

14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

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

Clase 11

Repaso Práctico en R e Introducción a Variables Instrumentales

Teoría

Módulos oficiales del MicroMaster:

M10 – Endogeneity, Instrumental Variables and Experimental Design

B.Sc. Cristian Lazo Quispe

✉ clazoq@uni.pe

[in linkedin.com/in/cristian-lazo-quispe](https://www.linkedin.com/in/cristian-lazo-quispe) · github.com/CristianLazoQuispe

Material del *Advanced Program in Data Science & Global Skills*, diseñado y desarrollado por **BREIT (Brescia Institute of Technology)**, iniciativa de **Aporta** —plataforma de innovación e impacto social del **Grupo Breca**— en alianza con el **MIT IDSS** (Institute for Data, Systems, and Society).

Basado en MIT 14.310x (E. Duflo & S. Ellison) y en diversas fuentes abiertas (ver bibliografía al final), bajo licencia CC BY-NC-SA 4.0.

4 de agosto de 2026

Índice

1. De qué trata esta clase	2
2. Endogeneity	2
3. Instrumental variables	3
4. El Wald estimator	4
5. Local Average Treatment Effect (LATE)	5
6. 2SLS: la generalización	5

1. De qué trata esta clase

En las Clases 9 y 10 aprendimos a estimar y a interpretar una regresión, y a detectar (y corregir) el **sesgo de variable omitida (OVB)** agregando los controles correctos. Pero a veces **no hay** un control que arregle el problema: la variable que nos interesa está “contaminada” por algo que no podemos observar ni medir. Hoy vemos qué hacer en ese caso: usar una **variable instrumental (instrument)**.

Controlar sirve cuando podemos *medir* lo que confunde. Cuando no podemos (por ejemplo, la “habilidad” o la “motivación” de una persona), necesitamos otra estrategia: encontrar algo que mueva a X **como si fuera un sorteo**, sin tocar a Y por ningún otro lado. A eso llamamos un **instrument**.

Glosario de siglas de esta clase

Aviso para no confundir **IV** con el número romano 4: aquí **todas** las siglas son iniciales en **inglés**.

Sigla	Significado
IV	Instrumental Variables (variables instrumentales)
2SLS	Two-Stage Least Squares (mínimos cuadrados en dos etapas)
LATE	Local Average Treatment Effect (efecto de tratamiento promedio local)
OLS	Ordinary Least Squares (lo que hace lm , ya visto en Clase 9)
RCT	Randomized Controlled Trial (experimento aleatorizado, Clase 8)
SE	Standard Error (error estándar)

De ahí sale también el nombre de la función de R: `ivreg()` = **instrumental variables regression**.

2. Endogeneity

Definición 2.1 (Endogeneity (endogeneidad)). Decimos que X es **endógena** cuando está correlacionada con el error ε del modelo: algo no observado afecta a la vez a X y a Y (o Y afecta de vuelta a X). En ese caso, OLS de Y sobre X **no** estima el efecto causal, aunque hayamos agregado todos los controles que se nos ocurran.

La raíz del problema, en fórmula (el OVB de Clase 10 es un caso particular)

Con $Y_i = \alpha + \beta X_i + \varepsilon_i$, el estimador OLS es $\hat{\beta}_{OLS} = \text{Cov}(Y, X) / \text{Var}(X)$. Sustituyendo Y :

$$\hat{\beta}_{OLS} = \beta + \frac{\text{Cov}(X, \varepsilon)}{\text{Var}(X)}.$$

OLS solo recupera el efecto causal β si $\text{Cov}(X, \varepsilon) = 0$. Ese segundo término es el sesgo, y **no depende del tamaño de muestra**: no es un problema de varianza (eso sí se cura con más datos, angostando el histograma), es un problema de **consistencia**. Con $n \rightarrow \infty$ convergirías con precisión perfecta al número **equivocado**.

La fórmula del OVB, $\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2 \cdot \delta$, es **el mismo resultado** para el caso particular en que lo que está metido en ε es una variable omitida W : ahí $\text{Cov}(X, \varepsilon) = \beta_2 \cdot \text{Cov}(X, W)$, y al dividir entre $\text{Var}(X)$ se obtiene exactamente $\beta_2 \cdot \delta$. La fórmula general de arriba cubre también los casos que el OVB de Clase 10 NO cubría: causalidad inversa, simultaneidad, o cualquier otra fuente de correlación entre X y ε (nuestra demo de policía y crimen es uno de esos casos, sin ninguna variable omitida de por medio). Un instrument Z funciona precisamente porque se elige para que $\text{Cov}(Z, \varepsilon) = 0$ (al venir de un sorteo), rompiendo ese término de sesgo.

