Econometría aplicada con Stata 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 Stata para la investigación
Guía práctica paso a paso para tesis, artículos científicos y análisis de datos
Serie Econometría sin miedo
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.° 2026-12544
ISBN: 978-612-03-3256-6
Libro digital disponible en: Amazon Kindle, Apple Books y Google Play Books.
Ninguna parte de esta obra puede reproducirse sin autorización escrita del autor, salvo citas breves con fines de reseña o docencia.
El código del libro se ofrece 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.
Stata es una marca registrada de StataCorp LLC. Este libro es una obra independiente: no está patrocinado, autorizado ni avalado por StataCorp. Python, R, EViews 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. Pensar como econometrista 1
Capítulo 1 ¿Por qué tu regresión miente? Correlación y causalidad 2
Capítulo 2 Tus datos listos en Stata en una tarde: de la ENAHO y el Banco Mundial a una base limpia 19
Capítulo 3 Prueba de hipótesis e intervalos de confianza: la estadística que tu investigación necesita 49
Parte Dos. La regresión, bien hecha 75
Capítulo 4 Regresión lineal simple paso a paso 76
Capítulo 5 Regresión lineal múltiple y el sesgo de variable omitida 104
Capítulo 6 Cómo interpretar una regresión sin pánico 129
Capítulo 7 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación 152
Capítulo 8 Regresión logística: modelos logit y probit 187
Parte Tres. Datos que se mueven en el tiempo 215
Capítulo 9 Datos de panel: efectos fijos vs. aleatorios y test de Hausman 216
Capítulo 10 Series de tiempo y modelo ARIMA paso a paso 246
Capítulo 11 Pronósticos con modelos VAR y cointegración 279
Parte Cuatro. Evaluación de impacto 313
Capítulo 12 Evaluación de impacto I: diferencias en diferencias 314
Capítulo 13 Evaluación de impacto II: variables instrumentales 344
Capítulo 14 Evaluación de impacto III: regresión discontinua y propensity score matching 371
Parte Cinco. Del resultado a la tesis y al artículo 398
Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico 399
APÉNDICE A. EQUIVALENCIAS STATA · PYTHON · R · EVIEWS 425
APÉNDICE B. BASES DE DATOS DEL LIBRO 426

Contenido

ACERCA DEL AUTOR iv

ÍNDICE DE TABLAS xi

PREFACIO: CÓMO USAR ESTE LIBRO xii

Parte Uno. Pensar como econometrista 1

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

Lo que vas a lograr 3

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 6

1.8 El error que comete casi todo tesista 13

1.9 El error que cometen los investigadores en la redacción de artículos 13

RESUMEN Y CONCLUSIONES 14

EJERCICIOS RESUELTOS 14

Para tu tesis o artículo 17

Referencias 17

Capítulo 2 Tus datos listos en Stata en una tarde: de la ENAHO y el Banco Mundial a una base limpia 19

Lo que vas a lograr 19

2.1 ¿Qué es la ENAHO y por qué te conviene? 20

2.2 El número que convierte tu muestra en un país 20

2.3 ¿Microdatos o datos por país? 21

2.4 Por qué un do-file y no clics 21

2.5 Manos a la obra 21

2.6 El error que arruina más investigaciones de lo que crees 43

2.7 El error que cometen los investigadores en la redacción de artículos 44

RESUMEN Y CONCLUSIONES 44

EJERCICIOS RESUELTOS 45

Para tu tesis o artículo 47

Referencias 47

Capítulo 3 Prueba de hipótesis e intervalos de confianza: la estadística que tu investigación necesita 49

