

# Repaso Práctico en R e Introducción a Variables Instrumentales

## Clase 11 — 14.310x: Data Analysis for Social Scientists

B.Sc. Cristian Lazo Quispe

BREIT (Brescia Institute of Technology) · Aporta — Grupo Breca

Programa SDS (Statistics and Data Science, MITx) · BREIT (Brescia Institute of Technology), iniciativa de Aporta —Grupo Breca—  
en alianza con MIT IDSS.

Basado en MIT 14.310x (CC BY-NC-SA 4.0).

✉ [clazoq@uni.pe](mailto:clazoq@uni.pe) · [in linkedin.com/in/cristian-lazo-quispe](https://www.linkedin.com/in/cristian-lazo-quispe) · [github.com/CristianLazoQuispe](https://github.com/CristianLazoQuispe)

## ¿Cómo va a ser hoy? (clase de problemas)

### El plan

Hoy es distinta: **pocos slides, mucha práctica**. El plan:

- 1 **Kahoot** de repaso (Clases 9 y 10) para arrancar.
- 2 Un repaso exprés + lo nuevo de hoy: **endogeneidad y variables instrumentales (IV)**, con pocas fórmulas y muchos ejemplos.
- 3 **Mini-Kahoot** de lo nuevo, apenas lo veamos.
- 4 **Breakout rooms**: resuelven 2 problemas en R.
- 5 Cerramos resolviendo juntos en plenaria.

## ¡Arrancamos con un Kahoot! (repaso Clases 9 y 10)

### Únete ahora

Entra a **kahoot.it** con el **PIN** en pantalla. Son **20 preguntas** de repaso: modelo lineal y OLS (Clase 9), y regresión en la práctica y OVB (Clase 10).

### Reglas del juego

- Responde rápido: la velocidad da más puntos.
- Cuidado con las **trampas** (el error clásico).
- Cada pregunta trae su imagen y, al revelar, la explicamos paso a paso.



## ¿Qué veremos hoy?

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra

## Repaso exprés: el modelo lineal (Clase 9)

| Idea                                                       | En una frase                                                               |
|------------------------------------------------------------|----------------------------------------------------------------------------|
| OLS ( $\hat{\beta}_1 = \text{Cov}(X, Y) / \text{Var}(X)$ ) | la pendiente que mejor ajusta la nube de puntos                            |
| <i>Ceteris paribus</i>                                     | el coeficiente se lee “manteniendo lo demás constante”                     |
| $R^2$                                                      | % de la variación de Y que el modelo explica                               |
| Gauss–Markov (CLRM)                                        | con esos supuestos, OLS es el mejor estimador lineal insesgado             |
| Dummy variable                                             | su coeficiente es la diferencia de medias entre el grupo 1 y el grupo 0    |
| Multicolinealidad perfecta                                 | dos regresores que se repiten exacto $\Rightarrow$ R no puede estimar (NA) |

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10**
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra

# Repaso exprés: regresión en la práctica y OVB (Clase 10)

| Idea                               | En una frase                                                                                                  |
|------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Log-lin                            | el coeficiente $\times 100$ es el % que cambia $Y$                                                            |
| Cuadrático                         | punto de retorno $X^* = -\beta_1 / (2\beta_2)$                                                                |
| <b>OVB (Omitted Variable Bias)</b> | $\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2 \delta$ ; el signo del sesgo es el de $\beta_2 \cdot \delta$ |
| Efectos fijos                      | factor(grupo) absorbe lo constante dentro del grupo (variación <i>within</i> )                                |
| Errores robustos                   | la heterocedasticidad no sesga $\hat{\beta}$ , pero sí su error estándar                                      |
| RDD                                | el salto $\tau$ en el umbral = efecto causal <i>local</i>                                                     |

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB**
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra

## La simulación que te debíamos: ¿el OVB (Omitted Variable Bias) es mala suerte, o sistemático?

La pregunta: apuesta antes de pasar de página

En Clase 10, **una** muestra dio un coeficiente de educación = 302 al omitir la habilidad (verdadero: 150). ¿Fue mala suerte de esa muestra?