Ejemplo 2.1 (Ejemplos de endogeneidad (MIT L21)). Precio y cantidad demandada (se determinan juntos); salud y ejercicio físico (¿el ejercicio mejora la salud, o solo quien ya está sano puede ejercitarse?); años de educación e ingreso futuro (la educación está ligada a la habilidad, la motivación, el contexto familiar... nada de eso se puede “controlar” del todo).

En Clase 10 la solución era *agregar un control observado*. Aquí el problema es justamente que **no tenemos** ese control (la habilidad no se mide). Por eso necesitamos una herramienta distinta: el instrument.

3. Instrumental variables

Definición 3.1 (Instrument (instrumento) Z). Un **instrument** Z es una variable que usamos para aislar la parte “limpia” de X : la parte que se mueve por el instrumento y no por lo que confunde. Un buen instrument cumple **3 condiciones**.

Las 3 condiciones de un buen instrument

1. **Relevance**: Z debe mover a X (efecto real y detectable en el **first stage**, la regresión de X sobre Z).
2. **Independence**: Z debe estar asignado tan “al azar” como un sorteo (no correlacionado con lo que confunde X e Y). Si Z viene de un RCT, esta condición está garantizada por diseño.
3. **Exclusion restriction**: Z NO debe afectar a Y por ninguna otra vía que no sea a través de X .

Relevance se verifica con datos (el first stage). Independence se garantiza por el diseño del sorteo. La **exclusion restriction** es distinta: es un supuesto que se **argumenta** caso por caso, y **nunca** se prueba con números. Una violación pequeña de esta condición puede arruinar por completo la estimación (lo vemos en la siguiente sección).

El instrument resuelve algo que la regresión simple no podía: separar direcciones de causalidad. La lógica tiene dos piezas de naturaleza distinta. **(1) Un supuesto que ponemos nosotros**, no que los datos prueban: Z no toca a Y por ningún otro camino que no sea a través de X (la exclusion restriction). No se demuestra con estadística: se **argumenta** –por ejemplo, “una beca de estudios no debería subir el ingreso de nadie por sí sola, salvo porque esa persona estudió más”. **(2) Lo que SÍ vemos en los datos**: que Z –sorteado, sin historia de causalidad inversa ni confusores propios– se correlaciona con Y (la reduced form). Si aceptamos (1) y observamos (2), la única explicación que queda en pie es que “ Z movió a Y porque Z movió a X , y X mueve a Y ”. Funciona justo donde la regresión simple fallaba: Z mismo no tiene el problema de endogeneidad, así que no puede “mentir” como sí mentiría una regresión directa de Y sobre X .

4. El Wald estimator

Ejemplo 4.1 (Motivación: una beca de estudios (MIT L21, Ghana)). Queremos el efecto de **terminar secundaria** (A) sobre el **ingreso futuro** (Y). Comparar directamente a quienes terminaron secundaria con quienes no está sesgado (la habilidad empuja las dos cosas). Pero una **beca de estudios se sorteó al azar** (Z): sube la probabilidad de terminar secundaria, y no debería afectar el ingreso por ninguna otra vía.

Wald estimator

Con un instrument binario, el efecto causal de X sobre Y se estima con:

$$\hat{\beta}_{\text{Wald}} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]} = \frac{\text{reduced form}}{\text{first stage}}.$$

El numerador (**reduced form**) es el efecto del instrument sobre Y ; el denominador (**first stage**) es el efecto del instrument sobre X . Ambos son creíbles porque Z fue sorteado: cada uno es un simple contraste tratamiento-control.

Ejemplo 4.2 (Aplicar la fórmula). Con la beca: first stage = 0.40 (sube 40 puntos porcentuales la probabilidad de terminar secundaria), reduced form = 320 (el ingreso sube en promedio 320 soles). Entonces

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

En palabras: el efecto de terminar secundaria sobre el ingreso es 800 soles.

Si el first stage es chico (el instrument casi no mueve a X), el Wald estimator se vuelve muy inestable: dividir entre un número cercano a cero amplifica cualquier ruido muestral. Por la misma razón, una violación pequeña de la exclusion restriction se “agrandar” muchísimo cuando el first stage es débil.

