.3 Dicotómicas: siempre contra alguien 116
5.4 ¿Sirve el R^2 para elegir el modelo? 116
5.5 Manos a la obra 116
5.6 El error que comete casi todo tesista: «mi coeficiente resistió los controles» 127
5.7 El error que cometen los investigadores en la redacción de artículos 128
RESUMEN Y CONCLUSIONES 128
EJERCICIOS RESUELTOS 129
Para tu tesis o artículo 132
Referencias 133
Capítulo 6 Cómo interpretar una regresión sin pánico 134
Lo que vas a lograr 134
6.1 Una ventana, tres bloques 134
6.2 ¿Qué dice un coeficiente? Depende de las unidades 135
6.3 El valor p responde una pregunta pequeña 135
6.4 Varios coeficientes a la vez: la prueba de Wald 136
6.5 Comparar sin hacer trampa: beta, Akaike y Schwarz 136
6.6 Manos a la obra 137
6.7 El error que comete casi todo tesista al leer su salida: el semáforo de estrellas 150
6.8 El error que cometen los investigadores en la redacción de artículos 151
RESUMEN Y CONCLUSIONES 151
EJERCICIOS RESUELTOS 152
Para tu tesis o artículo 154
Referencias 155
Capítulo 7 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación 156
Lo que vas a lograr 156
7.1 Unos supuestos cuidan el coeficiente; otros, su error estándar 157
7.2 Heterocedasticidad: el error no tiembla igual para todos 157
7.3 Errores agrupados: cuando los vecinos se parecen 158
7.4 Multicolinealidad: dos variables que cuentan la misma historia 158
7.5 Autocorrelación: el error de este año recuerda al del anterior 158
7.6 Normalidad y especificación: el supuesto que más asusta y el que más importa 159
7.7 Manos a la obra 159
7.8 El error que comete casi todo tesista con los supuestos: el anexo de cuatro pruebas 181
7.9 El error que cometen los investigadores en la redacción de artículos 182
RESUMEN Y CONCLUSIONES 182
EJERCICIOS RESUELTOS 183
Para tu tesis o artículo 186
Referencias 187

Capítulo 8 Regresión logística: modelos logit y probit 189
Lo que vas a lograr 189
8.1 ¿Qué pasa si tu variable dependiente solo vale 0 o 1? 190
8.2 Doblar la recta: logit y probit 190
8.3 Tres maneras de leer un logit 190
8.4 ¿Qué tan bueno es tu modelo? 191
8.5 Manos a la obra 191
8.6 El error que comete casi todo tesista con el logit 209
8.7 El error que cometen los investigadores en la redacción de artículos 209
RESUMEN Y CONCLUSIONES 210
EJERCICIOS RESUELTOS 210
Para tu tesis o artículo 214
Referencias 215
Parte Tres. Paneles y series de tiempo 216
Capítulo 9 Datos de panel: efectos fijos vs. aleatorios y test de Hausman 217
Lo que vas a lograr 217
9.1 ¿Qué tiene un panel que no tenga una encuesta? 218
9.2 Efectos fijos: cada país contra su propio promedio 218
9.3 Los choques comunes a toda la región 218
9.4 Efectos aleatorios: restar solo una parte 219
9.5 El test de Hausman y lo que de verdad compara 219
9.6 Manos a la obra 219
9.7 El error que comete casi todo tesista al elegir entre fijos y aleatorios 241
9.8 El error que cometen los investigadores en la redacción de artículos 242
RESUMEN Y CONCLUSIONES 242
EJERCICIOS RESUELTOS 243
Para tu tesis o artículo 245
Referencias 245
Capítulo 10 Series de tiempo y modelo ARIMA paso a paso 246
Lo que vas a lograr 246
10.1 En una serie, el orden de los datos es el dato 247
10.2 Tendencia y estaciones: los dos enemigos del promedio 247
10.3 ¿Hace falta diferenciar? Dos pruebas, dos hipótesis opuestas 247
10.4 Las piezas del ARIMA y su versión estacional 248
10.5 Identificar, estimar, diagnosticar, pronosticar 248
10.6 Manos a la obra 249
10.7 El error que comete casi todo tesista con ARIMA 269
10.8 El error que cometen los investigadores en la redacción de artículos 269
RESUMEN Y CONCLUSIONES 270
EJERCICIOS RESUELTOS 270
Para tu tesis o artículo 273
Referencias 273
Capítulo 11 Pronósticos con modelos VAR y cointegración 275
Lo que vas a lograr 275
11.1 Un VAR: cada variable con su pasado y el de las demás 276
11.2 Antes de leer nada: rezagos, estabilidad y residuos limpios 276
11.3 Causalidad de Granger: el pasado que ayuda a pronosticar 276
11.4 Impulso-respuesta y descomposición de varianza 277
11.5 Paseos aleatorios y regresión espuria 277
11.6 Cointegración y corrección de errores 277
11.7 Dos maneras de probar la cointegración 278
11.8 Manos a la obra 278
11.9 El error que comete casi todo tesista con VAR y cointegración 302

11.10 El error que cometen los investigadores en la redacción de artículos 303
RESUMEN Y CONCLUSIONES..... 303
EJERCICIOS RESUELTOS..... 303
Para tu tesis o artículo 305
Referencias 306
Parte Cuatro. Efectos, no solo asociaciones 307
Capítulo 12 Evaluación de impacto I: diferencias en diferencias..... 308
Lo que vas a lograr..... 308
12.1 Dos restas que, solas, mienten 309
12.2 La doble resta, en cuatro celdas..... 309
12.3 El supuesto que lo sostiene todo: tendencias paralelas 309
12.4 Los errores estándar y el problema de un solo tratado 310
12.5 Muchos periodos: efectos fijos bidireccionales y estudio de eventos 310
12.6 Cuando el tratamiento llega en fechas distintas..... 311
12.7 Manos a la obra..... 311
12.8 El error que cometen los tesistas al evaluar una política con DiD..... 330
12.9 El error que cometen los investigadores en la redacción de artículos 330
RESUMEN Y CONCLUSIONES..... 330
EJERCICIOS RESUELTOS..... 331
Para tu tesis o artículo 333
Referencias 334
Capítulo 13 Evaluación de impacto II: variables instrumentales..... 335
Lo que vas a lograr..... 335
13.1 Tres caminos por los que se cuela la endogeneidad 336
13.2 El instrumento: una palanca que solo toca la causa 336
13.3 La cuenta de Wald: dividir dos efectos ... 337
13.4 Mínimos cuadrados en dos etapas 337
13.5 ¿De quién es el efecto? Los cumplidores 337
13.6 Instrumentos débiles: cuando la palanca casi no mueve..... 338
13.7 Dos pruebas más: endogeneidad y sobreidentificación..... 338
13.8 Manos a la obra..... 339
13.9 El error que cometen los tesistas con variables instrumentales 357
13.10 El error que cometen los investigadores en la redacción de artículos 357
RESUMEN Y CONCLUSIONES..... 358
EJERCICIOS RESUELTOS..... 358
Para tu tesis o artículo 362
Referencias 362
Capítulo 14 Evaluación de impacto III: regresión discontinua y propensity score matching..... 363
Lo que vas a lograr..... 363
14.1 Una regla con un umbral 364
14.2 Lo único que debe cambiar en el corte 364
14.3 ¿Nítida o difusa?..... 364
14.4 Cuán cerca y con qué recta 364
14.5 Cuando la gente se acomoda al umbral ... 365
14.6 Manos a la obra: el PANES y el apoyo al gobierno 365
14.7 Sin corte ni sorteo: buscar parejas 376
14.8 Manos a la obra: emparejamiento paso a paso..... 377
14.9 ¿Cuál método elijo? 385
14.10 El error que cometen los tesistas con la discontinuidad y el emparejamiento 386
14.11 El error que cometen los investigadores en la redacción de artículos 386
RESUMEN Y CONCLUSIONES..... 386
EJERCICIOS RESUELTOS..... 387
Para tu tesis o artículo 389
Referencias 390
Parte Cinco. Reportar y dar un paso más..... 391

Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico 392
Lo que vas a lograr 392
15.1 Lo que tu lector necesita, en orden 393
15.2 La sección de método, en cinco párrafos . 393
15.3 Una tabla que se entiende sola 393
15.4 Cómo se reporta cada número 394
15.5 Lo que tus números no dicen 394
15.6 Manos a la obra 395
15.7 El error que comete casi todo tesista: pegar la ventana de EViews 410
15.8 El error que cometen los investigadores en la redacción de artículos 410
RESUMEN Y CONCLUSIONES 411
EJERCICIOS RESUELTOS 411
Para tu tesis o artículo 414
Referencias 414
Capítulo 16 Modelos ARDL y bounds test 415
Lo que vas a lograr 415
16.1 ¿Por qué otro modelo de cointegración? . 415
16.2 Un ARDL es una regresión con memoria 416
16.3 La forma de corrección de errores 416
16.4 El bounds test: dos cotas en lugar de una 417
16.5 Antes de creerle: supuestos, residuos y estabilidad 418
16.6 Manos a la obra 418
16.7 El error que comete casi todo tesista con ARDL 436
16.8 El error que cometen los investigadores en la redacción de artículos 437
RESUMEN Y CONCLUSIONES 437
EJERCICIOS RESUELTOS 437
Para tu tesis o artículo 440
Referencias 440
APÉNDICE A. EQUIVALENCIAS EVIEWS · STATA · PYTHON · R 441
APÉNDICE B. BASES DE DATOS DEL LIBRO 447

Capítulo 1

¿Por qué tu regresión miente? Correlación y causalidad

Hace unas décadas, los epidemiólogos tenían una pista prometedora contra el cáncer de pulmón. Las personas que comían muchas frutas y verduras ricas en carotenoides, y las que tenían niveles altos de betacaroteno en la sangre, enfermaban menos. Los estudios lo repetían una y otra vez.

¿Qué más hacía falta? Si el betacaroteno protege, basta con dárselo en una pastilla a quienes más lo necesitan: los fumadores.

En el suroeste de Finlandia alguien decidió comprobarlo como se debe. Se reclutó a 29 133 fumadores varones de 50 a 69 años. No se dejó que cada uno eligiera qué tomar. Se sorteó: unos recibieron vitamina E, otros betacaroteno, otros ambas cosas y otros una pastilla sin nada activo. Ni ellos ni quienes los atendían sabían qué le había tocado a cada uno. Los siguieron durante cinco a ocho años.

En 1994 se publicaron los resultados. Entre los hombres que recibieron betacaroteno hubo **más** cáncer de pulmón, no menos: un 18 % más que entre los que no lo recibieron. Su mortalidad total también fue mayor, en un 8 %. Los propios autores lo calificaron de inesperado (Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group, 1994).

Años de estudios decían «protege». El experimento dijo «no protege, y quizá daña». ¿Cómo pudo equivocarse tanta evidencia a la vez?

Piensa en quién tiene el betacaroteno alto en la sangre. Por lo general, alguien que come bien. Y quien come bien suele diferir en muchas otras cosas: cuánto fuma, cuánto se mueve, cuánto gana, cuán a menudo va al médico. Los estudios observacionales comparaban a personas distintas y le atribuían toda la diferencia a una sola molécula. El sorteo, en cambio, comparó a hombres iguales en todo, salvo en la pastilla.

Ahora piensa en tu investigación. Tienes una encuesta, una variable que te interesa y un programa que calcula una regresión en dos segundos. Te sale un coeficiente positivo con tres estrellas. ¿Cómo sabes que no te está pasando lo mismo?

De eso trata este capítulo. Y, en el fondo, de eso trata todo el libro.

Lo que vas a lograr

Al terminar este capítulo podrás:

- Distinguir cuándo un número describe una relación y cuándo mide un efecto.
- Explicar con una sola ecuación por qué una comparación simple puede engañarte.
- Simular en EViews, paso a paso, un caso donde la regresión miente y otro donde dice la verdad.

- Reconocer en tu propia tesis la pregunta que el jurado te va a hacer primero.

1.1 La correlación no te dice por qué

Empecemos por lo que ya sabes. Dos variables están correlacionadas cuando se mueven juntas: si una sube, la otra tiende a subir (o a bajar). La correlación es útil, porque te dice que hay algo. Sin embargo, no te dice **qué** hay.