El experimento: repetirlo 1000 veces

Generamos 1000 muestras **nuevas** con el mismo modelo y guardamos el coeficiente de educación cada vez, con y sin controlar la habilidad.

# La simulación que te debíamos: ¿el OVB (Omitted Variable Bias) es mala suerte, o sistemático?

La pregunta: apuesta antes de pasar de página

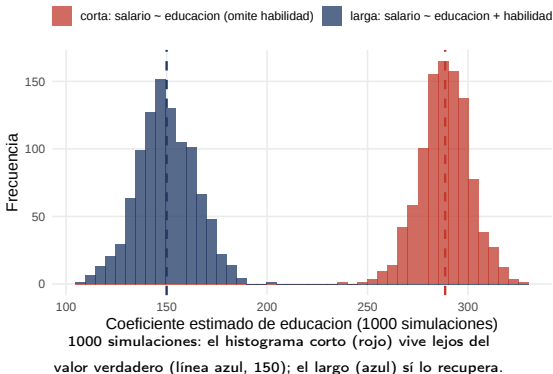
En Clase 10, **una** muestra dio un coeficiente de educación = 302 al omitir la habilidad (verdadero: 150). ¿Fue mala suerte de esa muestra?

El experimento: repetirlo 1000 veces

Generamos 1000 muestras **nuevas** con el mismo modelo y guardamos el coeficiente de educación cada vez, con y sin controlar la habilidad.

## Resultado

Promedio **corta** (omite habilidad)  $\approx 289$ . Promedio **larga** (controla habilidad)  $\approx 150$ . **No es mala suerte**: el sesgo se repite siempre, en la misma dirección.



# La simulación que te debíamos: ¿el OVB (Omitted Variable Bias) es mala suerte, o sistemático?

La pregunta: apuesta antes de pasar de página

En Clase 10, **una** muestra dio un coeficiente de educación = 302 al omitir la habilidad (verdadero: 150). ¿Fue mala suerte de esa muestra?

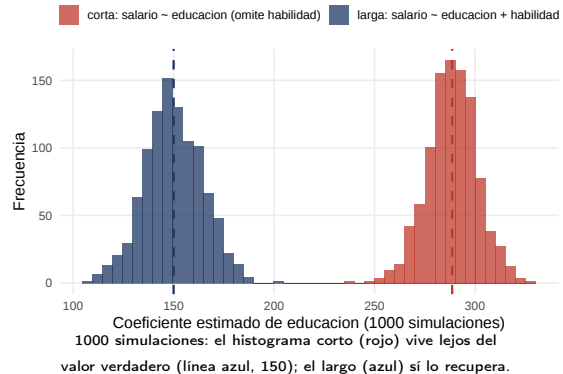
El experimento: repetirlo 1000 veces

Generamos 1000 muestras **nuevas** con el mismo modelo y guardamos el coeficiente de educación cada vez, con y sin controlar la habilidad.

## Resultado

Promedio **corta** (omite habilidad)  $\approx 289$ . Promedio **larga** (controla habilidad)  $\approx 150$ . **No es mala suerte**: el sesgo se repite siempre, en la misma dirección.

Ojo: esto NO “arregla” el sesgo



# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra

## Endogeneity: cuando $X$ e $Y$ se afectan mutuamente

### La idea, en palabras

Hasta ahora, el remedio al sesgo era **agregar un control**. Pero a veces no hay control que sirva: la variable que confunde **no se puede medir** (habilidad, motivación...), o peor:  $Y$  también afecta de vuelta a  $X$ . A eso se le llama **endogeneity (endogeneidad)**.

### Ejemplos (MIT L21)

Precio y cantidad (se determinan juntos); salud y ejercicio (¿el ejercicio mejora la salud, o solo quien ya está sano puede ejercitarse?); educación e ingreso (la educación viene ligada a la habilidad, la motivación, el contexto familiar...).

### La clave: la regresión NO distingue direcciones

$\text{Cov}(X, Y)$  es el **mismo número** exista o no exista causalidad de  $X$  a  $Y$ , de  $Y$  a  $X$ , o de ambas a la vez. Que  $X$  “pueda” afectar a  $Y$  no lo ves en los datos: lo **argumentas** con lo que sabes del contexto, *antes* de correr el modelo.

