

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

Soluciones

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

Niveles de dificultad (*difficulty levels*)

Cada problema está rotulado por color según su nivel, de lo más accesible (Nivel 1) a lo más difícil (Nivel 5). Empieza por donde te sientas cómodo y avanza a tu ritmo; los niveles 4–5 son extensiones para profundizar, no un requisito.

- **Nivel 1 – Fundamentos:** calentamiento muy guiado; para quien parte de casi cero.
- **Nivel 2 – Núcleo:** el nivel objetivo del curso/examen.
- **Nivel 3 – Aplicación:** combina varias ideas en contexto realista.
- **Nivel 4 – Reto:** difícil, integra varios temas.
- **Nivel 5 – Reto+:** muy difícil / fuera del scope típico.

1. Problemas y soluciones

Nivel 1 – Fundamentos (warm-up)

Para quien parte de casi cero: cálculos directos y guiados.

Ejercicio 1.1 (¿Esto es endogeneidad?). La **endogeneidad** (**endogeneity**) aparece cuando X y Y se determinan mutuamente, o cuando hay una tercera variable que mueve a las dos a la vez, y por eso no podemos leer la relación entre ellas como causal.

- (a) El precio de un producto y la cantidad vendida.
- (b) Una beca de estudios asignada por **sorteo (RCT)** y el año de secundaria alcanzado.

¿Cuál de los dos casos tiene un problema de endogeneidad? ¿Por qué el otro no?

Solución

Qué nos piden. Distinguir un caso “sospechoso” de uno “limpio”.

(a) Precio y cantidad son endógenos: el precio afecta cuánto se compra, pero también la cantidad demandada afecta el precio (oferta y demanda se determinan juntas). No hay manera de decir “cuál causa cuál” solo mirando los datos.

(b) Aquí NO hay endogeneidad: la beca se asignó por **sorteo**, así que no puede depender de nada de la persona (ni de su habilidad, ni de sus ganas de estudiar). El sorteo rompe cualquier relación de ida y vuelta.

En palabras. La pregunta clave siempre es: ¿pudo Y (o algo detrás de ambas) influir en X ? Si la respuesta es sí, hay endogeneidad y OLS simple no basta.

Ejercicio 1.2 (Aplicar la fórmula de Wald, paso 1). Una beca se asignó por sorteo. Al comparar tratamiento vs. control encontramos:

$$\text{first stage} = 0.40 \quad \text{reduced form} = 320.$$

Calcula el **Wald estimator**:

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

Solución

Qué nos piden. Solo dividir dos números que ya nos dieron.

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

En palabras. El sorteo subió el ingreso en 320 (**reduced form**) porque subió en 0.40 la probabilidad de terminar secundaria (**first stage**). Repartiendo los 320 entre ese 40 %, el efecto de terminar secundaria es 800.

Ejercicio 1.3 (Repaso: leer un coeficiente dummy con más controles). Una regresión de puntaje de examen sobre `shs_complete` (terminó secundaria) y dummies de región da:

$$\widehat{\text{puntaje}} = -0.21 + 0.63 \cdot \text{shs_complete} + 0.19 \cdot \text{region2} + \dots$$

¿Cómo se interpreta el 0.63? ¿Por qué agregamos dummies de región aunque no nos interesen por sí mismas?

Solución

El 0.63: a igualdad de región (*ceteris paribus*), terminar secundaria se asocia con 0.63 puntos más en el examen.

Las dummies de región: son **controles**. No nos interesa su valor, pero si la región afecta tanto la educación como el puntaje, omitirla dejaría un sesgo de variable omitida (OV) sobre el coeficiente que sí nos importa (`shs_complete`). Es la misma idea de Clase 10: agregar un buen control “limpia” el coeficiente de interés.

Ejercicio 1.4 (Repaso relámpago: el sesgo de OV). El modelo verdadero es $\text{salario} = \beta_0 + 120 \cdot \text{educación} + 300 \cdot \text{habilidad} + \varepsilon$. La pendiente de ($\text{habilidad} \sim \text{educación}$) es $\delta = 0.3$. ¿Cuánto vale el sesgo si omitimos la habilidad? ¿La regresión corta da un número mayor o menor que 120?

Solución

Sesgo = $\beta_{\text{hab}} \cdot \delta = 300 \times 0.3 = 90$. Como el sesgo es positivo, la regresión corta da $120 + 90 = 210$: **mayor** que el verdadero 120 (sobrestima).

Nivel 2 – Núcleo (core)

El nivel objetivo del curso/examen; todos deberían llegar aquí.

Ejercicio 1.5 (Las 3 condiciones de un buen *instrument*). Un médico recuerda al azar (sorteo) a algunos pacientes que es temporada de gripe y deberían vacunarse. Queremos usar ese recordatorio como **instrument (instrumento)** para medir el efecto de vacunarse sobre los días de enfermedad.