Franz Messerli, un cardiólogo de Nueva York, lo mostró con un ejemplo que dio la vuelta al mundo. Reunió datos de 23 países y comparó cuánto chocolate se consume por habitante con cuántos premios Nobel ha ganado cada país por habitante. La relación salió fuerte y muy significativa, con Suiza a la cabeza de ambas listas. Messerli publicó el trabajo en una de las revistas médicas más leídas del mundo, y su propósito era advertir lo fácil que es tomarse en serio una correlación que no tiene sentido (Messerli, 2012).

¿Por qué pasa algo así? Cuando dos variables se mueven juntas, hay cuatro explicaciones posibles. Conviene que las tengas a mano, porque el jurado las va a tener.

1. Causalidad. X provoca Y. Por ejemplo, más fertilizante sube el rendimiento de una parcela, y una helada reduce la cosecha de papa.

2. Causalidad inversa. Y provoca X. Imagina que encuentras que las empresas que más gastan en publicidad son las que más venden. ¿La publicidad genera las ventas? Puede ser, pero muchas empresas fijan su presupuesto de publicidad como una parte de lo que venden. Entonces vender más hace gastar más en publicidad, y no solo al revés.

3. Una tercera variable. Algo distinto mueve a las dos. En el caso del chocolate, una candidata obvia es la riqueza del país: un país rico consume más chocolate y también financia más universidades y laboratorios. En la historia del betacaroteno, comer bien iba de la mano con muchos otros hábitos saludables. A esa tercera variable los econometristas le dicen **variable de confusión** o **factor de confusión**.

4. El azar. Por ejemplo, las exportaciones de quinua y las lluvias de una ciudad lejana pueden moverse juntas durante algunos años sin tener nada que ver. Con suficientes variables, algunas se van a correlacionar por pura casualidad. Si cruzas cien pares de variables sin ninguna relación, unas cinco te saldrán «significativas» al 5 %. Así funciona ese umbral.

Fíjate en algo incómodo. Las cuatro explicaciones producen **la misma tabla** en tu programa estadístico. Un coeficiente, un error estándar, unas estrellas. El software no sabe cuál de las cuatro es la verdadera. Tú sí tienes que saberlo.

1.2 La pregunta que de verdad importa: ¿qué habría pasado si…?

Vamos a un caso más cercano. Supongamos que Diego prepara su tesis sobre el rendimiento escolar en Trujillo. Tiene datos de 5 000 estudiantes de quinto de secundaria: sus notas finales, en la escala de 0 a 20, y si pagaron o no clases particulares. Los que pagaron clases sacaron, en promedio, tres puntos más. ¿Las clases particulares suben la nota en tres puntos?

La pregunta que Diego quiere responder no es «¿quiénes sacaron más nota?». Es otra, más difícil: **¿cuánto habría sacado cada estudiante si no hubiera tomado clases?**

Esa nota que no vemos tiene nombre. Se llama **resultado potencial**. Cada estudiante tiene dos:

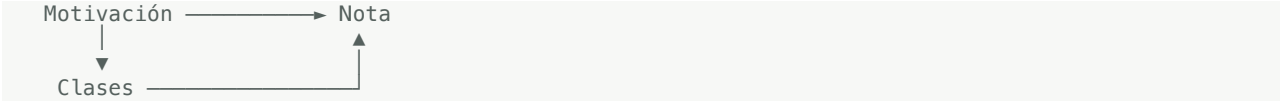
- La nota que sacaría **con** clases particulares.
- La nota que sacaría **sin** clases particulares.

El efecto de las clases para ese estudiante es la diferencia entre las dos. Simple. El problema es que la vida no nos deja ver las dos. Cada estudiante tomó clases o no las tomó. Uno de sus dos resultados potenciales quedó para siempre en el terreno del «¿qué habría pasado si…?». Los libros lo llaman el **problema fundamental de la inferencia causal** (Cunningham, 2021).

¿Entonces no se puede medir nada? Sí se puede, aunque no estudiante por estudiante: se mide en promedio, y con cuidado. Y el cuidado empieza por entender dónde está la trampa.

1.3 Dibuja tu problema antes de estimarlo

Antes de abrir EViews, toma una hoja y dibuja flechas. Parece un juego de niños, pero es una de las herramientas más serias de la inferencia causal moderna. Cada flecha significa «esto provoca aquello». El problema de Diego, dibujado, se ve así:



La flecha de abajo, de Clases a Nota, es la que Diego quiere medir. Las otras dos son el problema: la motivación empuja al estudiante hacia las clases y, por otro camino, empuja también su nota hacia arriba. Ese camino alternativo, que pasa por la motivación, se llama **camino de puerta trasera**. Mientras esté abierto, parte de la relación entre clases y nota viaja por él, y tu regresión la confunde con el efecto de las clases (Cunningham, 2021).

Dibujar te obliga a hacer explícito lo que crees sobre el mundo. Si alguien te dice «te faltó considerar el ingreso familiar», no discuten sobre estrellas y valores p: discuten sobre una flecha. ¿El ingreso familiar mueve la probabilidad de pagar clases? Seguro que sí. ¿Mueve la nota por otro camino? Probablemente, porque un hogar con más ingresos también compra libros, internet y tranquilidad. Entonces es otra puerta trasera que tienes que cerrar.

Hay tres maneras de cerrar una puerta trasera, y este libro las recorre todas. Puedes medir el confusor y controlarlo, usar datos donde el confusor no cambia o encontrar una fuente de variación que no pase por él. Por ejemplo, un sorteo. Por ahora basta con que dibujes. Vas a ver que la mitad de las preguntas del jurado se responden mirando ese dibujo.

1.4 La trampa, en una sola línea

Cuando Diego compara el promedio de los que tomaron clases con el de los que no, está calculando esto:

Diferencia observada = efecto causal + sesgo de selección

Esta ecuación es la más importante del libro. Angrist y Pischke la ponen en el centro de su manual, y con razón (Angrist & Pischke, 2014). Léela despacio.

El **efecto causal** es lo que Diego quiere: cuánto suben la nota las clases a quienes las tomaron.

El **sesgo de selección** es la diferencia que ya existía entre los dos grupos antes de las clases. ¿Quiénes pagan clases particulares? Los estudiantes más motivados. Los que tienen padres pendientes de sus notas. Los que tienen algo de dinero extra en casa. Esos estudiantes habrían sacado mejor nota **aunque no hubieran tomado ni una clase**.

Si el sesgo de selección es grande, la diferencia observada puede ser el triple del efecto real. O puede tener el signo contrario, como en la historia del betacaroteno. Y ningún valor p te va a avisar.

Por eso los econometristas hacemos siempre la misma pregunta incómoda: **¿los dos grupos eran comparables antes del tratamiento?** Si la respuesta es no, tu diferencia mezcla dos cosas y no sabes cuánto pesa cada una.

1.5 El truco del experimento

¿Cómo se elimina el sesgo de selección? Con una moneda.

Si Diego pudiera sortear quién recibe clases particulares, la motivación, el dinero y los padres atentos quedarían repartidos más o menos igual entre los dos grupos. Nadie eligió. El azar eligió. Entonces el sesgo de selección se va a cero y la diferencia observada mide el efecto causal. Eso hizo el estudio de Finlandia, y por eso encontró lo que años de estudios observacionales no vieron.

América Latina tiene su propio ejemplo famoso.

Colombia tuvo un programa, PACES, que entregaba vouchers para pagar parte de la pensión de un colegio secundario privado. Los vouchers se renovaban solo si el estudiante mantenía un rendimiento satisfactorio. Y los vouchers se repartieron **por sorteo** (Angrist et al., 2002).

Como el azar decidió quién ganaba, los ganadores y los perdedores eran comparables, y la diferencia entre ellos se podía leer como efecto del voucher. Tres años después de los sorteos, los ganadores tenían unos 10 puntos porcentuales más de probabilidad de haber terminado el 8.º grado, sobre todo porque repetían menos de año. Además, sacaban 0,2 desviaciones estándar más en las pruebas de rendimiento (Angrist et al., 2002).

Todo empezó con un sorteo.

1.6 ¿Y si no puedes hacer un experimento?

Aquí viene la mala noticia: casi nunca vas a poder. No puedes sortear quién estudia una maestría, quién vive en Lima o quién recibe un crédito. Tu investigación probablemente usa una encuesta que ya existe, como la ENAHO, o datos de una empresa, o series del Banco Central.

La buena noticia es que la econometría es, justamente, el arte de acercarse al experimento cuando no lo tienes. Este libro recorre las herramientas en orden de dificultad:

- **Controlar por lo que sí observas** (capítulos 4 a 8). Si el confusor está en tus datos, lo metes en la regresión.
- **Seguir a las mismas personas o empresas en el tiempo** (capítulo 9). Los datos de panel eliminan lo que no cambia, aunque no lo midas.
- **Buscar algo que se parezca al azar** (capítulos 12 a 14). Variables instrumentales, diferencias en diferencias, regresión discontinua y emparejamiento.

Cada herramienta tiene un supuesto. Tu trabajo, y el de este libro, es saber cuál es y defender por qué se cumple en tu caso.

Basta de teoría. Vamos a ver la trampa con tus propios ojos.

1.7 Manos a la obra

Abre EViews. En la parte superior verás la ventana **Command**: ahí escribes una instrucción, pulsas Enter y EViews la ejecuta. Copia los bloques en orden, línea por línea. Si prefieres tenerlos todos juntos, abre el archivo

`cap01.prg` del repositorio del libro con *File › Open › Programs* y pulsa el botón *Run*; el prefacio te explica cómo descargarlo. No necesitas instalar nada adicional: todo lo que usamos viene con EViews.

Vamos a hacer algo que en la vida real es imposible: **crear un mundo donde conocemos la verdad**. Simularemos a los 5 000 estudiantes de Diego. Decidiremos nosotros que las clases particulares suben la nota exactamente **1 punto**. Y luego veremos qué nos dice cada método. Los datos de esta sección son simulados, no provienen de ninguna encuesta.

La clave está en una variable que nadie mide en las encuestas: la motivación. En nuestro mundo, la motivación hace dos cosas. Sube la nota. Y aumenta la probabilidad de pagar clases.

Paso 1. Crea 5 000 estudiantes con una motivación oculta

```
wfcreate(wf=cap01) u 5000
rndseed 2026
series motivacion = nrnd
show motivacion.hist
```

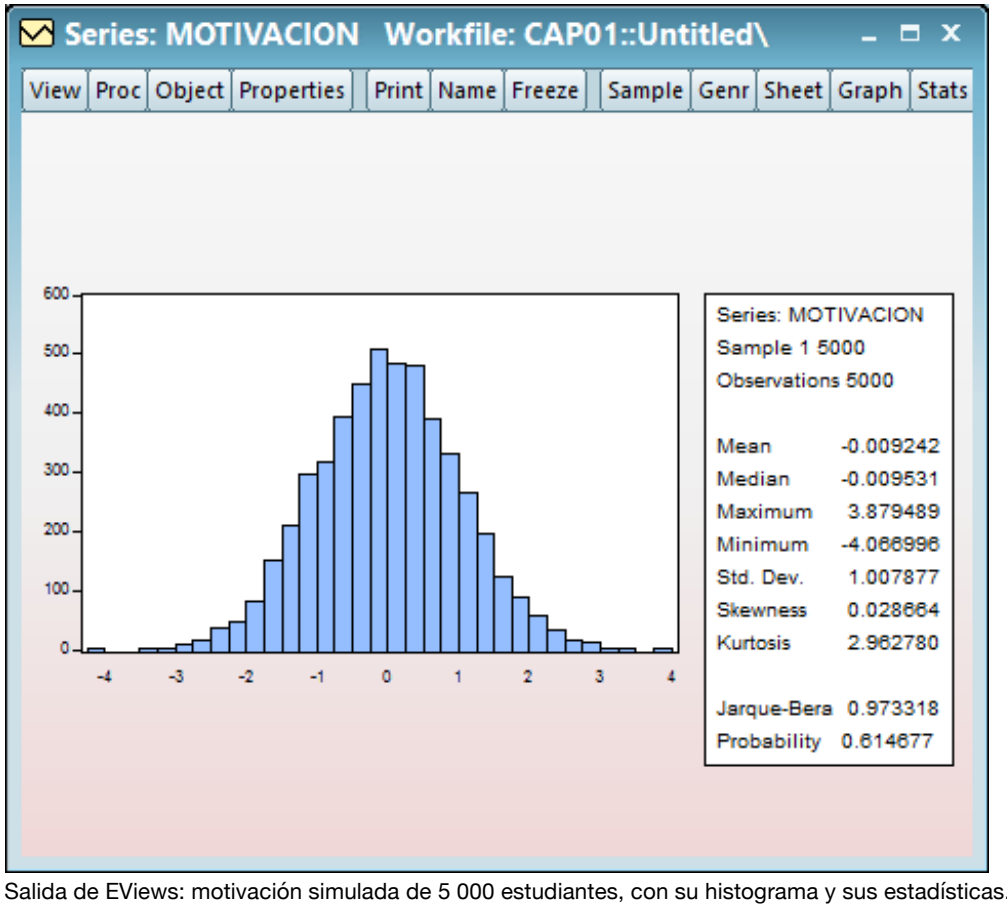


FIGURA 1.1 Salida de EViews: motivación simulada de 5 000 estudiantes, con su histograma y sus estadísticas. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

¿Qué hicimos? `wfcreate` crea un **workfile**, el archivo de trabajo de EViews: `wf=cap01` le pone nombre, `u` indica que los datos no tienen fechas (son personas, no meses) y `5000` crea 5 000 filas, una por estudiante.