## Demo: policía y crimen, ¿el signo puede estar al revés?

El problema: piénsalo antes de pasar de página

Supongamos que la policía **SÍ** reduce el crimen (efecto verdadero =  $-0.5$ ). Pero también las ciudades **reaccionan** al crimen contratando más policías. Ambas flechas conviven al mismo tiempo. **¿Qué signo esperas** si corres  $\text{lm}(\text{crimen} \sim \text{policia})$ ? ¿Positivo, negativo, o no se puede saber?

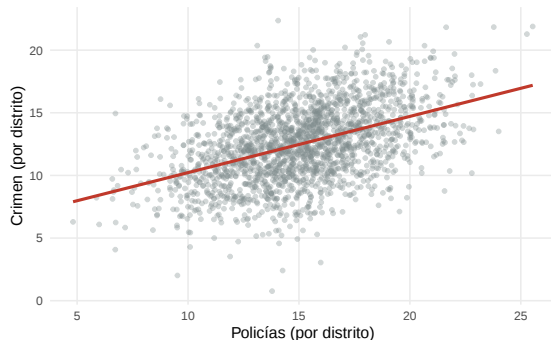
## Demo: policía y crimen, ¿el signo puede estar al revés?

El problema: piénsalo antes de pasar de página

Supongamos que la policía **SÍ** reduce el crimen (efecto verdadero =  $-0.5$ ). Pero también las ciudades **reaccionan** al crimen contratando más policías. Ambas flechas conviven al mismo tiempo. ¿Qué signo esperas si corres `lm(crimen ~ policia)`? ¿Positivo, negativo, o no se puede saber?

Resultado (simulado y verificado en R)

Efecto verdadero:  $-0.5$  (protege). Coeficiente que **de verdad obtienes** al correr la regresión:  $+0.45$ . ¡El signo se invirtió!



La recta ajustada (roja) sube, aunque el efecto verdadero de policía sobre crimen es negativo.

## Demo: policía y crimen, ¿el signo puede estar al revés?

El problema: piénsalo antes de pasar de página

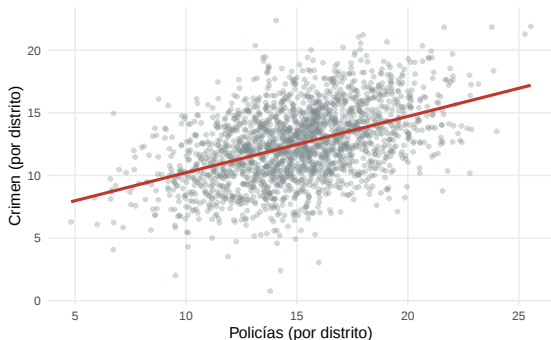
Supongamos que la policía **SÍ** reduce el crimen (efecto verdadero =  $-0.5$ ). Pero también las ciudades **reaccionan** al crimen contratando más policías. Ambas flechas conviven al mismo tiempo. ¿Qué signo esperas si corres `lm(crimen ~ policia)`? ¿Positivo, negativo, o no se puede saber?

Resultado (simulado y verificado en R)

Efecto verdadero:  $-0.5$  (protege). Coeficiente que **de verdad obtienes** al correr la regresión:  $+0.45$ . ¡El signo se invirtió!

### La lección

La regresión no “vio” el efecto protector: lo que ves es la **mezcla** de las dos flechas. Solo un instrument (algo que mueva a la policía sin depender del crimen local) puede aislar el  $-0.5$  verdadero.



La recta ajustada (roja) sube, aunque el efecto verdadero de policía sobre crimen es negativo.

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales**
- 6 Check de comprensión
- 7 Manos a la obra

## La idea del instrument: una palanca externa

### ¿Qué buscamos?

Algo que mueva a  $X$  **como un sorteo**, sin tocar a  $Y$  por ningún otro lado. Ejemplo (Ghana, L21): una **beca de estudios sorteada** ( $Z$ ) sube la probabilidad de terminar secundaria ( $X$ ), y solo debería afectar el ingreso futuro ( $Y$ ) a través de  $X$ .