Lo que vas a lograr 49

3.1 La pregunta: ¿las mujeres ganan menos? 50

3.2 Una muestra es solo una de muchas posibles 50

3.3 El intervalo de confianza: un rango honesto 50

3.4 La prueba de hipótesis funciona como un juicio 51

3.5 Lo que el valor p no te dice 52

3.6 Dos maneras de equivocarte: errores tipo I y II 52

3.7 Significativo no es lo mismo que importante 52

3.8 Manos a la obra 53

3.9 El error que comete casi todo el mundo con el valor p 68

3.10 El error que cometen los investigadores en la redacción de artículos 68

RESUMEN Y CONCLUSIONES 69

EJERCICIOS RESUELTOS 69

Para tu tesis o artículo 73

Referencias 74

Parte Dos. La regresión, bien hecha 75

Capítulo 4 Regresión lineal simple paso a paso 76

Lo que vas a lograr 76

4.1 Una recta para 45 392 personas 77

4.2 La recta que comete menos errores al cuadrado ... 77
4.3 ¿Por qué el logaritmo del ingreso? ... 78
4.4 Cómo leer la pendiente cuando la y está en logaritmos ... 78
4.5 El R^2: cuánto de la nube explica la recta ... 79
4.6 Valores ajustados y residuos: las dos mitades de cada dato ... 79
4.7 Manos a la obra ... 80
4.8 El error que comete casi todo el mundo con la pendiente ... 96
4.9 El error que cometen los investigadores en la redacción de artículos ... 97
RESUMEN Y CONCLUSIONES ... 97
EJERCICIOS RESUELTOS ... 98
Para tu tesis o artículo ... 102
Referencias ... 103
Capítulo 5 Regresión lineal múltiple y el sesgo de variable omitida ... 104
Lo que vas a lograr ... 104
5.1 Del «en promedio» al «entre personas parecidas» ... 105
5.2 La experiencia potencial y la ecuación de Mincer ... 105
5.3 El sesgo de variable omitida, en una fórmula ... 106
5.4 Variables dicotómicas y la trampa de las dummies ... 106
5.5 ¿El R^2 sirve para elegir modelos? ... 107
5.6 Manos a la obra ... 107
5.7 El error que comete casi todo tesista con los controles ... 121
5.8 El error que cometen los investigadores en la redacción de artículos ... 122
RESUMEN Y CONCLUSIONES ... 122
EJERCICIOS RESUELTOS ... 122
Para tu tesis o artículo ... 126
Referencias ... 127
Capítulo 6 Cómo interpretar una regresión sin pánico ... 129
Lo que vas a lograr ... 129
6.1 Tres zonas, tres preguntas ... 130
6.2 ¿Cómo se lee un coeficiente? Depende de dónde esté el logaritmo ... 130
6.3 Una hipótesis a la vez, o todas juntas ... 131
6.4 El coeficiente no cambia; lo que cambia es tu confianza ... 131
6.5 ¿Ponderar o no ponderar? ... 131
6.6 Beta, AIC y BIC: comparar sin hacer trampa ... 132
6.7 Manos a la obra ... 132
6.8 El error que comete casi todo tesista al leer su salida ... 146
6.9 El error que cometen los investigadores en la redacción de artículos ... 147
RESUMEN Y CONCLUSIONES ... 147
EJERCICIOS RESUELTOS ... 147
Para tu tesis o artículo ... 150
Referencias ... 151
Capítulo 7 Supuestos de la regresión: heterocedasticidad, multicolinealidad y autocorrelación ... 152
Lo que vas a lograr ... 152
7.1 Dos tipos de supuestos, dos tipos de daño ... 153
7.2 Heterocedasticidad: cuando el ruido crece con el nivel ... 153
7.3 Multicolinealidad: dos variables que dicen lo mismo ... 154
7.4 Autocorrelación: el error de hoy se parece al de ayer ... 154
7.5 Normalidad: el supuesto que más asusta y menos importa ... 155
7.6 Especificación: ¿falta algo en tu modelo? ... 155
7.7 Manos a la obra ... 155
7.8 El error que comete casi todo tesista con los supuestos ... 177

7.9 El error que cometen los investigadores en la redacción de artículos 178
RESUMEN Y CONCLUSIONES 178
EJERCICIOS RESUELTOS 178
Para tu tesis o artículo 184
Referencias 185
Capítulo 8 Regresión logística: modelos logit y probit 187
Lo que vas a lograr 187
8.1 ¿Qué pasa si tu y solo vale 0 o 1? 188
8.2 La curva en S: logit y probit 188
8.3 Por qué no puedes leer un logit como un MCO 189
8.4 ¿Qué tan bueno es tu modelo? 189
8.5 Manos a la obra 190
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 210
RESUMEN Y CONCLUSIONES 210
EJERCICIOS RESUELTOS 210
Para tu tesis o artículo 213
Referencias 214
Parte Tres. Datos que se mueven en el tiempo 215
Capítulo 9 Datos de panel: efectos fijos vs. aleatorios y test de Hausman 216
Lo que vas a lograr 216
9.1 ¿Qué es un panel y por qué vale tanto? 217
9.2 Efectos fijos: comparar a cada unidad consigo misma 217
9.3 Los efectos de tiempo: lo que le pasa a todos a la vez 218
9.4 Efectos aleatorios: el punto medio 218
9.5 El test de Hausman: qué compara de verdad 218
9.6 Manos a la obra 219
9.7 El error que comete casi todo tesista con el test de Hausman 240
9.8 El error que cometen los investigadores en la redacción de artículos 240
RESUMEN Y CONCLUSIONES 241
EJERCICIOS RESUELTOS 241
Para tu tesis o artículo 244
Referencias 245
Capítulo 10 Series de tiempo y modelo ARIMA paso a paso 246
Lo que vas a lograr 246
10.1 ¿Qué tiene de especial una serie de tiempo? 246
10.2 Estacionariedad: la condición para que el pasado enseñe algo 247
10.3 Las piezas del ARIMA: AR, I y MA 248
10.4 Box y Jenkins: identificar, estimar, diagnosticar, pronosticar 248
10.5 ¿Cómo se juzga un pronóstico? 249
10.6 Manos a la obra 249
10.7 El error que comete casi todo tesista con ARIMA 270
10.8 El error que cometen los investigadores en la redacción de artículos 271
RESUMEN Y CONCLUSIONES 271
EJERCICIOS RESUELTOS 271
Para tu tesis o artículo 276
Referencias 277
Capítulo 11 Pronósticos con modelos VAR y cointegración 279
Lo que vas a lograr 279
11.1 ¿Qué es un VAR y por qué se volvió estándar? 279
11.2 Los supuestos del VAR: rezagos, estabilidad y residuos limpios 280
11.3 Causalidad de Granger: predecir no es causar 280
11.4 Impulso-respuesta y descomposición de varianza 281