`rndseed 2026` fija la semilla del generador de números aleatorios: gracias a ella, tú obtendrás exactamente los mismos números que aparecen en el libro. La función `nrnd` le da a cada estudiante una motivación al azar, con promedio 0 y desviación estándar 1. Por último, `show motivacion.hist` abre la serie en su vista de histograma.

La ventana confirma lo que pedimos: una media de −0,009 y una desviación estándar de 1,008. Nadie mide esta motivación en una encuesta, pero en nuestro mundo existe.

Un consejo práctico: corre los bloques una sola vez y en orden. Si repites un bloque que sortea números, la secuencia cambia y tus cifras dejarán de coincidir con las del libro. Si te pasa, vuelve a empezar desde el paso 1.

Paso 2. Haz que los más motivados paguen clases

```
series clases = @rbinom(1, @logit(1.5*motivacion))
show clases.freq
```

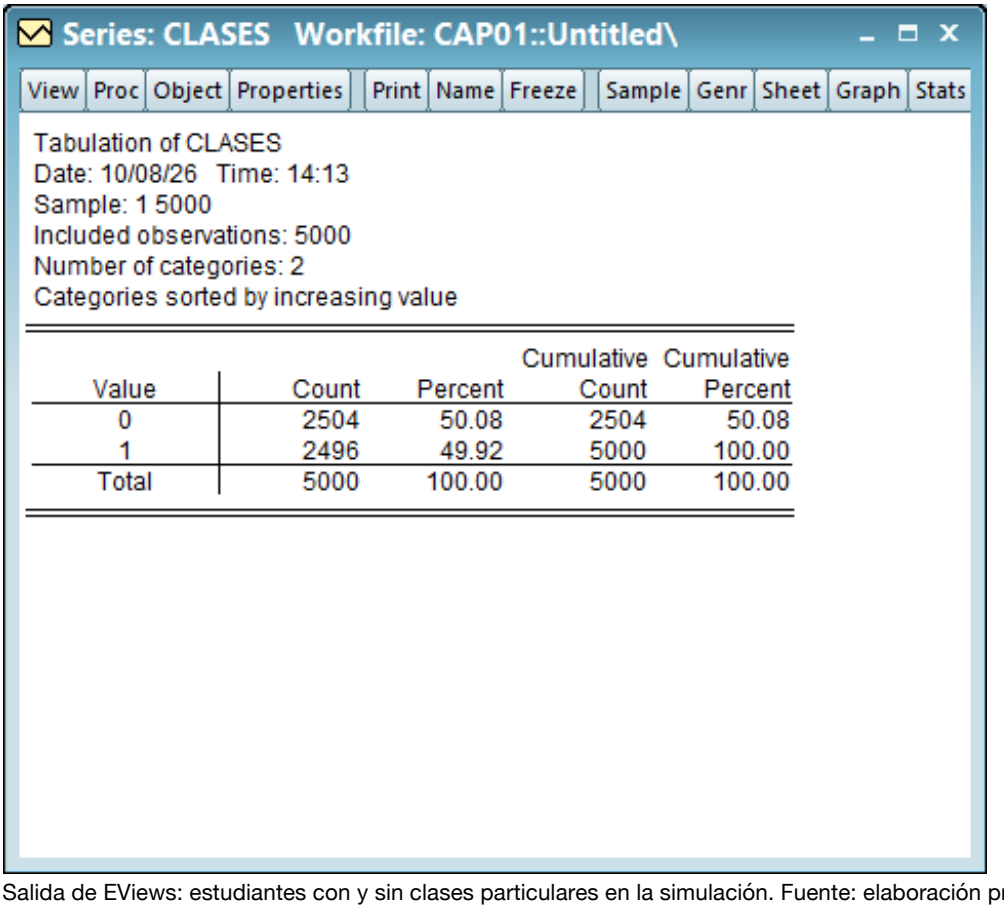
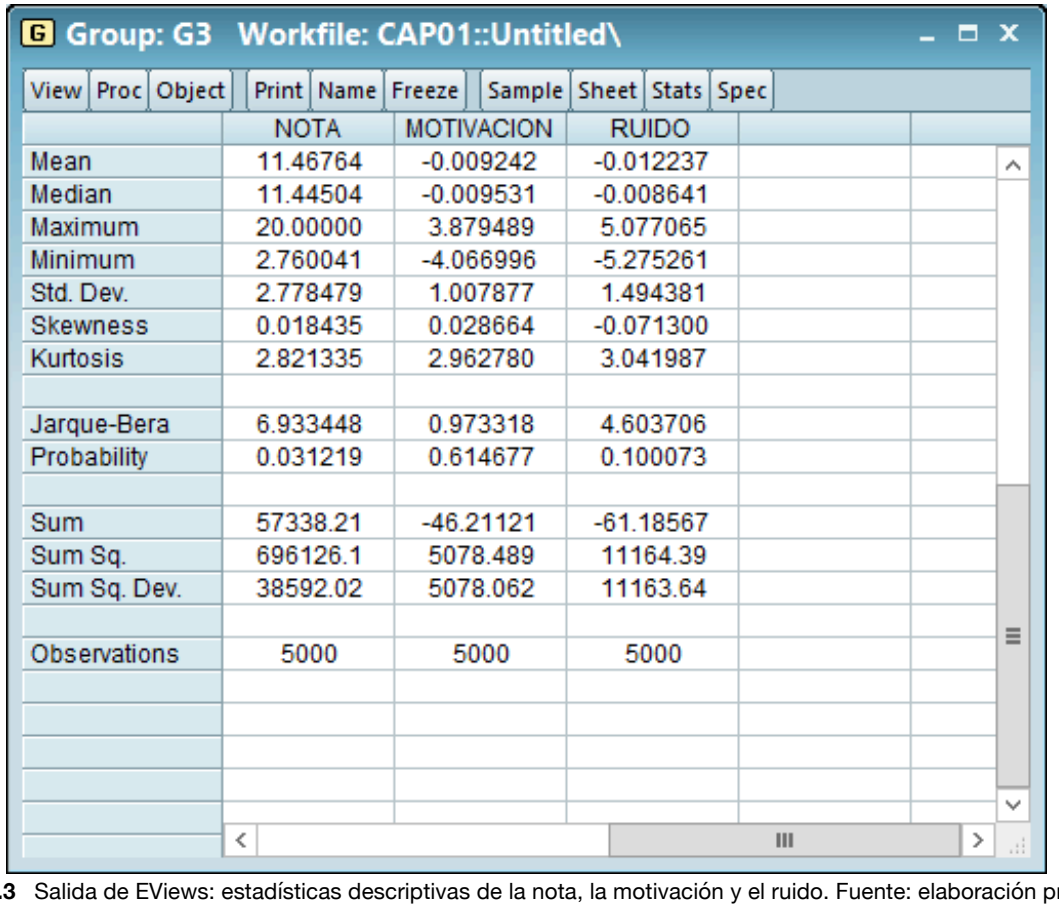


FIGURA 1.2 Salida de EViews: estudiantes con y sin clases particulares en la simulación. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Esta línea es el corazón del problema. La función `@logit()` convierte la motivación en una probabilidad entre 0 y 1. Por ejemplo, un estudiante con motivación 0 tiene 50 % de probabilidad de pagar clases, y uno muy motivado se acerca al 100 %. Luego `@rbinom(1, p)` lanza una moneda cargada con esa probabilidad y anota 1 si el estudiante pagó clases y 0 si no. La vista `freq` cuenta cuántos hay de cada tipo: 2 496 estudiantes (49,92 %) tomaron clases y 2 504 no.

Paso 3. Construye la nota con un efecto real de 1 punto

```
series ruido = 1.5*nrnd
series nota = 11 + 1*clases + 2*motivacion + ruido
nota = @recode(nota<0, 0, @recode(nota>20, 20, nota))
group g3 nota motivacion ruido
show g3.stats
```



	NOTA	MOTIVACION	RUIDO
Mean	11.46764	-0.009242	-0.012237
Median	11.44504	-0.009531	-0.008641
Maximum	20.00000	3.879489	5.077065
Minimum	2.760041	-4.066996	-5.275261
Std. Dev.	2.778479	1.007877	1.494381
Skewness	0.018435	0.028664	-0.071300
Kurtosis	2.821335	2.962780	3.041987
Jarque-Bera	6.933448	0.973318	4.603706
Probability	0.031219	0.614677	0.100073
Sum	57338.21	-46.21121	-61.18567
Sum Sq.	696126.1	5078.489	11164.39
Sum Sq. Dev.	38592.02	5078.062	11163.64
Observations	5000	5000	5000

FIGURA 1.3 Salida de EViews: estadísticas descriptivas de la nota, la motivación y el ruido. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Lee la segunda línea despacio, porque ahí está la verdad que luego intentaremos descubrir. La nota parte de 11 puntos. Suma **1 punto** si el estudiante tomó clases, 2 puntos por cada unidad de motivación y un poco de ruido, que representa la suerte del día del examen. La tercera línea recorta la nota para que no baje de 0 ni pase de 20: `@recode(condición, a, b)` devuelve `a` si la condición se cumple y `b` si no.

Las dos últimas líneas presentan un objeto nuevo. Un **group** junta varias series para verlas lado a lado, y su vista `stats` calcula las estadísticas de todas a la vez. El resumen muestra una nota promedio de 11,47 y una motivación promedio prácticamente nula (−0,009).