5. Local Average Treatment Effect (LATE)

Los 4 tipos de persona frente a un instrument binario

- **Always-takers:** toman el tratamiento con o sin el instrument.
- **Never-takers:** nunca lo toman.
- **Compliers:** lo toman **solo** si el instrument se los “empuja”.
- **Defiers:** al revés (se asume que no existen).

Los always-takers y never-takers no cambian su X pase lo que pase con Z : no aportan nada al first stage. Todo el first stage –y por lo tanto todo lo que el Wald estimator puede “ver”– viene de los **compliers**. Por eso el Wald estima el **Local Average Treatment Effect (LATE)**: el efecto promedio **solo sobre los compliers**, no sobre toda la población.

Dos instrumentos válidos para la misma pregunta pueden dar números distintos, porque cada uno “empuja” a un grupo distinto de compliers. Si el efecto verdadero varía de persona a persona (efectos heterogéneos), no hay razón para que los LATE de dos instrumentos coincidan entre sí ni con el OLS. Esto es exactamente lo que encuentran Angrist & Evans (1998, MIT Homework 10) al usar dos instrumentos distintos para el efecto de un tercer hijo sobre el trabajo de la madre.

6. 2SLS: la generalización

Two-stage least squares (2SLS)

Cuando hay **varios** instrumentos o controles adicionales, el Wald estimator se generaliza al **2SLS**, en dos pasos:

1. **First stage:** regresión de X sobre el/los instrument(s) y los controles; se guardan los valores **predichos** $\widehat{X}$.
2. **Second stage:** regresión de Y sobre $\widehat{X}$ (en vez de sobre el X original).

Con un solo instrument binario, el 2SLS da **exactamente el mismo** número que el Wald estimator.

En R: Wald estimator y 2SLS

```
# Wald estimator (a mano, con base R):
first_stage <- coef(lm(secundaria ~ beca, data = datos))["beca"]
reduced_form <- coef(lm(ingreso ~ beca, data = datos))["beca"]
wald <- reduced_form / first_stage

# 2SLS con el paquete AER (una sola linea, no hace falta programar
# las dos etapas a mano):
library(AER)
m_iv <- ivreg(ingreso ~ secundaria | beca, data = datos)
summary(m_iv)
```

El módulo oficial del MIT (M10) también cubre **diseño experimental** (cómo elegir qué y a quién aleatorizar, phase-in, encouragement design, efectos de equilibrio general). Es contenido **conceptual**, sin fórmulas ni cálculos: lo mencionamos en los slides, pero el foco de esta clase –y de la práctica– es el Wald estimator y las 3 condiciones de un buen instrument.

Fuentes y atribución

Material de repaso **basado en MIT 14.310x – Data Analysis for Social Scientists** (Esther Duflo & Sara Fisher Ellison), MIT OpenCourseWare, licencia **CC BY-NC-SA 4.0** (uso no comercial, con atribución, compartir bajo la misma licencia), y en las fuentes de la bibliografía. Los enunciados se basan en el material del curso y en diversas fuentes abiertas; cuando un problema se adapta de una fuente específica, se cita en el enunciado.

Para esta clase (M10 — Endogeneidad y Variables Instrumentales) se usaron, además, como *referencia/inspiración* (sin copiar texto ni código): MIT 14.310x lección **L21** (Prof. E. Duflo) y **Homework 10** (Angrist & Evans 1998, *Children and Their Parents' Labor Supply*); y los repositorios abiertos de la comunidad `mynameisjanus/14310xDataAnalysis` (Module 10) —sin licencia explícita, usado solo como referencia, no como copia. Textos de apoyo: Wooldridge, *Introductory Econometrics*; Angrist & Pischke, *Mostly Harmless Econometrics* (variables instrumentales, LATE).

Bibliografía.

- Duflo, E. & Ellison, S. F. *14.310x: Data Analysis for Social Scientists*. MIT OpenCourseWare (slides, lecture notes y problem sets). ocw.mit.edu.
- Larsen, R. J. & Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 6.^a ed., Pearson, 2018.
- DeGroot, M. H. & Schervish, M. J. *Probability and Statistics*, 4.^a ed., Pearson, 2012.
- Blitzstein, J. & Hwang, J. *Introduction to Probability* (Harvard Stat 110); W. Chen, *Probability Cheatsheet*.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics*. openintro.org.
- Materiales abiertos de la comunidad del curso en GitHub: `gevorg-hunanyan/probability`; `hanmochen/Probability-Theory-2020`; `mynameisjanus/14310xDataAnalysis`.

creativecommons.org/licenses/by-nc-sa/4.0