- (a) Nombra las 3 condiciones que debe cumplir un buen instrumento.
- (b) ¿Cuál de las 3 NO se puede comprobar con los datos?

Solución

(a) **Relevance**: el instrumento debe mover a X (el recordatorio sí debe subir la tasa de vacunación). **Independence**: el instrumento debe estar tan bien asignado como un sorteo (aquí, sí: fue al azar). **Exclusion restriction**: el instrumento NO debe afectar a Y por ninguna otra vía que no sea X (el recordatorio no debería, por sí solo, cambiar los días de enfermedad de otra forma).

(b) La **exclusion restriction**. Relevance se ve en el **first stage** (con datos); independence se garantiza por el diseño del sorteo. Pero que el recordatorio NO tenga ningún otro efecto sobre la salud es un supuesto que se **argumenta**, no se prueba con números (así lo remarca la lección L21 del MIT).

Ejercicio 1.6 (First stage vs. reduced form: ¿cuál es cuál?). Tenemos Z (instrumento, sorteado), X (variable endógena) y Y (resultado). Dos regresiones:

- (a) X sobre Z .
- (b) Y sobre Z .

¿Cuál es el **first stage** y cuál el **reduced form**?

Solución

(a) es el **first stage**: mide cuánto mueve el instrumento a la variable endógena. (b) es el **reduced form**: mide cuánto mueve el instrumento al resultado final (sin pasar por X explícitamente). El Wald estimator es, justamente, (b) dividido entre (a).

Ejercicio 1.7 (Aplicar la fórmula de Wald, paso 2 – y comparar con OLS). Un programa de capacitación laboral se asignó por sorteo (Z = elegido). X = completó la capacitación, Y = % de probabilidad de conseguir empleo.

$$\text{first stage} = 0.50 \quad \text{reduced form} = 12 \quad \text{OLS ingenuo (Y sobre X)} = 30.$$

- (a) Calcula el Wald estimator.
- (b) ¿El OLS ingenuo sobrestima o subestima frente al Wald? ¿Por qué podría pasar eso?

Solución

(a) $\text{Wald} = 12/0.50 = 24$.

(b) El OLS ingenuo (30) sobrestima frente al Wald (24). Es plausible: quien **elige** completar la capacitación puede tener de entrada más motivación o habilidades para conseguir empleo (una omitida correlacionada con X). El Wald usa solo la variación que vino del sorteo, libre de esa contaminación.

Ejercicio 1.8 (¿Qué mide el *Local Average Treatment Effect* (LATE)?). Con un instrumento binario hay, en principio, 4 tipos de personas: **always-takers** (siempre toman el tratamiento, con o sin instrumento), **never-takers** (nunca lo toman), **compliers** (lo toman SOLO si el instrumento se los “empuja”), y **defiers** (al revés; se asume que no existen).

(a) ¿Para cuál de estos grupos estima un efecto el Wald/IV estimator?

(b) ¿Por qué los always-takers y never-takers NO aportan información al first stage?

Solución

(a) Solo para los **compliers**: por eso se llama efecto **local** (LATE), no el efecto promedio de toda la población.

(b) Un always-taker toma el tratamiento SIN importar el valor del instrumento (su X no cambia si Z cambia); lo mismo un never-taker (nunca cambia). El first stage mide *cuánto cambia X cuando cambia Z* : para estos dos grupos ese cambio es cero. Todo el first stage (y, por lo tanto, todo lo que el IV puede “ver”) viene de los compliers.

Ejercicio 1.9 (Instrumento binario: clasifica los 4 grupos). En el programa de becas, clasifica a cada estudiante como **complier**, **always-taker** o **never-taker** según lo que haría con y sin beca:

Estudiante	¿Termina secundaria SIN beca?	¿Termina secundaria CON beca?
A	Sí	Sí
B	No	No
C	No	Sí

Solución

A = **always-taker** (termina de todas formas). B = **never-taker** (no termina ni con beca). C = **complier** (la beca es justo lo que lo “empuja”): es a estudiantes como C a quienes describe el efecto que estima el Wald estimator.

Nivel 3 – Aplicación

Combina dos o más ideas en contexto realista; consolida lo aprendido.

Ejercicio 1.10 (¿Por qué un first stage chico “agrandar” el Wald?). Dos escenarios con la MISMA reduced form (=20), pero distinto first stage:

(a) First stage = 0.50.

(b) First stage = 0.05.

Calcula el Wald en cada caso. ¿Qué problema anticipas con el escenario (b) si el 0.05 viene de una muestra chica (y por lo tanto es un número poco preciso)?

Solución

(a) Wald = $20/0.50 = 40$. (b) Wald = $20/0.05 = 400$.

Con first stage chico, **dividir entre casi cero amplifica todo**: si el 0.05 estimado en realidad fuera 0.04 o 0.06 por puro ruido de muestra, el Wald saltaría de 400 a 333 o 500. Un instrumento **débil** (first stage chico) hace que el estimador sea muy inestable, y también agranda cualquier violación pequeña de la exclusion restriction.