Paso 4. Primer intento: comparar promedios

Esto es lo que haría la mayoría. Separar en dos grupos y comparar.

```
show nota.statby clases
```

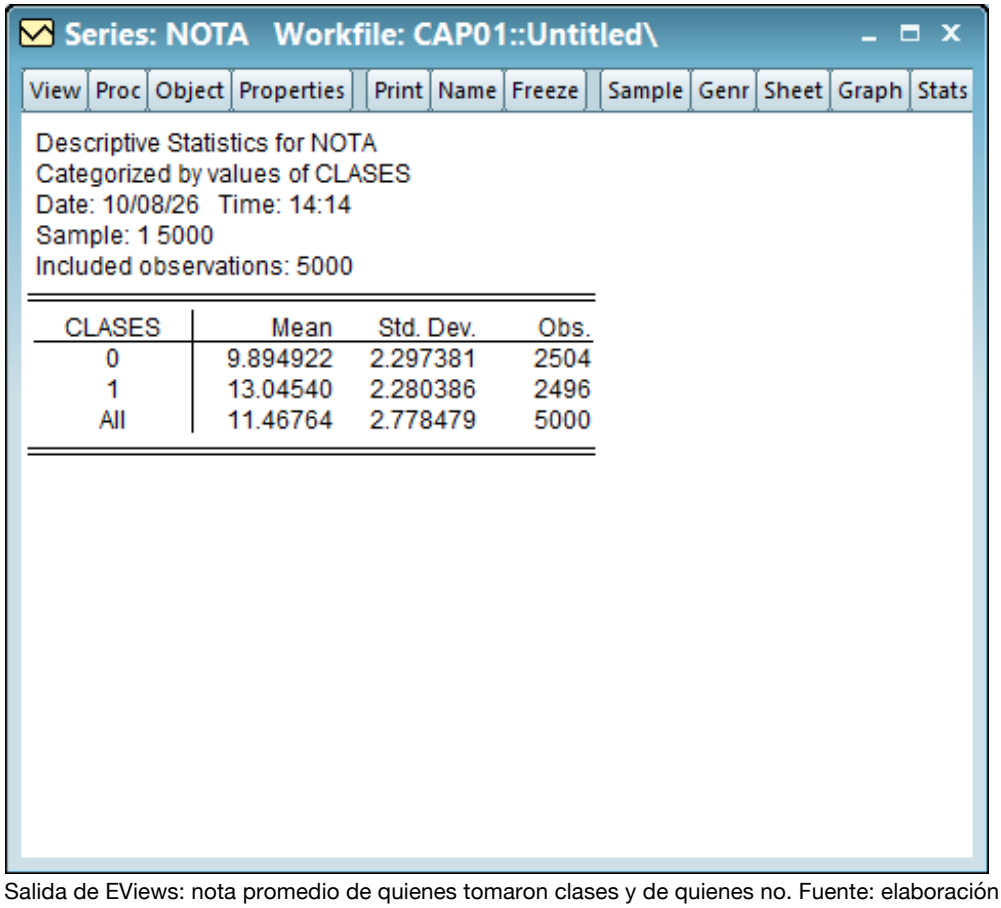


FIGURA 1.4 Salida de EViews: nota promedio de quienes tomaron clases y de quienes no. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

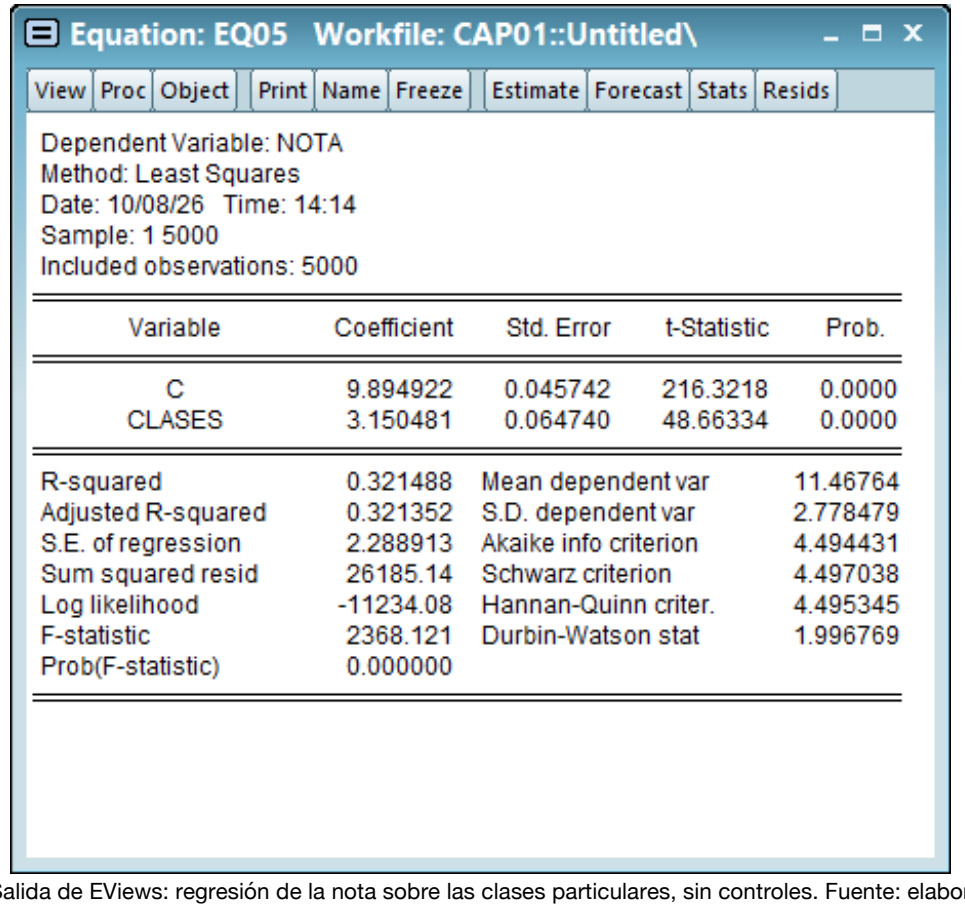
La vista `statby` calcula las estadísticas de una serie por cada valor de otra. Los que no tomaron clases promedian 9,89; los que sí, 13,05. La diferencia es de 3,15 puntos. Sabemos que la verdad es uno. La comparación simple triplicó el efecto.

Un detalle de lectura: EViews escribe los decimales con punto, como en inglés. En el texto del libro usamos la coma, así que 13.04540 en la pantalla es 13,05 en la prosa.

Paso 5. Segundo intento: la regresión

«Bueno», dirás, «pero yo no comparo promedios, yo corro una regresión». Veamos.

```
equation eq05.ls nota c clases
show eq05
```



Equation: EQ05 Workfile: CAP01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: NOTA
Method: Least Squares
Date: 10/08/26 Time: 14:14
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.894922	0.045742	216.3218	0.0000
CLASES	3.150481	0.064740	48.66334	0.0000

R-squared	0.321488	Mean dependent var	11.46764
Adjusted R-squared	0.321352	S.D. dependent var	2.778479
S.E. of regression	2.288913	Akaike info criterion	4.494431
Sum squared resid	26185.14	Schwarz criterion	4.497038
Log likelihood	-11234.08	Hannan-Quinn criter.	4.495345
F-statistic	2368.121	Durbin-Watson stat	1.996769
Prob(F-statistic)	0.000000		

FIGURA 1.5 Salida de EViews: regresión de la nota sobre las clases particulares, sin controles. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

La primera línea crea un objeto **equation** llamado `eq05` y lo estima por mínimos cuadrados (`ls`). Después del punto va la variable dependiente y luego las explicativas; la `c` es la constante. EViews no la agrega solo: si la olvidas, estimas una recta que pasa por el origen.

El coeficiente de `CLASES` es 3,15. El valor p (columna `Prob.`) es 0,0000. ¿Y el intervalo de confianza? EViews no lo muestra en esta tabla, pero lo tiene a un comando de distancia.

```
eq05.cinterval 0.95
```

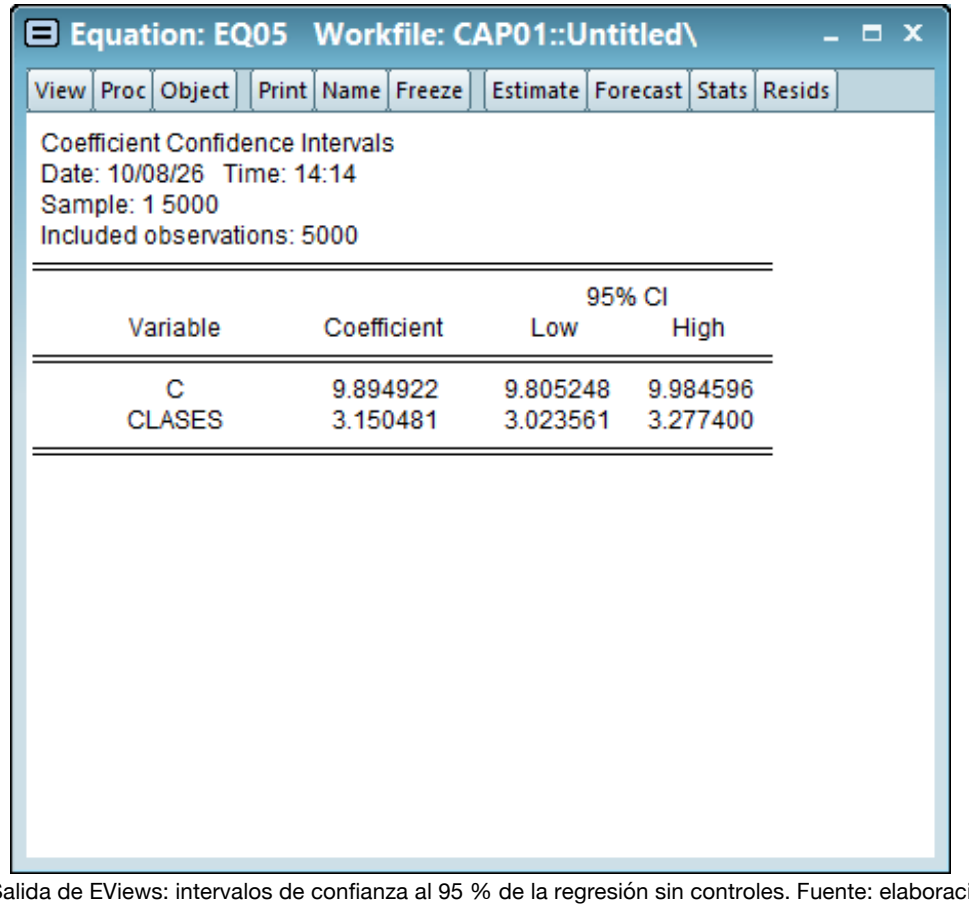


FIGURA 1.6 Salida de EViews: intervalos de confianza al 95 % de la regresión sin controles. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

El intervalo de confianza al 95 % va de 3,02 a 3,28: ni siquiera se acerca al 1 verdadero. Es la tabla más convincente que verás en tu vida, y está equivocada.

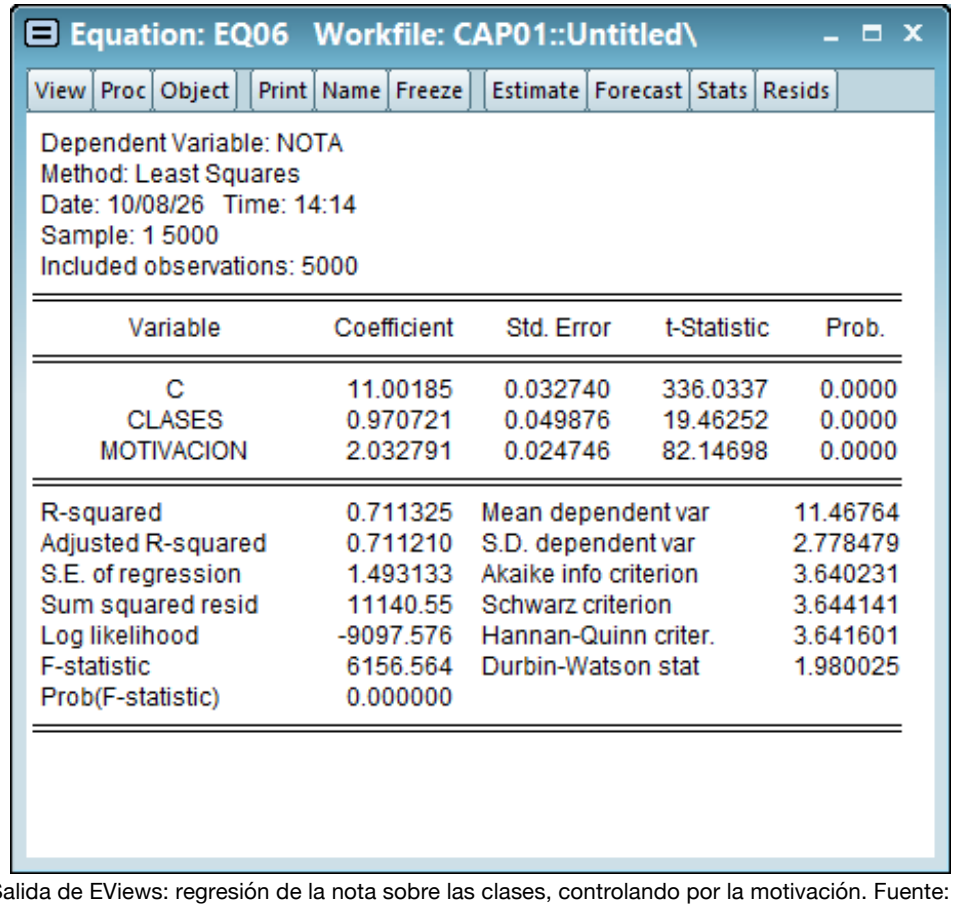
Mira además la constante, `C`: 9,89. Es exactamente el promedio de los que no tomaron clases en el paso 4. Y 9,89 más 3,15 da 13,05: el promedio de los que sí las tomaron.