11.5 Paseos aleatorios, regresión espuria y cointegración 281
11.6 Dos maneras de probar la cointegración.. 282
11.7 Manos a la obra..... 282
11.8 El error que comete casi todo tesista con VAR y cointegración 305
11.9 El error que cometen los investigadores en la redacción de artículos 306
RESUMEN Y CONCLUSIONES..... 306
EJERCICIOS RESUELTOS..... 306
Para tu tesis o artículo 310
Referencias 311
Parte Cuatro. Evaluación de impacto 313
Capítulo 12 Evaluación de impacto I: diferencias en diferencias..... 314
Lo que vas a lograr..... 314
12.1 Dos comparaciones que mienten 315
12.2 La idea: restar dos veces..... 315
12.3 El supuesto que lo sostiene todo: tendencias paralelas 316
12.4 Los errores estándar: la lección del capítulo 6 316
12.5 Muchos años: efectos fijos bidireccionales y estudio de eventos..... 317
12.6 Cuidado con la adopción escalonada..... 317
12.7 Manos a la obra..... 317
12.8 El error que comete casi todo tesista con diferencias en diferencias 338
12.9 El error que cometen los investigadores en la redacción de artículos 338
RESUMEN Y CONCLUSIONES..... 339
EJERCICIOS RESUELTOS..... 339
Para tu tesis o artículo 342
Referencias 342
Capítulo 13 Evaluación de impacto II: variables instrumentales..... 344
Lo que vas a lograr..... 344
13.1 ¿Por qué tu variable explicativa está contaminada? 345
13.2 El instrumento: un empujón que no toca el resultado..... 345
13.3 La cuenta más simple: dividir dos diferencias..... 346
13.4 Mínimos cuadrados en dos etapas 346
13.5 ¿A quién se aplica el efecto? Los cumplidores 347
13.6 Instrumentos débiles: cuando el empujón es un susurro 347
13.7 Dos pruebas útiles y sus límites..... 348
13.8 Manos a la obra..... 348
13.9 El error que comete casi todo tesista con variables instrumentales 364
13.10 El error que cometen los investigadores en la redacción de artículos 364
RESUMEN Y CONCLUSIONES..... 364
EJERCICIOS RESUELTOS..... 365
Para tu tesis o artículo 369
Referencias 370
Capítulo 14 Evaluación de impacto III: regresión discontinua y propensity score matching..... 371
Lo que vas a lograr..... 371
14.1 La idea: una regla con un umbral 372
14.2 El supuesto: nada más salta en el umbral 372
14.3 Nítida o difusa 372
14.4 Ancho de banda y polinomio: cuán cerca y con qué curva..... 373
14.5 La trampa: cuando la gente se mueve de lado 373
14.6 Manos a la obra: la ventaja de ganar una elección..... 374
14.7 Cuando no hay umbral: emparejar a los parecidos..... 383
14.8 Manos a la obra: tabaco y peso al nacer .. 384
14.9 ¿Cuál método elijo? 390
14.10 El error que comete casi todo tesista con el emparejamiento 391

14.11 El error que cometen los investigadores en la redacción de artículos 391
RESUMEN Y CONCLUSIONES..... 391
EJERCICIOS RESUELTOS..... 392
Para tu tesis o artículo 396
Referencias 397
Parte Cinco. Del resultado a la tesis y al artículo .. 398
Capítulo 15 Cómo redactar los resultados de tu tesis o artículo científico 399
Lo que vas a lograr 399
15.1 ¿Qué debe responder tu capítulo de resultados? 400
15.2 La anatomía de una tabla de regresión 400
15.3 El párrafo de resultados: magnitud, incertidumbre y contexto 401
15.4 Las palabras que prometen causalidad 401
15.5 Limitaciones y robustez: dilo tú antes que el jurado 402
15.6 Tu investigación debe poder correr otra vez 402
15.7 La sustentación: las preguntas que sí vas a escuchar 402
15.8 ¿Y si es un artículo científico? 403
15.9 Manos a la obra..... 403
15.10 El error que comete casi todo tesista: narrar estrellas 418
15.11 El error que cometen los investigadores en la redacción de artículos 419
RESUMEN Y CONCLUSIONES..... 419
EJERCICIOS RESUELTOS..... 420
Para tu tesis o artículo 422
Referencias 423
APÉNDICE A. EQUIVALENCIAS STATA · PYTHON · R · EVIEWS..... 425
APÉNDICE B. BASES DE DATOS DEL LIBRO426

Capítulo 1

¿Por qué tu regresión miente? Correlación y causalidad

En 1991, dos investigadores de Boston reunieron casi todo lo que la medicina sabía sobre una pregunta que importaba a muchísimas mujeres. ¿Tomar estrógenos después de la menopausia protege el corazón? Meir Stampfer y Graham Colditz juntaron estudios de hospitales, estudios de población y dieciséis estudios que siguieron a mujeres durante años.