### Las 3 condiciones de un buen instrument

**Relevance:**  $Z$  **SÍ** debe mover a  $X$  (si no, el Wald divide entre cero). **Independence:**  $Z$  está asignado al azar (o “as good as random”). **Exclusion restriction:**  $Z$  **NO** debe afectar a  $Y$  por ninguna otra vía. Piensa en un tubo  $Z \rightarrow X \rightarrow Y$ : el instrument empuja fuerte por un extremo, y no debe haber ninguna cañería lateral hacia  $Y$ . Esta última condición NO se puede comprobar con datos: se **argumenta**.



$Z$  sí puede llegar a  $Y$ , pero solo pasando por  $X$ . La flecha roja de  $Z$  directo a  $Y$  es justo lo que la exclusion restriction prohíbe.

## Un instrument NO tiene que venir de un sorteo real

### La idea

La fórmula del Wald no exige un experimento: solo necesita una variable  $Z$  que empuje a  $X$  “como si fuera un sorteo”.  $Z$  puede venir de un RCT real (la beca), o puede estar **ya en tus datos** si puedes **argumentar** que es “as good as random”.

### Ejemplo: un instrument “encontrado”

En Perú, la ley obliga a entrar a primaria si cumples 6 años antes del 31 de marzo: los nacidos en enero-marzo entran antes y terminan con más años de educación a la misma edad.  $Z_i = 1$  si nació entre enero y marzo: **nadie la sorteó**, ya estaba en la encuesta.

### La lección

Con datos observacionales, la tarea es **buscar** variación arbitraria (fechas de corte legal, distancias, cohortes de una reforma) y **argumentar** por qué es “limpia”. Si no encuentras nada creíble: reporta el OLS honestamente como una *cota*, no inventes un instrument.

## First stage, reduced form y el Wald estimator

Dos regresiones simples, sobre el instrument

**First stage:**  $X$  sobre  $Z$  (¿cuánto mueve el instrument a  $X$ ?). **Reduced form:**  $Y$  sobre  $Z$  (¿cuánto mueve el instrument a  $Y$ ?). El **Wald estimator** combina las dos:

$$\hat{\beta}_{\text{Wald}} = \frac{\text{reduced form}}{\text{first stage}}.$$

Aplicar la fórmula (beca  $\rightarrow$  secundaria  $\rightarrow$  ingreso)

First stage = 0.40 (la beca sube 40 puntos porcentuales la probabilidad de terminar secundaria). Reduced form = 320 (la beca sube el ingreso en 320 soles, en promedio).

$$\hat{\beta}_{\text{Wald}} = \frac{320}{0.40} = 800.$$

El efecto de terminar secundaria sobre el ingreso es 800 soles.

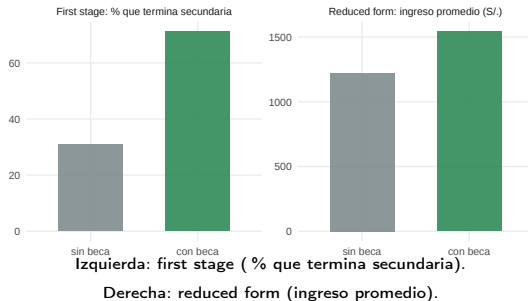
**Cuidado: instrumento débil**

Si el first stage es chico (cerca de 0), dividir entre él **amplifica** el ruido y cualquier violación pequeña de la exclusion restriction. Un buen instrument necesita un first stage **visible**.

## Demo: la beca de estudios, OLS vs. Wald estimator

Problema: respóndelo antes de pasar de página

Comparar directamente a quienes terminaron secundaria con quienes no da 1141 soles (**OLS ingenuo**). ¿Por qué no confiar en ese número?