Detente aquí un momento, porque esta es la lección del capítulo. **Una regresión con una sola variable es exactamente la misma comparación de promedios**, solo que con más decimales. Ni el valor p ni el tamaño de la muestra corrigen el sesgo de selección. Con 50 000 estudiantes, el error sería igual de grande; solo estarías más seguro de él.

Paso 6. Tercer intento: si pudieras medir la motivación

En nuestro mundo simulado tenemos algo que Diego no tiene: la motivación de cada estudiante. Metámosla en la regresión.

```
equation eq06.ls nota c clases motivacion
show eq06
```



Equation: EQ06 Workfile: CAP01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: NOTA
Method: Least Squares
Date: 10/08/26 Time: 14:14
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.00185	0.032740	336.0337	0.0000
CLASES	0.970721	0.049876	19.46252	0.0000
MOTIVACION	2.032791	0.024746	82.14698	0.0000

R-squared	0.711325	Mean dependent var	11.46764
Adjusted R-squared	0.711210	S.D. dependent var	2.778479
S.E. of regression	1.493133	Akaike info criterion	3.640231
Sum squared resid	11140.55	Schwarz criterion	3.644141
Log likelihood	-9097.576	Hannan-Quinn criter.	3.641601
F-statistic	6156.564	Durbin-Watson stat	1.980025
Prob(F-statistic)	0.000000		

FIGURA 1.7 Salida de EViews: regresión de la nota sobre las clases, controlando por la motivación. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Ahora el coeficiente de `CLASES` es 0,97, con un error estándar de 0,05: el 1 verdadero queda a menos de un error estándar. La regresión dejó de mentir. ¿Qué cambió? Le dimos la variable de confusión. Al comparar estudiantes con **la misma** motivación, el sesgo de selección desapareció. Eso es lo que significa «controlar» por una variable, y lo verás a fondo en el capítulo 5.

Mira también el coeficiente de `MOTIVACION`: 2,03. Es casi justo lo que pusimos al crear los datos.

Paso 7. Cuarto intento: el experimento

¿Y si, en lugar de dejar que cada estudiante eligiera, sorteáramos las clases? Vamos a usar a los mismos estudiantes, con la misma motivación y la misma suerte en el examen. Lo único que cambia es quién decide: ahora decide una moneda.

```
series clases_azar = @rbinom(1, 0.5)
series nota_azar = 11 + 1*clases_azar + 2*motivacion + ruido
nota_azar = @recode(nota_azar<0, 0, @recode(nota_azar>20, 20, nota_azar))
equation bal07.ls motivacion c clases_azar
show bal07
```

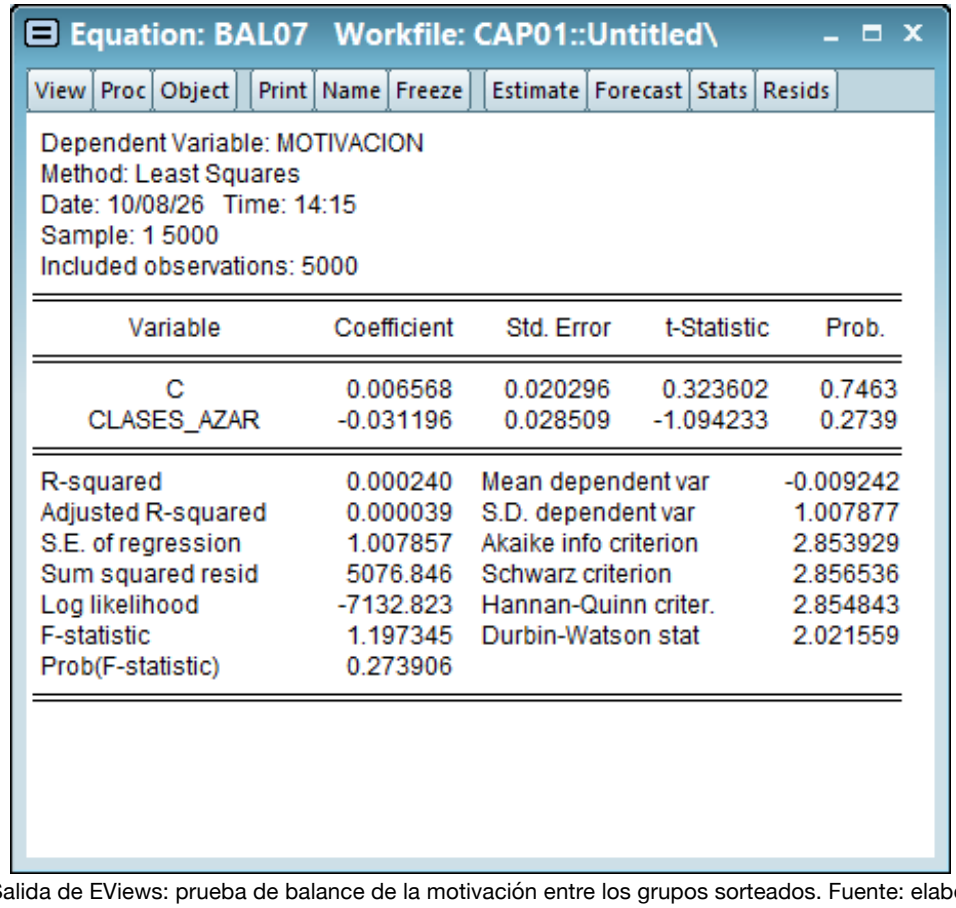


FIGURA 1.8 Salida de EViews: prueba de balance de la motivación entre los grupos sorteados. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Antes de estimar la nota, revisamos el **balance**: ¿los dos grupos sorteados tienen la misma motivación? Para comprobarlo usamos el truco del paso 5 al revés. Una regresión de la motivación sobre la variable del sorteo compara los promedios de los dos grupos: la constante es el promedio del grupo sin clases y el coeficiente, la diferencia.

La diferencia es de apenas −0,03, y su valor p es 0,27. No hay ninguna diferencia que merezca ese nombre. Si prefieres la prueba t clásica, la vista `motivacion.testby clases_azar` te da el mismo estadístico t, con el signo cambiado.

El sorteo repartió la motivación de forma pareja **sin que tuviéramos que medirla**. Ahora sí, la regresión simple.

```
equation eq08.ls nota_azar c clases_azar
show eq08
```

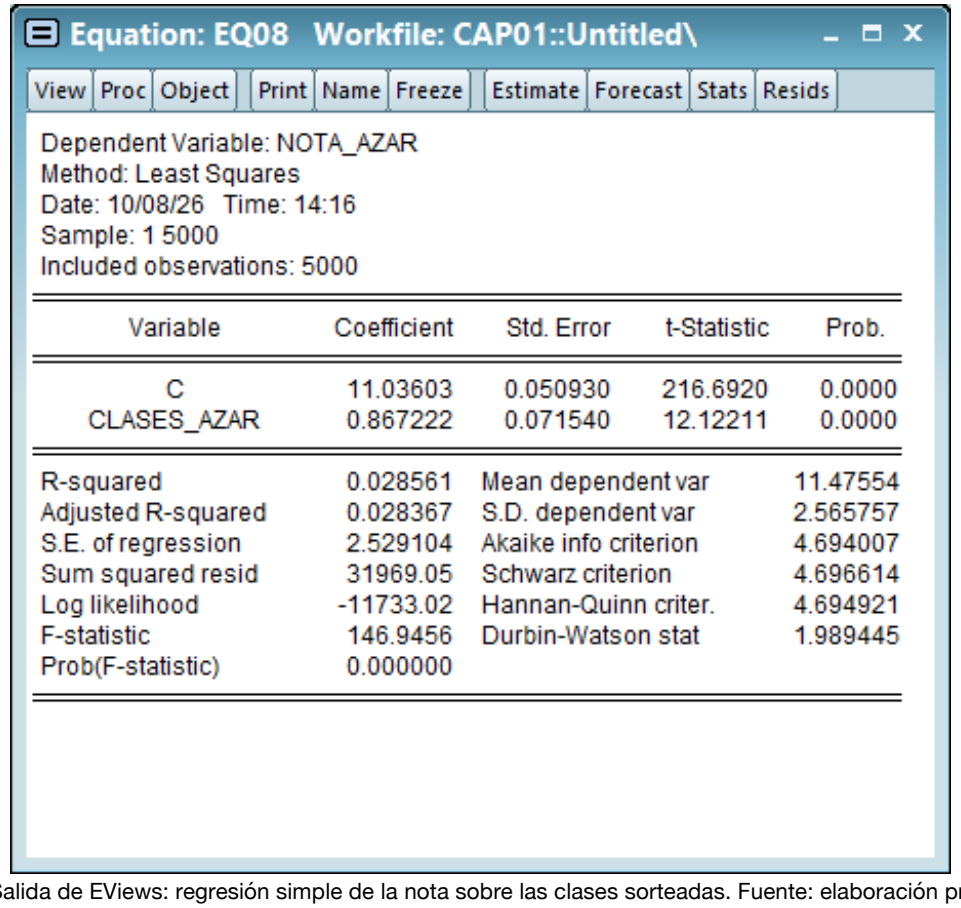


FIGURA 1.9 Salida de EViews: regresión simple de la nota sobre las clases sorteadas. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

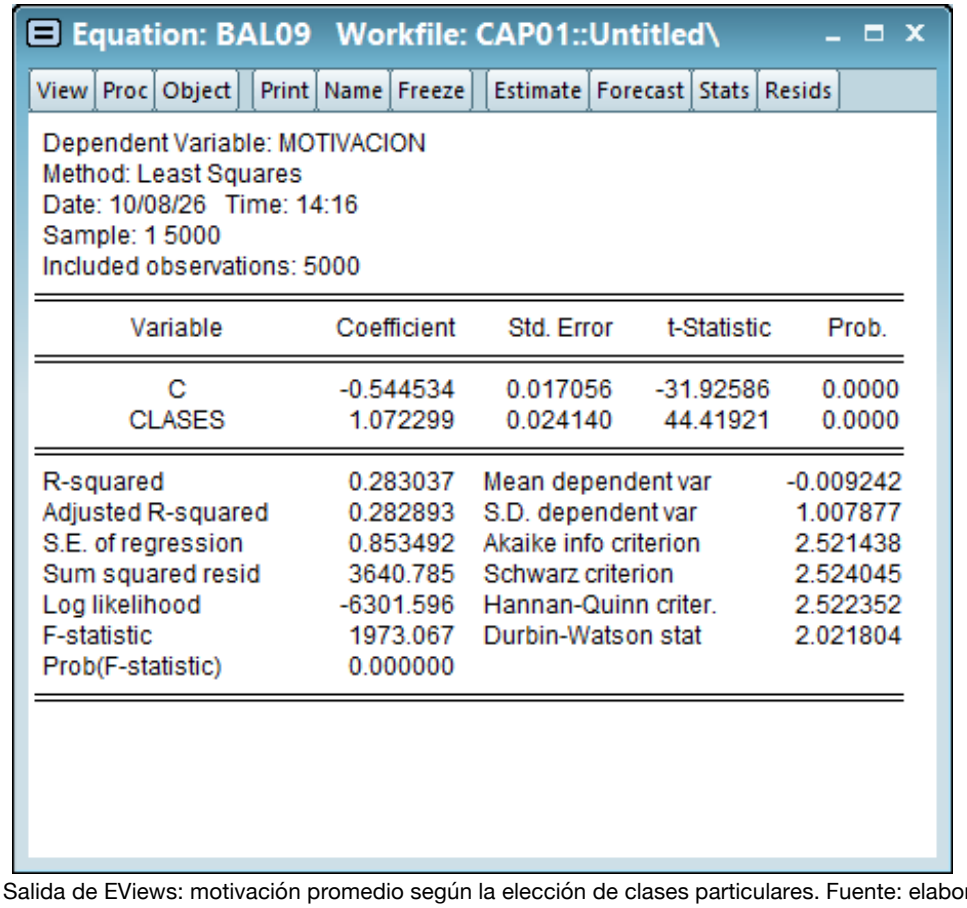
Sin ningún control, el coeficiente es 0,87, con un error estándar de 0,07. Queda a menos de dos errores estándar del 1 verdadero: el intervalo de confianza al 95 % lo contiene. El azar hizo el trabajo que en el tercer intento hizo el control. ¿Por qué no sale 1,00 exacto? Porque un solo sorteo nunca reparte todo de forma perfecta; con otra semilla saldría un poco por encima o un poco por debajo.

