Tarea 4

João Marcos Garcia

2025-06-09

Ejercicio 1 · El Retorno Causal a la Educación

Una de las preguntas más fundamentales en economía laboral es el efecto causal de la educación sobre los ingresos. Una regresión OLS simple de salarios sobre años de escolaridad probablemente esté sesgada debido a factores de confusión no observados como la habilidad individual, motivación o antecedentes familiares. Los individuos con mayor habilidad innata pueden ser más propensos tanto a buscar más educación como a obtener salarios más altos, llevando a OLS a **sobrestimar** el verdadero retorno a la escolaridad.

David Card (1995) propuso usar la *proximidad a una universidad* como variable instrumental. La idea es que los individuos que crecieron cerca de una universidad de 4 años enfrentan menores costos de asistencia y por lo tanto tienen más probabilidades de matricularse.

Analizaremos esta estrategia usando los cuatro supuestos clave de IV. Sean:

- **Y** = logaritmo del salario por hora, `ln(wage)`
- **X** = años de escolaridad, `education`
- **Z** = 1 si el individuo creció cerca de una universidad de 4 años; 0 en caso contrario

Para cada supuesto a continuación, **(i)** establezca qué significa en este contexto y **(ii)** proporcione un argumento plausible de por qué podría ser violado.

1 · Exogeneidad

- ¿Qué debe ser cierto sobre la relación entre la proximidad universitaria *(Z)* y los determinantes no observados de los salarios U?
- Describa un escenario que viole la exogeneidad.

2 · Restricción de exclusión

- ¿A través de qué canales puede Z afectar los salarios *(Y)* bajo la hipótesis nula?
- Describa un escenario que viole la exclusión.

3 · Monotonicidad

- ¿Quiénes son los **desafiantes** descartados por este supuesto?
- Dé un ejemplo hipotético de un desafiante.

4 · Interpretación LATE

- Identifique a los **cumplimentadores** en este contexto.
- Explique por qué su retorno a la educación puede diferir del retorno promedio en la población.

Ejercicio 2 · Por Qué Importa la Monotonicidad

Este ejercicio muestra qué recupera el estimador IV con y sin desafiantes. Considere un instrumento de diseño de incentivo Z (ej. una carta informativa) que incentiva el tratamiento X (ej. vacunación), que a su vez afecta un resultado Y (ej. un índice de salud).

La población se divide en cuatro tipos latentes:

Grupo	$E[Y(0)]$	$E[Y(1)]$	δ
Cumplimentadores (C)	10	15	+5
Siempre-Tratados (ST)	8	12	+4
Nunca-Tratados (NT)	12	18	+6
Desafiantes (D)	5	7	+2

Parte 1 · Un Mundo Sin Desafiantes (La monotonicidad se cumple)

Participaciones poblacionales:

- Cumplimentadores 40% · Siempre-Tratados 20% · Nunca-Tratados 40% · Desafiantes 0%

a) Primera etapa

Calcule $E[X \mid Z = 1] - E[X \mid Z = 0]$.

b) Intención-de-Tratar (ITT)

Calcule $E[Y \mid Z = 1] - E[Y \mid Z = 0]$ usando la Ley de Expectativas Iteradas.

c) Estimación IV

Calcule $\widehat{\beta}_{IV} = \dfrac{\text{ITT}}{\text{Primera etapa}}$.

d) Interpretación

¿A qué parámetro corresponde la estimación en (c)? Sea explícito.

Parte 2 · Un Mundo Con Desafiantes (Monotonicidad violada)

Nuevas participaciones poblacionales:

- Cumplimentadores 40% · Siempre-Tratados 10% · Nunca-Tratados 30% · Desafiantes 20%

  Nota: Los efectos causales en la tabla anterior permanecen sin cambios.

e)-g)

Repita los pasos (a)–(c) para la nueva población.

h) Discusión

¿La nueva estimación IV identifica el efecto promedio del tratamiento para los cumplimentadores? Explique qué representa y por qué eso es un problema.

Ejercicio 3 · Oferta Laboral y Tamaño Familiar (Angrist & Evans 1998)

En clase discutimos cómo Angrist & Evans (1998) usan la composición sexual de los primeros hijos como instrumento para tener un tercer hijo. En este ejercicio replicarás dos tablas centrales del artículo usando datos PUMS de EE.UU. de 1990 y el paquete ***AER*** *en R. El objetivo es practicar limpieza de datos, regresiones ponderadas, y 2SLS en un contexto del mundo real.*

El artículo está disponible en http://piketty.pse.ens.fr/files/AngristEvans1998.pdf. Descarga los datos acompañantes (`AngEv98.zip`) de https://economics.mit.edu/sites/default/files/publications/AngEv98.zip y extrae el archivo **`m_d_903.sas7bdat`**.