Quince de esos dieciséis encontraron menos riesgo entre las que tomaban estrógenos. El resultado parecía contundente. Las mujeres que tomaban estrógenos tenían casi la mitad del riesgo de sufrir una enfermedad coronaria: un riesgo relativo de 0,56. Los autores fueron más lejos. Escribieron que la evidencia apoyaba con fuerza un efecto protector y que era poco probable que se explicara por otros factores (Stampfer & Colditz, 1991).

Con evidencia así, ¿quién iba a dudar? Los datos lo decían.

Entonces alguien hizo lo que nadie había hecho.

Entre 1993 y 1998, la Women's Health Initiative reclutó a 16 608 mujeres de 50 a 79 años en 40 centros de Estados Unidos. Esta vez no se dejó que cada mujer decidiera. Se sorteó quién recibía hormonas y quién recibía una pastilla idéntica sin nada activo.

El 31 de mayo de 2002, el comité que vigilaba el estudio recomendó detenerlo antes de tiempo. Las mujeres que recibieron hormonas tuvieron **más** enfermedad coronaria, no menos: un cociente de riesgo de 1,29 (Writing Group for the Women's Health Initiative Investigators [WHI], 2002).

Decenas de estudios decían «protege». El experimento dijo «daña». ¿Cómo pudo equivocarse tanta evidencia a la vez?

La respuesta no estaba en las hormonas. Estaba en las mujeres. Las que elegían tomar hormonas no eran iguales a las que no lo hacían.

Un estudio con 4 286 mujeres de 60 a 79 años encontró que quienes habían vivido en peores condiciones socioeconómicas a lo largo de su vida, incluso desde la infancia, usaban menos terapia hormonal. Esa diferencia no desaparecía al tomar en cuenta los factores de riesgo de la edad adulta (Lawlor et al., 2004). Dicho de otro modo: las usuarias de hormonas ya partían con ventaja, y los estudios le atribuyeron esa ventaja a 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 Stata, 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.

Robert Matthews lo mostró con un ejemplo que se volvió famoso. Tomó datos de países europeos y encontró una correlación altamente significativa entre la cantidad de cigüeñas y la tasa de natalidad. El título de su artículo lo dice todo: *Las cigüeñas traen a los bebés (p = 0,008)*. Su conclusión también: las cigüeñas no traen bebés, pero leer correlaciones y valores p sin pensar sí trae conclusiones poco confiables (Matthews, 2000).

¿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, estudiar más horas sube tu nota, y una sequía reduce la cosecha de papa.

2. Causalidad inversa. Y provoca X. Imagina que encuentras que los distritos con más comisarías tienen más denuncias. ¿Las comisarías generan delitos? Lo más probable es lo contrario: donde hay más delitos, se ponen más comisarías. Y donde hay una comisaría cerca, además, es más fácil denunciar.

3. Una tercera variable. Algo distinto mueve a las dos. En el caso de las cigüeñas, una candidata obvia es el tamaño del país: más territorio, más nidos y más personas. Las mujeres con mejor posición socioeconómica usaban más hormonas y tenían mejor salud. A esa tercera variable los econometristas le dicen **variable de confusión** o **factor de confusión**.

4. El azar. Por ejemplo, el precio del cobre y el número de goles de tu equipo en una temporada pueden moverse juntos 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 Lucía prepara su tesis sobre el rendimiento escolar en Arequipa. 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 Lucía 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:

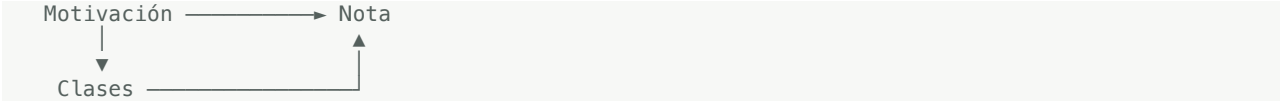
- La nota que sacaría **con** clases particulares.
- La nota que sacaría **sin** clases particulares.

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 Stata, 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 Lucía, dibujado, se ve así:



La flecha de abajo, de Clases a Nota, es la que Lucía 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 Lucía 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 Lucía 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 de las hormonas. 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 Lucía 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 la Women's Health Initiative, y por eso encontró lo que décadas de estudios observacionales no vieron.

América Latina tiene su propio ejemplo famoso.