Fíjate en el R cuadrado: apenas 0,03. La regresión explica poco de la nota, porque la motivación y la suerte pesan mucho. Y aun así el efecto está bien medido. Un R cuadrado bajo no significa que el coeficiente esté mal. Imagina un programa que sube la nota un punto a cada estudiante: sería un gran programa, aunque explique poco de por qué unos sacan 8 y otros 18.

Paso 8. La prueba del delito

Volvamos a los datos originales, donde cada uno eligió. Comparemos la motivación de los dos grupos con la misma prueba de balance.

```
equation bal09.ls motivacion c clases
show bal09
```



Equation: BAL09 Workfile: CAP01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: MOTIVACION
Method: Least Squares
Date: 10/08/26 Time: 14:16
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.544534	0.017056	-31.92586	0.0000
CLASES	1.072299	0.024140	44.41921	0.0000

R-squared	0.283037	Mean dependent var	-0.009242
Adjusted R-squared	0.282893	S.D. dependent var	1.007877
S.E. of regression	0.853492	Akaike info criterion	2.521438
Sum squared resid	3640.785	Schwarz criterion	2.524045
Log likelihood	-6301.596	Hannan-Quinn criter.	2.522352
F-statistic	1973.067	Durbin-Watson stat	2.021804
Prob(F-statistic)	0.000000		

FIGURA 1.10 Salida de EViews: motivación promedio según la elección de clases particulares. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Ahí está el sesgo de selección, a la vista. Los que no tomaron clases tenían, en promedio, una motivación de −0,54 (la constante). Los que sí, 1,07 más: alrededor de 0,53. El estadístico t es 44,42 y el valor p es 0,0000.

Como cada unidad de motivación vale 2 puntos de nota, esa sola diferencia explica unos 2,14 de los 3,15 puntos que vimos en el primer intento. Haz la cuenta: 1 punto de efecto real más 2,14 de sesgo. La ecuación de la trampa, con números.

Paso 9. Controlar no es magia: la proxy ruidosa

El tercer intento funcionó porque teníamos la motivación exacta de cada estudiante. En tu investigación no la vas a tener. Ninguna encuesta pregunta «¿cuán motivado estás, en una escala perfecta?». Lo que sí suele haber es algo parecido, por ejemplo las horas de estudio que el estudiante declara. A esa variable parecida se le llama **proxy** o variable sustituta.

¿Basta con controlar por la proxy? Probemos. Vamos a crear unas horas de estudio que dependen de la motivación, pero con mucho ruido, como pasa con cualquier respuesta de encuesta.

```
series horas = 6 + 2*motivacion + 3*nrnd
horas = @recode(horas<0, 0, horas)
group g10 motivacion horas
show g10.cor
```

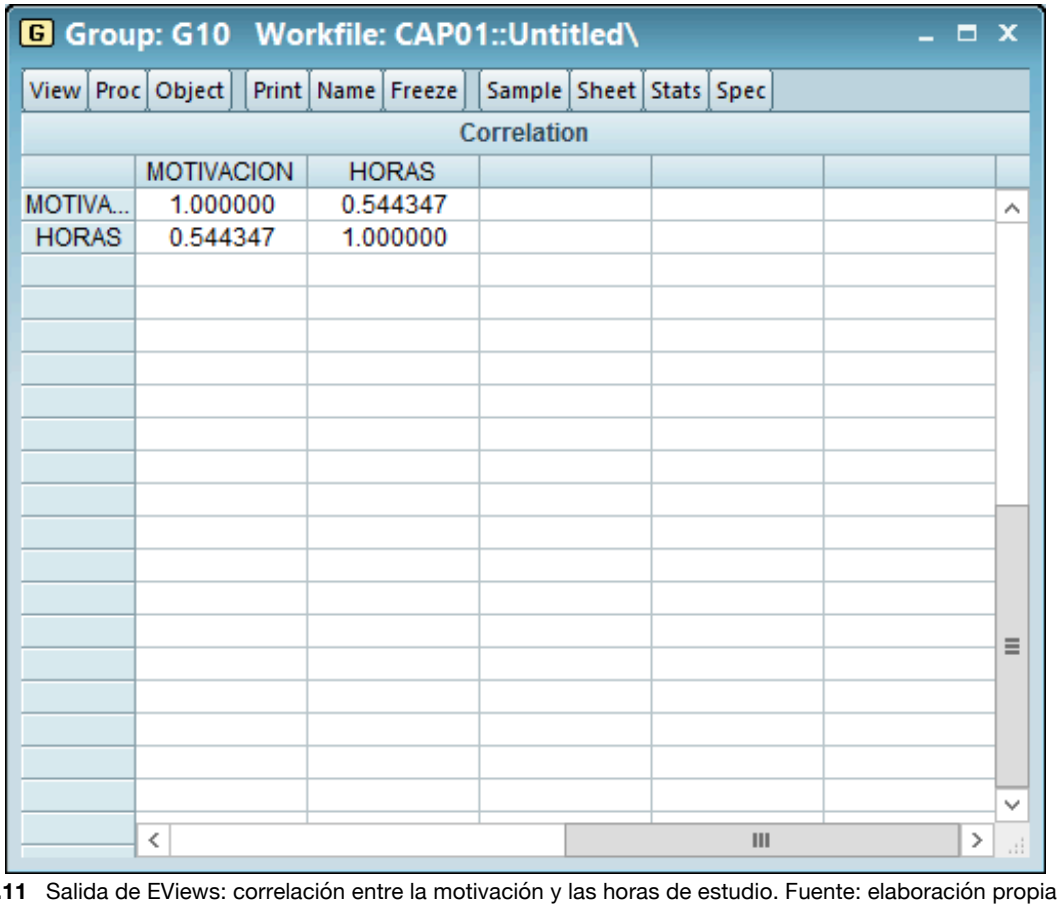


FIGURA 1.11 Salida de EViews: correlación entre la motivación y las horas de estudio. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Las horas de estudio se parecen a la motivación: la correlación es 0,54. ¿Será suficiente?

```
equation eq11.ls nota c clases horas
show eq11
```

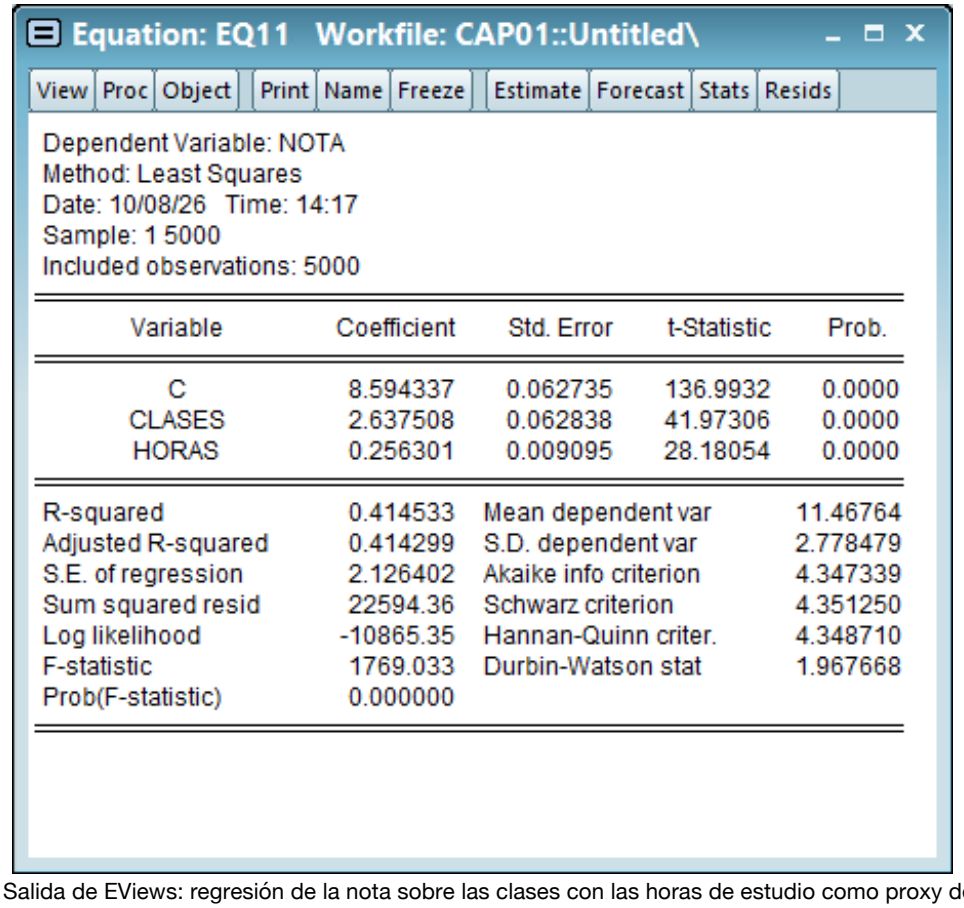


FIGURA 1.12 Salida de EViews: regresión de la nota sobre las clases con las horas de estudio como proxy de la motivación. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

No lo es. El coeficiente de las clases apenas bajó de 3,15 a 2,64. Seguimos muy lejos del 1 verdadero, y la tabla sigue luciendo impecable, con su valor p de 0,0000.

La lección es incómoda, pero vale oro: **controlar por una medida imperfecta del confusor solo corrige una parte del sesgo**. Cuanto peor mide la proxy al confusor, más sesgo queda. Por eso un buen econometrista no se conforma con «ya controlé por todo lo que tenía». Se pregunta qué tan bien mide cada control lo que de verdad le preocupa y, cuando la respuesta es «más o menos», busca una estrategia distinta. Esas estrategias llenan la última parte del libro.

Pongamos todos los intentos lado a lado.

TABLA 1.1 Cinco formas de estimar el efecto de las clases particulares (efecto real: 1 punto)

Método	Efecto estimado	¿Acierta?
Comparar promedios	3,15	No: triplica el efecto
Regresión simple	3,15	No: es lo mismo con más decimales
Regresión con el confusor	0,97	Sí, si mides el confusor
Regresión con una proxy del confusor	2,64	No: corrige solo una parte
Experimento (sorteo)	0,87	Sí, dentro de su margen de error, sin medir nada más

Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg); cifras de las figuras 1.4, 1.5, 1.7, 1.12 y 1.9.

1.8 El error que comete casi todo tesista

Abre el repositorio de tesis de cualquier universidad y busca títulos que empiecen con «Influencia de…» o «Impacto de…». Vas a encontrar muchos. Muchos se apoyan en una encuesta aplicada una sola vez, a un solo grupo, sin sorteo y sin ninguna estrategia para enfrentar el sesgo de selección. Y concluyen que X «influye significativamente» en Y porque el valor p salió menor que 0,05.

El error no está en el cálculo. Está en la palabra. «Significativo» quiere decir que la relación probablemente no es producto del azar: descarta la cuarta explicación de nuestra lista. No dice nada de las otras tres. Un coeficiente puede ser significativo y estar contaminado por causalidad inversa o por un confusor, como acabas de ver con el 3,15.

¿Cómo lo evitas? Con dos hábitos.

Primero, elige el verbo según tu diseño. Si tu estrategia no elimina el sesgo de selección, escribe «se asocia con», «se relaciona con», «se observa una diferencia de». Reserva «efecto», «impacto» e «influye» para cuando tengas un experimento o uno de los métodos de los capítulos 12 a 14. Tu jurado lo va a notar. Y si no lo nota, lo notará quien lea tu tesis después.