A continuación se presenta **código inicial** que carga el conjunto de datos y lo limpia. Los comentarios en el código explican cada paso para que entiendas *qué* se está filtrando y *por qué*.

```
library(haven)       # Leer SAS
library(tidyverse)   # verbos dplyr + ggplot2
library(AER)         # ivreg()

# 1. Leer extracto PUMS crudo (1990, mujeres casadas 21-35 con mas que 2 hijos)
file_path <- file.path("/Users",
                       "joaomarcosgarcia",
                       "Downloads",
                       "AngEv98",
                       "m_d_903.sas7bdat")  # <-- ajustar si es necesario

pums_data <- read_sas(file_path)

# 2. Convertir variables etiquetadas de SAS seleccionadas a numéricas por conveniencia
vars_to_convert <- c("AGEM", "KIDCOUNT", "AGE2NDK", "AAGE", "AAGE2ND",
                     "ASEX", "ASEX2ND", "FERTIL", "AGEK", "SEXK", "SEX2NDK",
                     "WEEK89M", "HOUR89M", "INCOMEM1", "FAMINC")

pums_data <- pums_data %>%
  mutate(across(all_of(vars_to_convert), as.numeric))

# 3. Reproducir restricciones de muestra de Angrist-&-Evans
pums_data <- pums_data %>%
  filter(
    AGEM >= 21 & AGEM <= 35,    # Mujeres de 21-35 años
    KIDCOUNT >= 2,              # Al menos dos hijos
    AGE2NDK >= 1,               # Segundo hijo de al menos 1 año
    AAGE == 0 & AAGE2ND == 0,   # Edad del esposo no codificada al máximo
    ASEX == 0 & ASEX2ND == 0,   # Sexo de los primeros dos hijos conocido
    PWGTM1 > 0                  # Peso personal positivo
  ) %>%
  mutate(
    AGEFSTM   = AGEM - AGEK,                    # Edad al primer nacimiento
    SAMESEX   = as.numeric(SEXK == SEX2NDK),    # Instrumento (1=mismo sexo)
    MOREKIDS  = as.numeric(KIDCOUNT > 2),       # Tratamiento (tiene mas que 2 hijos)
    WORKEDFPAY= ifelse(WEEK89M > 0, 1, 0),      # Trabajó por pago en 1989
    LNFAMILYINC = log(FAMINC),                  # Log ingreso familiar
    LABORINCOME = INCOMEM1,                     # Ingreso laboral de la mujer
    # Demografía (HISPM como carácter es correcto en este archivo)
    HISP       = ifelse(HISPM != "000", 1, 0),
```

```r
  BLACK      = ifelse(RACEM == "002" & HISP == 0, 1, 0),
  OTHERRACE  = ifelse(!RACEM %in% c("001", "002") & HISP == 0, 1, 0)
) %>%
filter(AGEFSTM >= 15) %>%                    # Excluir nacimientos implausiblemente tempranos
# Mantener solo variables usadas en la replicación
select(KIDCOUNT, MOREKIDS, SAMESEX, PWGTM1, AGEM, AGEFSTM,
       WORKEDFPAY, WEEK89M, HOUR89M, SEXK, SEX2NDK,
       LNFAMILYINC, LABORINCOME, HISP, BLACK, OTHERRACE)
```

> **¿Por qué estos pasos?** Imitamos el marco muestral del artículo (mujeres casadas en edad reproductiva con al menos dos hijos) y construimos variables para el instrumento (`SAMESEX`), tratamiento (`MOREKIDS`), resultados (horas trabajadas `HOUR89M`, indicador `WORKEDFPAY`, log ingreso), y controles demográficos.

a) Reproduciendo Tabla 3

Estima, vía una regresión ponderada (`weights = PWGTM1`), la fracción de mujeres que tienen un tercer hijo **(MOREKIDS)** después de:

1. Un primogénito niño y niña (`SAMESEX = 0`), y
2. Dos hijos del mismo sexo (`SAMESEX = 1`).

Reporta la diferencia entre las dos fracciones. *Pista:* una regresión solo con intercepto con variables dummy de `SAMESEX` es suficiente.

b) Reproduciendo Tabla 7, columnas (1) y (2)

Vamos reproducir la Tabla 7, columnas (1) y (2).

Ejecuta:

1. **OLS** de los resultados sobre `MOREKIDS` más controles.
2. **2SLS** (usando `SAMESEX` como instrumento) del mismo resultado.

Resultados:

- Worked for pay, Weeks worked, Hours/week, Labor Income, Log(family income)

Controles (incluir en ambas etapas):

- Edad de la mujer (`AGEM`)
- Edad al primer nacimiento (`AGEFSTM`)
- Sexo del primer hijo (`SEXK`) y del segundo hijo (`SEX2NDK`)
- Hispana (`HISP`), Negra (`BLACK`), Otra raza (`OTHERRACE`)

Recuerda aplicar pesos personales. Usa `lm()` para OLS e `ivreg()` para 2SLS. No se requiere igualdad exacta con el artículo; las magnitudes deben ser similares.

c) Interpretación

Discute los signos y magnitudes de ambos coeficientes. ¿Qué implican sobre el efecto de un hijo adicional en el ingreso familiar de las madres?

d) Sesgo de variable omitida

Asumiendo que el instrumento es válido, ¿qué te dice la diferencia entre los coeficientes OLS y 2SLS sobre el **signo** del sesgo en la estimación OLS? Explica cuidadosamente.