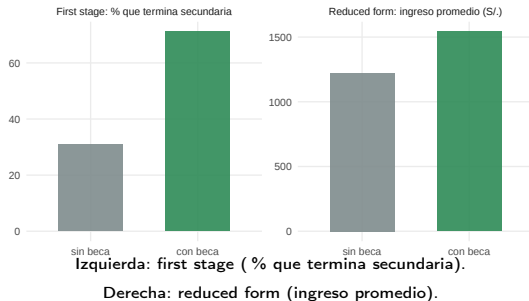
## Demo: la beca de estudios, OLS vs. Wald estimator

Problema: respóndelo antes de pasar de página

Comparar directamente a quienes terminaron secundaria con quienes no da 1141 soles (**OLS ingenuo**). ¿Por qué no confiar en ese número?

### Solución

La habilidad (no observada) empuja secundaria e ingreso a la vez: el OLS ingenuo está contaminado, igual que el OVB de Clase 10. El Wald estimator (800) solo usa la variación que vino del **sorteo** de la beca, libre de esa contaminación.



## LATE: ¿el efecto de qué grupo estamos midiendo?

### 4 tipos de persona frente a un instrument binario

**Always-takers:** toman el tratamiento pase lo que pase. **Never-takers:** nunca lo toman. **Compliers:** lo toman *solo* si el instrument los empuja. **Defiers:** al revés (asumimos que no existen).

### ¿Por qué “local”?

Los always-takers y never-takers no cambian su  $X$  aunque cambie  $Z$ : no aportan nada al first stage. **Todo** el first stage viene de los compliers. Por eso el Wald estima el **Local Average Treatment Effect (LATE)**: el efecto solo sobre los compliers, no sobre toda la población.

### Dos instrumentos, dos LATE distintos

Si el efecto verdadero varía de persona a persona, dos instrumentos válidos (que empujan a grupos distintos de compliers) pueden dar números distintos entre sí, y distintos del OLS. Es justo lo que pasa en MIT Homework 10 (Angrist & Evans 1998).

## 2SLS: la generalización (y cómo se ve en R)

### Con más de un instrument o controles adicionales

El Wald estimator se generaliza al **2SLS** (two-stage least squares): primero regresionas  $X$  sobre el/los instrument(s) y controles (**first stage**), guardas los valores predichos  $\hat{X}$ , y luego regresionas  $Y$  sobre  $\hat{X}$  (**second stage**). Con un solo instrument binario, da **exactamente lo mismo** que el Wald estimator.

### En R (una sola línea, sin programar las dos etapas a mano)

```
library(AER)
m_iv <- ivreg(ingreso ~ secundaria | beca, data = datos)
summary(m_iv)
```

### ¿Y el diseño experimental?

El módulo M10 también cubre diseño experimental (qué y a quién aleatorizar, *phase-in*, *encouragement design*). Es contenido conceptual, sin fórmulas: lo dejamos como lectura opcional, el foco de hoy es el Wald estimator.

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión**
- 7 Manos a la obra

## Mini-Kahoot: endogeneidad e IV

### Únete de nuevo

Ahora sí, **10 preguntas** sobre lo que acabamos de ver: endogeneity, instrument, first stage, reduced form, Wald estimator, LATE y compliers.

### Recuerda

Cuidado con las trampas: dividir al revés (first stage / reduced form en vez de al revés), confundir first stage con reduced form, y olvidar que el Wald estima un efecto **local** (LATE), no el efecto para todos.

# Contenido

- 1 Repaso exprés: Clase 9
- 2 Repaso exprés: Clase 10
- 3 La deuda: simular el OVB
- 4 ¿Y si controlar no basta?
- 5 Variables instrumentales
- 6 Check de comprensión
- 7 Manos a la obra**

## Breakout rooms: 2 problemas en R

- Formamos grupos de 3–4 personas.
- Cada grupo resuelve los **2 problemas** de las siguientes láminas (`practica_R/p1_mc_ovb.R` y `p2_wald_beca.R`, con espacios \_\_\_\_ para completar).
- Si se traban, revisen el repaso exprés o pregunten.
- Al final, resolvemos juntos en plenaria con el archivo `_sol.R`.