A fines de los años noventa, México puso en marcha Progresa, un programa que daba dinero a madres pobres de zonas rurales a condición de que sus hijos fueran a la escuela. El programa no llegó a todas las comunidades a la vez. De 495 localidades pobres censadas, 314 se eligieron **al azar** para recibirlo primero, y las otras 181 esperaron y sirvieron de comparación (Schultz, 2001, 2004).

Como el azar decidió el orden, la diferencia entre ambos grupos se podía leer como efecto del programa. Entre los niños que habían terminado sexto grado, la matrícula en secundaria subió 11,1 puntos porcentuales: de 58 % a cerca de 69 % (Schultz, 2001).

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 Stata y escribe `doedit` en la ventana Command: se abre el Do-file Editor, el cuaderno donde vas a guardar tus comandos. Copia los bloques en orden. Para correr un bloque, selecciona sus líneas y pulsa el botón *Execute (do)* o Ctrl+D en Windows. El do-file completo está en el archivo `cap01.do` del repositorio del libro; el prefacio te explica cómo descargarlo. No necesitas instalar ningún paquete adicional: todo lo que usamos viene con Stata 16.

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 Lucía. 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

```
clear all
set obs 5000
set seed 2026
generate motivacion = rnormal(0, 1)
```

```
. clear all

. set obs 5000
number of observations (_N) was 0, now 5,000

. set seed 2026

. generate motivacion = rnormal(0, 1)

.
```

FIGURA 1.1 Salida de Stata: creación de 5 000 estudiantes con semilla fija y motivación simulada. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

¿Qué hicimos? `clear all` limpia la memoria y `set obs 5000` crea 5 000 filas vacías, una por estudiante. `set seed 2026` fija la semilla del generador de números aleatorios: gracias a ella, tú obtendrás exactamente los mismos números que yo. Por último, `rnormal(0, 1)` le da a cada estudiante una motivación al azar, con promedio 0 y desviación estándar 1. Nadie la mide 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

```
generate clases = rbinomial(1, invlogit(1.5*motivacion))
tabulate clases
```

```
. generate clases = rbinomial(1, invlogit(1.5*motivacion))

. tabulate clases
```

clases	Freq.	Percent	Cum.
0	2,489	49.78	49.78
1	2,511	50.22	100.00
Total	5,000	100.00	

```
.
```

FIGURA 1.2 Salida de Stata: estudiantes con y sin clases particulares en la simulación. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Esta línea es el corazón del problema. La función `invlogit()` 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 `rbinomial(1, p)` lanza una moneda cargada con esa probabilidad y anota 1 si el estudiante pagó clases y 0 si no. La tabla muestra que 2 511 estudiantes (50,22 %) tomaron clases y 2 489 no.

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

```
generate ruido = rnormal(0, 1.5)
generate nota = clip(11 + 1*clases + 2*motivacion + ruido, 0, 20)
summarize nota motivacion ruido
```

```
. generate ruido = rnormal(0, 1.5)

. generate nota = clip(11 + 1*clases + 2*motivacion + ruido, 0, 20)

. summarize nota motivacion ruido
```

Variable	Obs	Mean	Std. Dev.	Min	Max
nota	5,000	11.51737	2.753705	1.966217	20
motivacion	5,000	-.01212	1.005548	-3.561762	3.53755
ruido	5,000	.0396956	1.503762	-6.092394	5.979718

```
.
```

FIGURA 1.3 Salida de Stata: estadísticas descriptivas de la nota, la motivación y el ruido. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

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 función `clip(x, 0, 20)` recorta la nota para que no baje de 0 ni pase de 20. El resumen muestra una nota promedio de 11,52 y una motivación promedio prácticamente nula (−0,012).

Paso 4. Primer intento: comparar promedios

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

```
tabstat nota, by(clases) statistics(mean sd n)
```

```
. tabstat nota, by(clases) statistics(mean sd n)

Summary for variables: nota
     by categories of: clases

  clases |      mean        sd         N
---------+------------------------------
       0 |  9.974979  2.304854      2489
       1 |  13.04625  2.266952      2511
---------+------------------------------
   Total |  11.51737  2.753705      5000
------------------------------------------

.
```

FIGURA 1.4 Salida de Stata: nota promedio de quienes tomaron clases y de quienes no. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Los que no tomaron clases promedian 9,97; los que sí, 13,05. La diferencia es de 3,07 puntos. Sabemos que la verdad es uno. La comparación simple triplicó el efecto.