Ejercicio 1.11 (OLS vs. dos instrumentos distintos: ¿por qué no dan lo mismo? (adaptado de MIT HW10)). Angrist & Evans (1998) estiman el efecto de tener un tercer hijo sobre si la madre trabaja, usando dos instrumentos distintos (embarazo múltiple en el segundo parto, y que los dos primeros hijos sean del mismo sexo). Encuentran:

$$\hat{\alpha}^{IV\text{-mismo sexo}} > \hat{\alpha}^{OLS} > \hat{\alpha}^{IV\text{-embarazo múlt.}}$$

Si el efecto de tener un tercer hijo fuera EXACTAMENTE el mismo para todas las madres, los tres números deberían coincidir. ¿Qué explica que sean distintos?

Solución

Cada instrumento “empuja” a un grupo distinto de **compliers** a tener el tercer hijo: las madres cuyos dos primeros hijos son del mismo sexo NO son el mismo tipo de madres que las que tuvieron un embarazo múltiple. Si el efecto verdadero varía de persona a persona (efectos **heterogéneos**), cada IV estima el LATE de *su propio* grupo de compliers, y no hay razón para que coincidan entre sí ni con el OLS (que promedia sobre otra población distinta: quienes *eligieron* tener un tercer hijo).

Ejercicio 1.12 (Una historia sospechosa de exclusion restriction). En una escuela con clases superpobladas, algunos niños reciben al azar pastillas antiparasitarias (*deworming*). Se propone usar “estar en una escuela tratada” como instrumento para “estar desparasitado”, y así medir el efecto de la desparasitación sobre la asistencia escolar de CADA niño (incluso los que no recibieron la pastilla). ¿Qué podría romper la exclusion restriction en este caso?

Solución

Los parásitos son **contagiosos**: tratar a algunos niños de la escuela reduce el contagio de TODA la escuela, incluidos quienes no tomaron la pastilla. Entonces “estar en una

escuela tratada” sí podría afectar la asistencia de un niño por una vía distinta a su propio tratamiento (los efectos de *spillover* sobre sus compañeros) – eso viola la exclusion restriction si queremos leer el efecto como “efecto individual de tomar la pastilla”.

Nivel 4 – Reto

Difícil: integra varios temas en un mismo problema.

Ejercicio 1.13 (De Wald a 2SLS, en palabras). Con un solo instrumento binario, el Wald estimator es un cociente de dos números. Cuando hay **varios** instrumentos o controles adicionales, el método se generaliza y se llama **2SLS** (*two-stage least squares*). Describe, en tus propias palabras (SIN código), los dos pasos del 2SLS.

Solución

Primera etapa (first stage): regresionar la variable endógena X sobre el/los instrumento(s) (y los controles); guardar los valores **predichos** de X (la parte de X que viene “limpia”, explicada solo por el instrumento). **Segunda etapa (second stage):** regresionar Y sobre esos valores predichos (en vez de sobre el X original). En la práctica, en R esto se hace con **una sola línea** (p.ej. `ivreg(y ~ x | z)`), sin necesidad de programar las dos etapas a mano.

Nivel 5 – Reto+

Opcional, fuera del alcance típico del examen: para quien quiera ver "por qué funciona" la fórmula.

Ejercicio 1.14 (Wald = 2SLS con un instrumento binario (opcional)). Con un solo instrumento Z , el 2SLS y el Wald estimator dan **exactamente el mismo** número. Sigue estos pasos (MIT L21) para verlo:

$$\hat{\beta}_{2SLS} = \frac{\text{Cov}(Y, \hat{X})}{\text{Var}(\hat{X})}, \quad \hat{X} = \pi_0 + \pi_1 Z.$$

Sustituye $\hat{X}$ y simplifica hasta llegar a γ_1/π_1 (reduced form sobre first stage).

Solución

$$\hat{\beta}_{2SLS} = \frac{\text{Cov}(Y, \pi_0 + \pi_1 Z)}{\text{Var}(\pi_0 + \pi_1 Z)} = \frac{\pi_1 \text{Cov}(Y, Z)}{\pi_1^2 \text{Var}(Z)} = \frac{\text{Cov}(Y, Z)/\text{Var}(Z)}{\pi_1} = \frac{\gamma_1}{\pi_1}.$$

Como γ_1 (coeficiente de $Y \sim Z$) es la reduced form y π_1 (coeficiente de $X \sim Z$) es el first stage, esto es exactamente el Wald estimator. **No hace falta memorizar esta**

derivación: lo importante es que Wald y 2SLS son la misma idea, vista de dos formas.

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) y `dnackat/data-analysis-for-social-scientists-mitx` (Homework 10) —sin licencia explícita, usados 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