Segundo, antes de correr cualquier regresión, escribe tu lista de confusores. Pregúntate: ¿qué cosa podría mover al mismo tiempo mi variable explicativa y mi resultado? Si la tienes en los datos, contrólala. Si no la tienes, dilo en tus limitaciones y explica hacia dónde sesgaría tu resultado. Esa sola página hace más por tu investigación que diez pruebas de normalidad.

Y su hermano gemelo: controlar por todo

Hay un error opuesto, igual de frecuente. El tesista que ya aprendió que debe controlar mete en la regresión todas las variables que encuentra en el workfile, «por si acaso». Suena prudente, pero no lo es.

Algunas variables no deben controlarse nunca, porque son consecuencia del tratamiento que estudias. Supongamos que quieres medir cuánto sube el ingreso de un agricultor por recibir un crédito y controlas por cuántas hectáreas sembró esa campaña. Parece razonable, hasta que te preguntas para qué sirvió el crédito: en buena parte, para sembrar más.

Al comparar solo a agricultores que sembraron lo mismo, borras justo uno de los caminos por los que el crédito sube el ingreso, y además comparas grupos que ya no son comparables. Angrist y Pischke llaman a estas variables **malos controles** (Angrist & Pischke, 2009, pp. 64-68).

La regla práctica cabe en una línea: **controla por lo que ocurrió antes del tratamiento, no por lo que el tratamiento provocó**. Vuelve a tu dibujo de flechas. Si una variable recibe una flecha desde tu tratamiento, déjala fuera.

1.9 El error que cometen los investigadores en la redacción de artículos

El error típico aparece en el resumen: «Los resultados muestran que la educación tiene un impacto positivo en el ingreso». Tres páginas después, en las limitaciones, el autor admite que sus datos son observacionales. El revisor lee primero el resumen y luego busca el diseño, y cuando no lo encuentra, el artículo vuelve con la palabra «causalidad» subrayada.

Primero, los verbos cambian a mitad del texto. El método dice «asociación», los resultados dicen «efecto» y las conclusiones recomiendan una política como si el efecto estuviera probado. Un artículo tiene que usar el mismo verbo desde el título hasta la última línea.

Segundo, la limitación no es un conjuro. Escribir al final «no se puede afirmar causalidad» no corrige un resumen que la afirma. Si tu diseño no elimina el sesgo de selección, dilo en la introducción, donde el lector decide si sigue leyendo.

Cómo se escribe bien: «Quienes tomaron clases particulares obtienen 3,15 puntos más de nota, pero esa diferencia cae a 0,97 al comparar estudiantes con la misma motivación. Como los datos son observacionales, interpretamos la relación como una asociación y discutimos qué confusores podrían explicarla».

RESUMEN Y CONCLUSIONES

1. Una correlación puede venir de causalidad, causalidad inversa, un confusor o el azar. El software no distingue entre las cuatro.
2. El efecto causal compara dos resultados potenciales de la misma persona, y uno de ellos nunca se observa.
3. **Diferencia observada = efecto causal + sesgo de selección.** Si los grupos no eran comparables, la diferencia te engaña.
4. Una regresión simple es una comparación de promedios. Un valor p pequeño no corrige el sesgo.
5. Controlar ayuda si mides bien el confusor; con una proxy imperfecta queda sesgo, y controlar por consecuencias del tratamiento lo empeora.
6. El sorteo elimina el sesgo de selección. Cuando no puedes sortear, la econometría te da herramientas para acercarte, cada una con su supuesto.

En el capítulo 2 dejamos la simulación y trabajamos con datos reales: vas a descargar la ENAHO y los indicadores del Banco Mundial, y a dejarlos listos en EViews en una tarde.

EJERCICIOS RESUELTOS

Los tres están al final de `cap01.prg`, bajo el rótulo `EJERCICIOS RESUELTOS`. Córrelos después de los pasos 1 a 9 y en este orden, porque usan el workfile que ya tienes abierto. Intenta predecir el resultado antes de correr cada uno.

Ejercicio 1. ¿Y si la motivación no decidiera quién toma clases?

Supongamos que los estudiantes siguen eligiendo, pero ahora lo que decide es la distancia entre su casa y la academia, de 0 a 10 kilómetros. Cuanto más lejos viven, menos probable es que paguen clases. La distancia no tiene nada que ver con la nota ni con la motivación. ¿Cuánto dará ahora la comparación simple de promedios?

```
series distancia = 10*rnd
series clases1 = @rbinom(1, @logit(2 - 0.4*distancia))
series nota1 = 11 + 1*clases1 + 2*motivacion + ruido
nota1 = @recode(nota1<0, 0, @recode(nota1>20, 20, nota1))
show nota1.statby clases1
```

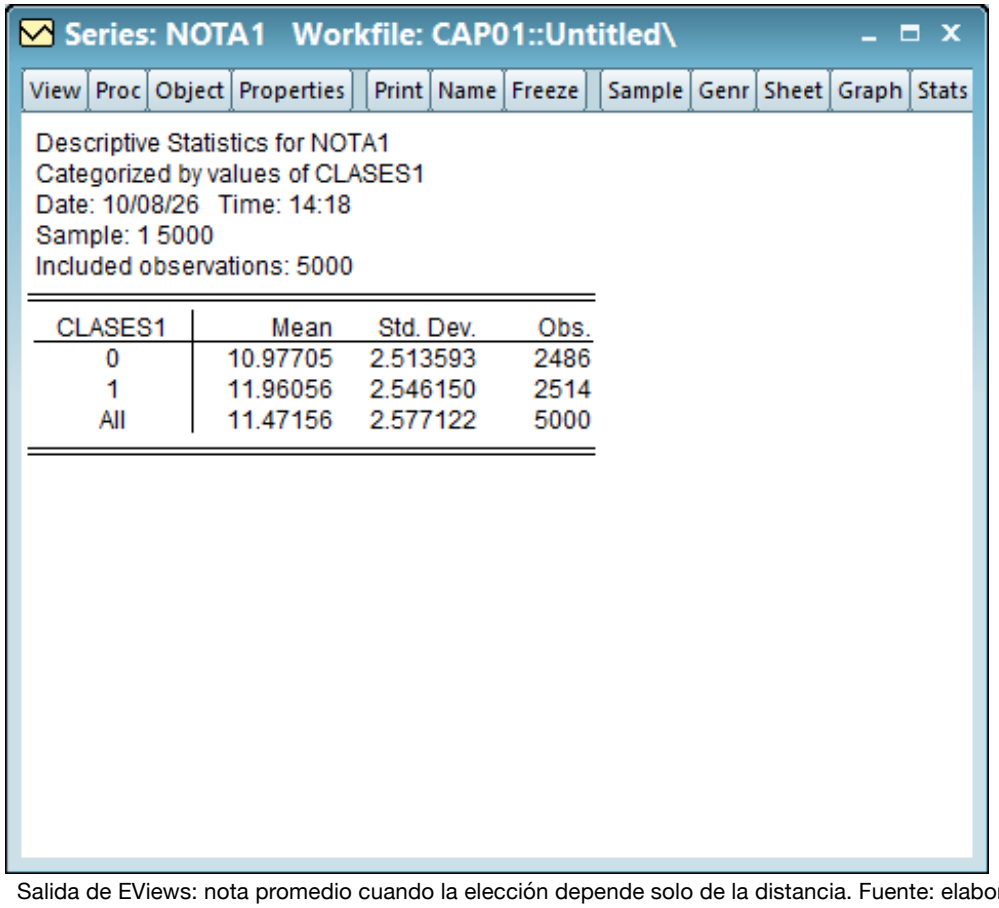


FIGURA 1.13 Salida de EViews: nota promedio cuando la elección depende solo de la distancia. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

```
show motivacion.statby clases1
```

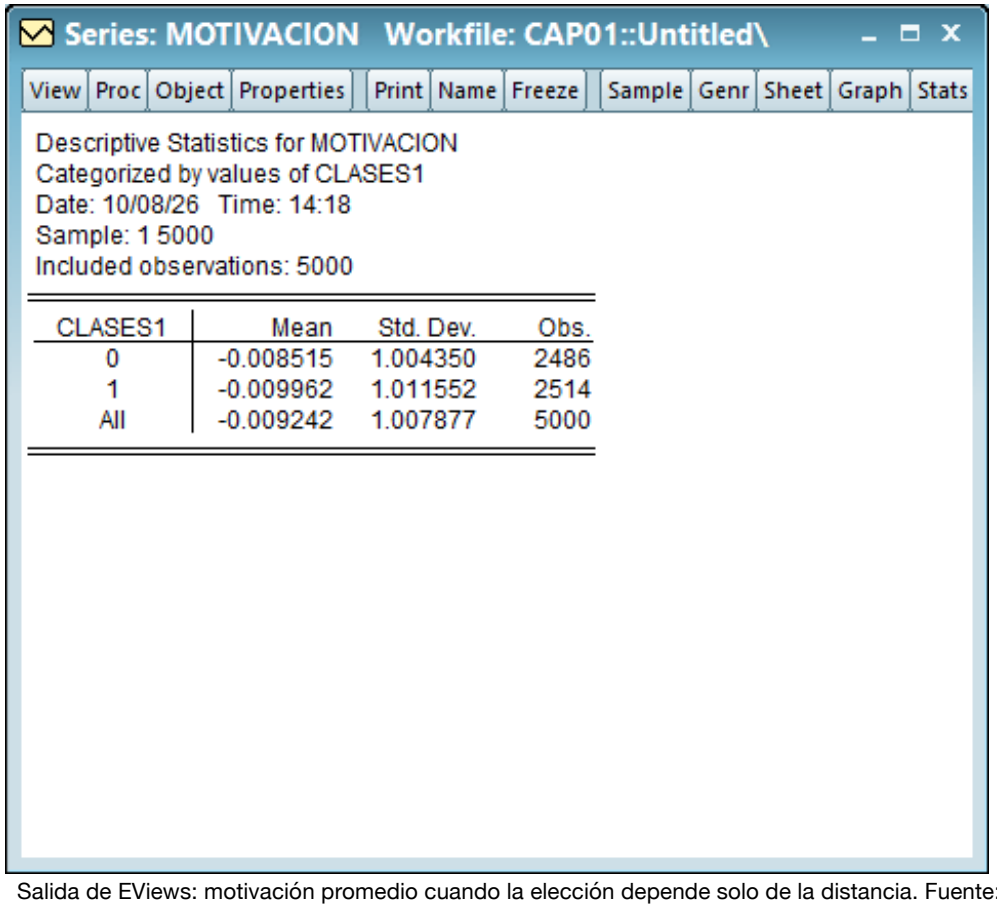


FIGURA 1.14 Salida de EViews: motivación promedio cuando la elección depende solo de la distancia. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Solución. La función `rnd` sortea un número entre 0 y 1, así que `10*rnd` reparte la distancia al azar entre 0 y 10 kilómetros. Los que tomaron clases promedian 11,96 y los que no, 10,98: una diferencia de 0,98, casi el 1 verdadero. Y eso que la motivación sigue subiendo la nota 2 puntos por unidad. La segunda tabla explica por qué: la motivación promedio es −0,010 en un grupo y −0,009 en el otro, prácticamente lo mismo.

Nadie lanzó una moneda y, aun así, la comparación acierta. Un confusor solo sesga tu comparación si mueve **las dos cosas**: el tratamiento y el resultado. Si lo que decide la elección no toca la nota, deja de ser un problema. Ojo: en tu investigación tendrías que defender ese supuesto con argumentos, porque los datos no pueden probarlo. Pensemos en la distancia real: ¿no viven más cerca de las academias las familias de barrios con más ingresos? Si es así, la distancia ya no sería inocente.

Ejercicio 2. Pon número al sesgo de selección.