Un detalle de lectura: Stata escribe los decimales con punto, como en inglés. En el texto del libro usamos la coma, así que 13.04625 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.

```
regress nota clases
```

```
. regress nota clases

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(1, 4998)      =   2256.45
       Model |  11790.676         1  11790.676     Prob > F        =    0.0000
    Residual |  26116.2084    4,998  5.22533182   R-squared       =    0.3110
-------------+----------------------------------   Adj R-squared   =    0.3109
       Total |  37906.8845    4,999  7.58289347   Root MSE        =    2.2859

------------------------------------------------------------------------------
        nota |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clases |   3.071274   .0646556    47.50   0.000     2.944521    3.198027
       _cons |   9.974979   .0458189   217.70   0.000     9.885154    10.0648
------------------------------------------------------------------------------

.
```

FIGURA 1.5 Salida de Stata: regresión de la nota sobre las clases particulares, sin controles. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

El coeficiente de `clases` es 3,07. El valor p (columna `P>|t|`) es 0,000. El intervalo de confianza al 95 % va de 2,94 a 3,20: ni siquiera se acerca al 1 verdadero. Es la tabla más convincente que verás en tu vida, y está equivocada.

Mira además la constante, `_cons`: 9,97. Es exactamente el promedio de los que no tomaron clases en el paso 4. Y 9,97 más 3,07 da 13,04: el promedio de los que sí las tomaron, salvo el redondeo.

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 Lucía no tiene: la motivación de cada estudiante. Metámosla en la regresión.

```
regress nota clases motivacion
```

```
. regress nota clases motivacion

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(2, 4997)      =   5887.35
       Model |  26612.8168         2  13306.4084   Prob > F        =    0.0000
    Residual |  11294.0676     4,997  2.26016963   R-squared       =    0.7021
-------------+----------------------------------   Adj R-squared   =    0.7019
       Total |  37906.8845     4,999  7.58289347   Root MSE        =    1.5034

------------------------------------------------------------------------------
        nota |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clases |   .9554947   .0499077    19.15   0.000     .8576537    1.053336
  motivacion |   2.009827   .0248184    80.98   0.000     1.961172    2.058482
       _cons |   11.06188   .0329879   335.33   0.000     10.99721    11.12655
------------------------------------------------------------------------------

.
```

FIGURA 1.6 Salida de Stata: regresión de la nota sobre las clases, controlando por la motivación. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Ahora el coeficiente de `clases` es 0,96, y el intervalo (0,86 a 1,05) contiene el 1 verdadero. 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`: 2,01. 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.

```
generate clases_azar = rbinomial(1, 0.5)
generate nota_azar = clip(11 + 1*clases_azar + 2*motivacion + ruido, 0, 20)
ttest motivacion, by(clases_azar)
```

```
. generate clases_azar = rbinomial(1, 0.5)

. generate nota_azar = clip(11 + 1*clases_azar + 2*motivacion + ruido, 0, 20)

. ttest motivacion, by(clases_azar)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   2,490   -.0109411    .0202772    1.011828   -.0507029    .0288208
       1 |   2,510   -.0132895    .0199497    .9994794   -.0524092    .0258301
---------+--------------------------------------------------------------------
combined |   5,000     -.01212    .0142206    1.005548   -.0399986    .0157586
---------+--------------------------------------------------------------------
    diff |            .0023485    .0284443               -.0534148    .0581117
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.0826
Ho: diff = 0                                     degrees of freedom =     4998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5329         Pr(|T| > |t|) = 0.9342          Pr(T > t) = 0.4671

.
```

FIGURA 1.7 Salida de Stata: prueba de balance de la motivación entre los grupos sorteados. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Antes de estimar nada, revisamos el **balance**: ¿los dos grupos sorteados tienen la misma motivación? La prueba `ttest` compara los promedios. En la columna `Mean`, el grupo sin clases tiene −0,011 y el grupo con clases, −0,013. La diferencia (`diff`) es de apenas 0,002, y el valor p de la prueba bilateral, en `Pr(|T| > |t|)`, es 0,93. No hay ninguna diferencia que merezca ese nombre.

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

```
regress nota_azar clases_azar
```

```
. regress nota_azar clases_azar

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(1, 4998)      =    194.32
       Model |  1224.63517         1  1224.63517   Prob > F        =    0.0000
    Residual |  31498.5173     4,998  6.30222436   R-squared       =    0.0374
-------------+----------------------------------   Adj R-squared   =    0.0372
       Total |  32723.1525     4,999  6.54593969   Root MSE        =    2.5104

------------------------------------------------------------------------------
   nota_azar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
clases_azar |     .98981   .0710061    13.94   0.000      .850607    1.129013
       _cons |   11.02052   .0503092   219.06   0.000      10.9219    11.11915
------------------------------------------------------------------------------

.
```

FIGURA 1.8 Salida de Stata: regresión simple de la nota sobre las clases sorteadas. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Sin ningún control, el coeficiente es 0,99, con un intervalo de 0,85 a 1,13. El azar hizo el trabajo que en el tercer intento hizo el control.