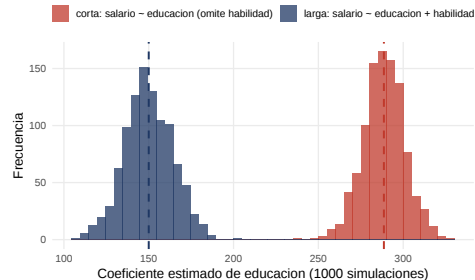
# Práctica 1 en R: Monte Carlo simulation del OVB

## Problema — resuélvelo *desde cero* en R

Repita 1000 veces el modelo  $\text{salario} = 800 + 150 \cdot \text{educacion} + 400 \cdot \text{habilidad} + \text{error}$  (mismo modelo de Clase 10), guardando el coeficiente de educación con y sin controlar habilidad.

- 1 ¿Cuánto da el promedio de la regresión *corta*?
- 2 ¿Cuánto da el promedio de la *larga*?
- 3 ¿El sesgo desaparece si subes  $N$ ?

Desde cero. Si te trabas: `practica_R/p1_mc_ovb.R` (con vacíos). Solución al final: `p1_mc_ovb_sol.R`.



1000 simulaciones: corta vs. larga.

## Práctica 2 en R: Wald estimator (beca de estudios)

### Problema — resuélvelo *desde cero* en R

La verdad (formal): el dataset es simulado, así que aquí SÍ conocemos el modelo que generó los datos (en la vida real nunca se observa):

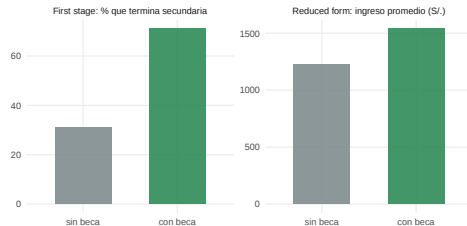
$$\text{ingreso} = 1000 + 800 \cdot \text{secundaria} + 700 \cdot \text{habilidad} + e,$$

con *habilidad* no observada y *beca* sorteada. El reto: recuperar el 800 usando solo *beca.csv* (columnas *beca*, *secundaria*, *ingreso*).

- 1 First stage:  $\text{secundaria} \sim \text{beca}$ . Anota el coeficiente.
- 2 Reduced form:  $\text{ingreso} \sim \text{beca}$ . Anota el coeficiente.
- 3 Calcula el Wald estimator (reduced form / first stage).
- 4 Compáralo con el OLS ingenuo ( $\text{ingreso} \sim \text{secundaria}$ ). ¿Cuál es más creíble?

Desde cero. Si te trabas: `practica_R/p2_wald_beca.R`. Solución al final:

`p2_wald_beca_sol.R`.



First stage y reduced form de la beca.

## Checklist: ¿cuándo necesito un instrument?

| Pregunta                                        | Riesgo si la ignoras                             | Remedio                                 |
|-------------------------------------------------|--------------------------------------------------|-----------------------------------------|
| ¿Puedo medir el confusor?                       | si sí, es solo OVB de Clase 10                   | agrégalo como control                   |
| ¿No lo puedo medir, o hay causalidad de vuelta? | endogeneity: ni el mejor control arregla esto    | buscar un instrument (Z)                |
| ¿El instrument mueve a $X$ ?                    | sin esto, no hay first stage                     | verificalo con datos                    |
| ¿El instrument toca $Y$ por otra vía?           | exclusion restriction violada, sesgo amplificado | argumentarlo con cuidado, caso por caso |

### La pregunta que resume la clase

Antes de creerle a un coeficiente: “¿puedo medir lo que confunde, o necesito una palanca externa (instrument) para aislar el efecto?”

## ¿Y ahora?

- **OVB (Clase 10):** se arregla *midiendo* y controlando la omitida. La simulación de hoy confirmó que el sesgo es sistemático, no mala suerte.
- **Endogeneity:** cuando NO puedes medir el confusor, o  $Y$  afecta de vuelta a  $X$ . La regresión simple no distingue direcciones de causalidad.
- **Instrument:** una palanca externa que mueve  $X$  como un sorteo, sin tocar  $Y$  por otra vía (relevance, independence, exclusion restriction).
- **Wald estimator** = reduced form / first stage. Estima el **LATE**: el efecto solo sobre los **compliers**.
- **2SLS** generaliza el Wald a varios instrumentos/controles; en R, `ivreg(y ~ x | z)`.

¡A practicar!