Con los datos originales, calcula el sesgo de selección de dos formas. La primera, como la diferencia observada menos el efecto real. La segunda, como la diferencia de motivación entre grupos multiplicada por su efecto en la nota (2 puntos).

```
equation e1.ls nota c clases
scalar sesgo_a = e1.@coefs(2) - 1
equation e2.ls motivacion c clases
scalar sesgo_b = 2*e2.@coefs(2)
table(3,2) tsesgo
tsesgo.setwidth(1) 24
tsesgo.setwidth(2) 12
tsesgo(1,1) = "Forma de calcularlo"
tsesgo(1,2) = "Sesgo"
tsesgo(2,1) = "Observado - real"
tsesgo(2,2) = sesgo_a
tsesgo(3,1) = "Motivacion x 2"
tsesgo(3,2) = sesgo_b
show tsesgo
```

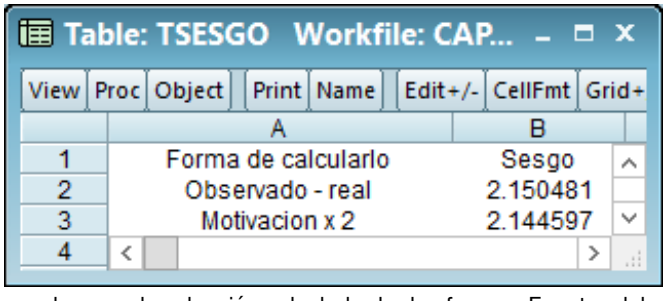


FIGURA 1.15 Salida de EViews: el sesgo de selección calculado de dos formas. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

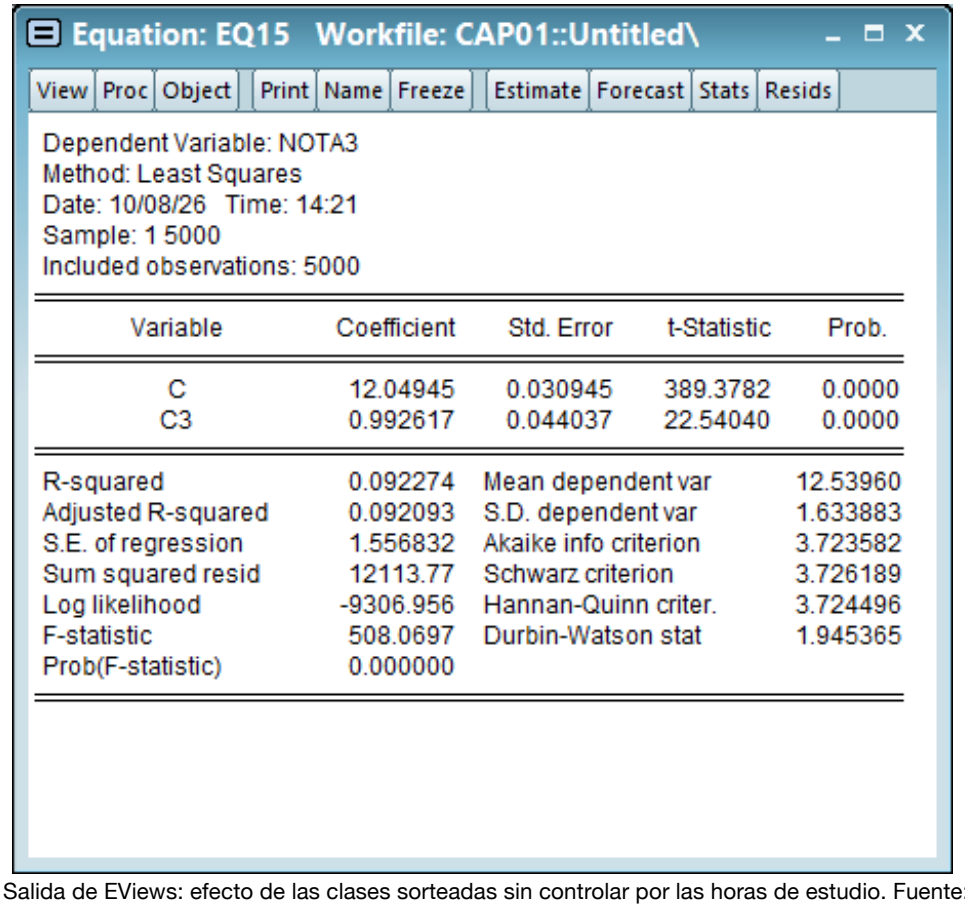
Solución. Aquí aparecen tres herramientas nuevas. Una ecuación estimada con `equation` sin `show` se calcula en silencio: queda guardada en el workfile, pero no abre ventana. `e1.@coefs(2)` es el segundo coeficiente de esa ecuación (el primero es la constante). Y `scalar` lo guarda en un número con nombre. Para verlos juntos armamos una **table** de 3 filas y 2 columnas, le damos ancho a cada columna con `setwidth` y llenamos cada celda con `tsesgo(fila, columna)`.

Las dos cuentas casi coinciden: 2,150 y 2,145. Es la ecuación de la trampa comprobada con números: 3,15 observados = 1 de efecto + unos 2,15 de sesgo. La pequeña diferencia entre ambas cuentas viene del ruido y del recorte de la nota entre 0 y 20. En tus datos reales nunca podrás hacer esta cuenta, porque no conoces el efecto real. Por eso conviene entrenar la intuición aquí, donde sí lo conoces.

Ejercicio 3. Un mal control en acción.

Ahora las clases se sortean, y además hacen que el estudiante estudie 3 horas más por semana. Cada hora sube la nota 0,2 puntos, y las clases suman 0,4 puntos directos. ¿Qué pasa si «controlas» por las horas?

```
series c3 = @rbinom(1, 0.5)
series horas3 = 5 + 3*c3 + 2*nrnd
series nota3 = 11 + 0.4*c3 + 0.2*horas3 + 1.5*nrnd
nota3 = @recode(nota3<0, 0, @recode(nota3>20, 20, nota3))
equation eq15.ls nota3 c c3
show eq15
```



Equation: EQ15 Workfile: CAP01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

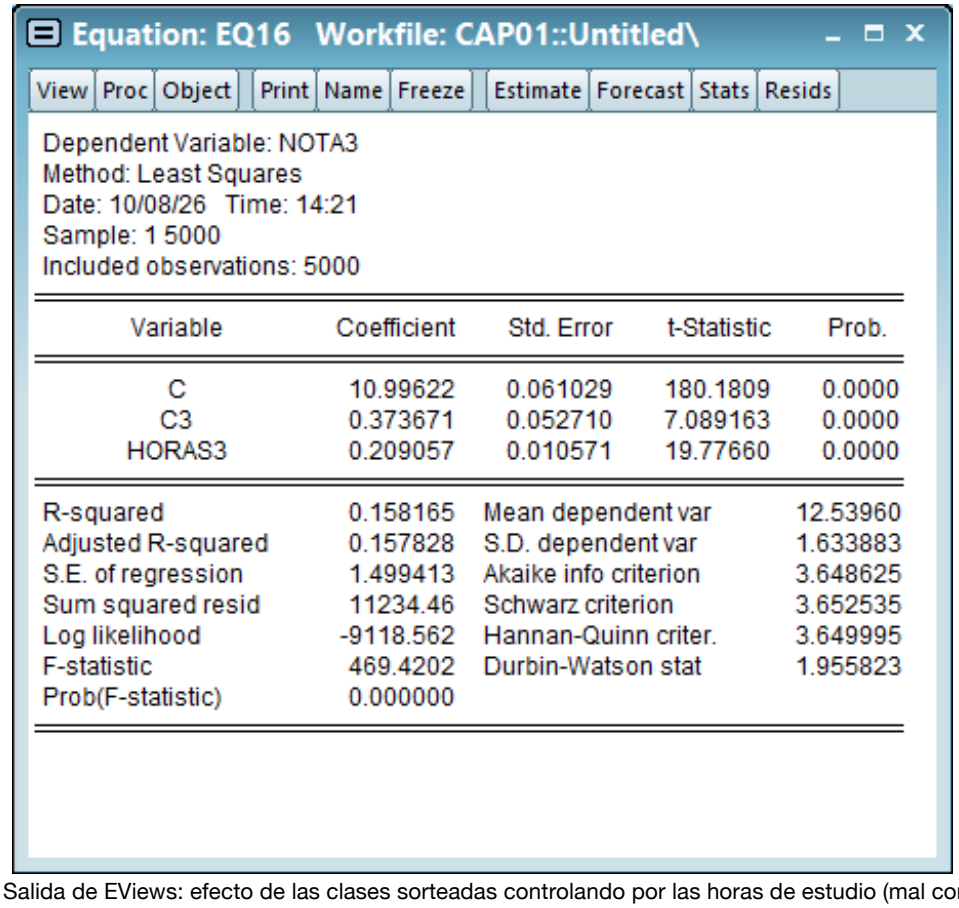
Dependent Variable: NOTA3
Method: Least Squares
Date: 10/08/26 Time: 14:21
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.04945	0.030945	389.3782	0.0000
C3	0.992617	0.044037	22.54040	0.0000

R-squared	0.092274	Mean dependent var	12.53960
Adjusted R-squared	0.092093	S.D. dependent var	1.633883
S.E. of regression	1.556832	Akaike info criterion	3.723582
Sum squared resid	12113.77	Schwarz criterion	3.726189
Log likelihood	-9306.956	Hannan-Quinn criter.	3.724496
F-statistic	508.0697	Durbin-Watson stat	1.945365
Prob(F-statistic)	0.000000		

FIGURA 1.16 Salida de EViews: efecto de las clases sorteadas sin controlar por las horas de estudio. Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

```
equation eq16.ls nota3 c c3 horas3
show eq16
```



Equation: EQ16 Workfile: CAP01::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: NOTA3
Method: Least Squares
Date: 10/08/26 Time: 14:21
Sample: 1 5000
Included observations: 5000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.99622	0.061029	180.1809	0.0000
C3	0.373671	0.052710	7.089163	0.0000
HORAS3	0.209057	0.010571	19.77660	0.0000

R-squared	0.158165	Mean dependent var	12.53960
Adjusted R-squared	0.157828	S.D. dependent var	1.633883
S.E. of regression	1.499413	Akaike info criterion	3.648625
Sum squared resid	11234.46	Schwarz criterion	3.652535
Log likelihood	-9118.562	Hannan-Quinn criter.	3.649995
F-statistic	469.4202	Durbin-Watson stat	1.955823
Prob(F-statistic)	0.000000		

FIGURA 1.17 Salida de EViews: efecto de las clases sorteadas controlando por las horas de estudio (mal control). Fuente: elaboración propia con EViews 14 y datos simulados (cap01.prg).

Solución. El efecto total de las clases es 1 punto: 0,4 directos más 0,6 que llegan porque el estudiante estudia más (3 horas × 0,2). La regresión sin controles lo estima bien, en 0,99, con un error estándar de 0,04. Al controlar por las horas, el coeficiente cae a 0,37. Le quitaste a las clases el camino por el que producen buena parte de su efecto. Las horas recibían una flecha desde el tratamiento, así que eran un mal control.

Para tu tesis o artículo

1. **Escribe tu pregunta como un «¿qué habría pasado si…?».** Por ejemplo: «¿Cuánto habría ganado este agricultor sin el crédito?». Si no puedes escribirla así, quizás tu investigación es descriptiva. Eso también está bien: solo cambia los verbos.

2. **Haz tu lista de tres confusores.** Para cada uno, anota si está en tus datos y hacia dónde movería tu resultado si lo omites.

3. **Revisa el título de tu tesis o de tu artículo.** Si dice «influencia», «efecto» o «impacto», pregúntate qué parte de tu diseño elimina el sesgo de selección.

Lo que te va a preguntar el jurado: «¿Cómo sabe que esa relación es causal y no una correlación?». Después de este capítulo tienes cómo responder: nombra el confusor que más te preocupa y explica qué hiciste con él.

Si es un artículo: En la revista, la pregunta del jurado la hace el revisor, y la hace antes: si el título o el resumen dicen «efecto» o «impacto», buscará en los primeros párrafos de la introducción qué diseño elimina el

sesgo de selección. Si tu diseño no lo elimina, cambia desde el título los verbos a «se asocia con»: es más barato que un rechazo.

Referencias

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *The New England Journal of Medicine, 330*(15), 1029-1035. https://doi.org/10.1056/NEJM199404143301501

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review, 92*(5), 1535-1558. https://doi.org/10.1257/000282802762024629

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691120355/mostly-harmless-econometrics

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691152844/mastering-metrics

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://mixtape.scunning.com/

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. *The New England Journal of Medicine, 367*(16), 1562-1564. https://doi.org/10.1056/NEJMon1211064