Fíjate en el R cuadrado: apenas 0,04. 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.

```
ttest motivacion, by(clases)
```

```
. ttest motivacion, by(clases)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   2,489   -.5407944     .017097    .8529687   -.5743203   -.5072685
       1 |   2,511    .5119225    .0171754    .8606557    .4782431    .5456018
---------+--------------------------------------------------------------------
combined |   5,000     -.01212    .0142206    1.005548   -.0399986    .0157586
---------+--------------------------------------------------------------------
    diff |           -1.052717    .0242353               -1.100229   -1.005205
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -43.4374
Ho: diff = 0                                     degrees of freedom =     4998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

.
```

FIGURA 1.9 Salida de Stata: motivación promedio según la elección de clases particulares. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Ahí está el sesgo de selección, a la vista. Los que tomaron clases tenían, en promedio, una motivación de 0,51; los que no, de −0,54. Stata calcula la diferencia como el grupo 0 menos el grupo 1 (`diff = mean(0) − mean(1)`), por eso sale negativa: −1,05. El estadístico t es −43,44, y el valor p es 0,000.

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

```
generate horas = max(0, 6 + 2*motivacion + rnormal(0, 3))
correlate motivacion horas
regress nota clases horas
```

```
. generate horas = max(0, 6 + 2*motivacion + rnormal(0, 3))

. correlate motivacion horas
(obs=5,000)

             | motiva~n    horas
-------------+------------------
  motivacion |   1.0000
       horas |   0.5453   1.0000


. regress nota clases horas

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(2, 4997)      =   1655.25
       Model |  15105.7299         2  7552.86495   Prob > F        =    0.0000
    Residual |  22801.1546     4,997  4.5629687   R-squared       =    0.3985
-------------+----------------------------------   Adj R-squared   =    0.3983
       Total |  37906.8845     4,999  7.58289347   Root MSE        =    2.1361

------------------------------------------------------------------------------
        nota |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clases |     2.57682   .0631424    40.81   0.000     2.453034    2.700607
       horas |    .2454684    .009107    26.95   0.000     .2276147     .263322
       _cons |    8.738126   .062761    139.23   0.000     8.615086    8.861165
------------------------------------------------------------------------------

.
```

FIGURA 1.10 Salida de Stata: correlación entre motivación y horas de estudio, y regresión con las horas como proxy. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Las horas de estudio se parecen a la motivación: la correlación es 0,55. Aun así, el coeficiente de las clases apenas bajó de 3,07 a 2,58. Seguimos muy lejos del 1 verdadero, y la tabla sigue luciendo impecable, con su valor p de 0,000.

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	3,07	No: triplica el efecto
Regresión simple	3,07	No: es lo mismo con más decimales
Regresión con el confusor	0,96	Sí, si mides el confusor
Regresión con una proxy del confusor	2,58	No: corrige solo una parte
Experimento (sorteo)	0,99	Sí, sin medir nada más

Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do); cifras de las figuras 1.4, 1.5, 1.6, 1.10 y 1.8.

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 3,07.

¿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 la base, «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 salario por terminar la universidad y controlas por la ocupación del trabajador. Parece razonable, hasta que te preguntas por qué alguien trabaja de analista y no de ayudante: en buena parte, porque terminó la universidad.

Al comparar solo a personas con la misma ocupación, borras justo uno de los caminos por los que la universidad sube el salario, 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 tienen más años de estudio ganan más, aun entre personas de la misma edad y región. Como los datos son observacionales, interpretamos esta 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 Stata en una tarde.

EJERCICIOS RESUELTOS

Los tres están al final de `cap01.do`, bajo el rótulo `EJERCICIOS RESUELTOS`. Córrelos después de los pasos 1 a 9 y en este orden, porque usan la base que ya tienes en memoria. 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?

```
generate distancia = runiform(0, 10)
generate clases1 = rbinomial(1, invlogit(2 - 0.4*distancia))
generate nota1 = clip(11 + 1*clases1 + 2*motivacion + ruido, 0, 20)
tabstat nota1 motivacion, by(clases1) statistics(mean n)
```

```
. generate distancia = runiform(0, 10)

. generate clases1 = rbinomial(1, invlogit(2 - 0.4*distancia))

. generate nota1 = clip(11 + 1*clases1 + 2*motivacion + ruido, 0, 20)

. tabstat nota1 motivacion, by(clases1) statistics(mean n)

Summary statistics: mean, N
  by categories of: clases1

 clases1 |     nota1  motiva~n
---------+--------------------
       0 |  11.01039 -.0211742
         |      2545      2545
---------+--------------------
       1 |  12.02065 -.0027338
         |      2455      2455
---------+--------------------
   Total |  11.50643   -.01212
         |      5000      5000
-----------------------------

.
```

FIGURA 1.11 Salida de Stata: nota y motivación promedio cuando la elección depende solo de la distancia. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Solución. La función `runiform(0, 10)` reparte la distancia al azar entre 0 y 10 kilómetros. Los que tomaron clases promedian 12,02 y los que no, 11,01: una diferencia de 1,01, casi el 1 verdadero. Y eso que la motivación sigue subiendo la nota 2 puntos por unidad. La columna de la motivación explica por qué: −0,021 en un grupo y −0,003 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).

```
quietly regress nota clases
scalar sesgo_a = _b[clases] - 1
quietly regress motivacion clases
scalar sesgo_b = 2*_b[clases]
display "Sesgo = diferencia observada - efecto real: " %6.3f sesgo_a
display "Sesgo = diferencia de motivacion x 2:       " %6.3f sesgo_b
```

```
. quietly regress nota clases

. scalar sesgo_a = _b[clases] - 1

. quietly regress motivacion clases

. scalar sesgo_b = 2*_b[clases]

. display "Sesgo = diferencia observada - efecto real: " %6.3f sesgo_a
Sesgo = diferencia observada - efecto real:  2.071

. display "Sesgo = diferencia de motivacion x 2:       " %6.3f sesgo_b
Sesgo = diferencia de motivacion x 2:       2.105

.
```

FIGURA 1.12 Salida de Stata: el sesgo de selección calculado de dos formas. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

Solución. Aquí aparecen tres herramientas nuevas. El prefijo `quietly` corre la regresión sin mostrarla. `_b[clases]` guarda el coeficiente de `clases` de la última regresión. Y `scalar` lo guarda en un número con nombre, que luego muestras con `display` y el formato `%6.3f` (tres decimales).

Las dos cuentas casi coinciden: 2,071 y 2,105. Es la ecuación de la trampa comprobada con números: 3,07 observados = 1 de efecto + unos 2,07 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?

```
generate c3 = rbinomial(1, 0.5)
generate horas3 = 5 + 3*c3 + rnormal(0, 2)
generate nota3 = clip(11 + 0.4*c3 + 0.2*horas3 + rnormal(0, 1.5), 0, 20)
regress nota3 c3
```

```
. generate c3 = rbinomial(1, 0.5)

. generate horas3 = 5 + 3*c3 + rnormal(0, 2)

. generate nota3 = clip(11 + 0.4*c3 + 0.2*horas3 + rnormal(0, 1.5), 0, 20)

. regress nota3 c3

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(1, 4998)      =    497.03
       Model |  1182.23238         1  1182.23238   Prob > F        =    0.0000
    Residual |  11888.1938     4,998  2.37859019   R-squared       =    0.0905
-------------+----------------------------------   Adj R-squared   =    0.0903
       Total |  13070.4261     4,999  2.61460815   Root MSE        =    1.5423

------------------------------------------------------------------------------
       nota3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c3 |   .9725405   .0436231    22.29   0.000     .8870201    1.058061
       _cons |   12.05354   .0307349   392.18   0.000     11.99329    12.1138
------------------------------------------------------------------------------

.
```

FIGURA 1.13 Salida de Stata: efecto de las clases sorteadas sin controlar por las horas de estudio. Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

```
regress nota3 c3 horas3
```

```
. regress nota3 c3 horas3

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(2, 4997)      =    403.34
       Model |  1816.72739         2  908.363694   Prob > F        =    0.0000
    Residual |  11253.6988     4,997  2.25209101   R-squared       =    0.1390
-------------+----------------------------------   Adj R-squared   =    0.1387
       Total |  13070.4261     4,999  2.61460815   Root MSE        =    1.5007

------------------------------------------------------------------------------
       nota3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c3 |   .4310146   .0533164     8.08   0.000     .3264911    .5355381
      horas3 |   .1788297   .0106541    16.78   0.000     .1579429    .1997165
       _cons |   11.15163   .0614953   181.34   0.000     11.03107    11.27218
------------------------------------------------------------------------------

.
```

FIGURA 1.14 Salida de Stata: efecto de las clases sorteadas controlando por las horas de estudio (mal control). Fuente: elaboración propia con Stata 16 y datos simulados (cap01.do).

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 0,97, con un intervalo de 0,89 a 1,06. Al controlar por las horas, el coeficiente cae a 0,43. 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 trabajador sin terminar la universidad?». 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

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/

Lawlor, D. A., Davey Smith, G., & Ebrahim, S. (2004). Socioeconomic position and hormone replacement therapy use: Explaining the discrepancy in evidence from observational and randomized controlled trials. *American Journal of Public Health, 94*(12), 2149-2154. https://doi.org/10.2105/ajph.94.12.2149

Matthews, R. (2000). Storks deliver babies (p = 0.008). *Teaching Statistics, 22*(2), 36-38. https://doi.org/10.1111/1467-9639.00013

Schultz, T. P. (2001). *School subsidies for the poor: Evaluating the Mexican Progresa poverty program* (Economic Growth Center Discussion Paper No. 834). Yale University. http://www.econ.yale.edu/growth_pdf/cdp834.pdf

Schultz, T. P. (2004). School subsidies for the poor: Evaluating the Mexican Progresa poverty program. *Journal of Development Economics, 74*(1), 199-250. https://doi.org/10.1016/j.jdeveco.2003.12.009

Stampfer, M. J., & Colditz, G. A. (1991). Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Preventive Medicine, 20*(1), 47-63. https://doi.org/10.1016/0091-7435(91)90006-p

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *JAMA, 288*(3), 321-333. https://doi.org/10.1001/jama.288.3.321