Ejercicio En Clase

Joao Garcia

2025-06-04

Replicación de Angrist y Krueger

El objetivo de este ejercício es de replicar y analisar los resultados del artículo **Does Compulsory Schooling Affect School Attendance and Earnings?**. El paper está disponible aquí.

Vamos a enfocarnos en el grupo nacido entre 1930 y 1939. Vamos a reproducir las figuras II y III, y algunos números de la Tabla V.

Cargar y Limpiar los Datos

Su primera tarea es descargar los datos de este enlace:

https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive

Encuentre el archivo llamado **NEW7080.rar** en la sección de **Angrist and Krueger (1991)**. Descárguelo y descomprímalo.

A continuación, siga este modelo para limpiar los datos.

```
library(tidyverse)
library(haven) # Para leer los datos que vienen de Stata
library(AER) # Para usar ivreg

file_path <- "/Users/joaomarcosgarcia/Downloads/NEW7080.dta" # Cambie por la ruta en su computador
ak_data <- read_dta(file_path)
```

Primero vamos nombrar las variables. Después, vamos enfocarnos en el grupo nacido entre 1930 y 1939.

```
ak_data <- ak_data %>%
  rename(
    AGE = v1, AGEQ = v2,
    EDUC = v4,
    ENOCENT = v5, ESOCENT = v6,
    LWKLYWGE = v9,
    MARRIED = v10,
    MIDATL = v11,
    MT = v12,
    NEWENG = v13,
    CENSUS = v16,
    QOB = v18,
    RACE = v19,
    SMSA = v20,
    SOATL = v21,
    WNOCENT = v24,
    WSOCENT = v25,
    YOB = v27
```

```
  )

ak_data <- ak_data %>%
  filter(YOB >= 30 & YOB <= 39)

# Creación de variables
ak_data <- ak_data %>%
  mutate(
    AGEQ = ifelse(CENSUS == 80, AGEQ - 1900, AGEQ),
    AGEQSQ = AGEQ * AGEQ
  )
```

Reproduzir las Figuras

Vamos empezar con las figuras. Tenemos que calcular la educación y los salarios medios para cada trimestre de nacimiento

```
# Prepare data for graphing: mean EDUC and LWKLYWGE by Year-Quarter of Birth
graph_data <- ak_data %>%
  mutate(
    YQOB_plot = 1900 + YOB + (QOB - 1) / 4 # Create a continuous Year.Quarter variable for plotting
  ) %>%
  group_by(YQOB_plot, QOB) %>% # Group by the plotting variable and actual QOB for labels
  summarize(
    mean_educ = mean(EDUC, na.rm = TRUE),
    mean_wage = mean(LWKLYWGE, na.rm = TRUE),
    .groups = 'drop' # Ungroup after summarizing
  ) %>%
  arrange(YQOB_plot) # Ensure data is sorted for line plots
```

Desde aquí, puede utilizar graph_data para crear las figuras.

Para ayudarle, debería ser algo como:

```
# ggplot(graph_data, aes(#########)) +
#  geom_line() +
#  geom_point() +
#  geom_text(aes(label = QOB), nudge_y = 0.05)
```

Reproduzir los Resultados

Vamos empezar con el modelo MCO en la columna (1) de la Tabla V. Para incluir varias dummies de año de nacimiento, vamos utilizar as.factor(YOB). Esto crea una dummy para cada valor de la variable YOB.

```
formula <- as.formula("LWKLYWGE ~ EDUC + as.factor(YOB)")

res <- lm(formula, data=ak_data)

summary(res)
```

```
##
## Call:
## lm(formula = formula, data = ak_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.7490 -0.2354  0.0726  0.3378  4.6448
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.017348   0.005471 917.144  < 2e-16 ***
## EDUC               0.071081   0.000339 209.674  < 2e-16 ***
## as.factor(YOB)31 -0.006387   0.005039  -1.268 0.204971
## as.factor(YOB)32 -0.014838   0.004972  -2.984 0.002844 **
## as.factor(YOB)33 -0.017583   0.005032  -3.494 0.000476 ***
## as.factor(YOB)34 -0.020999   0.004985  -4.213 2.52e-05 ***
## as.factor(YOB)35 -0.032895   0.004952  -6.643 3.07e-11 ***
## as.factor(YOB)36 -0.031781   0.004956  -6.413 1.43e-10 ***
## as.factor(YOB)37 -0.036712   0.004908  -7.480 7.47e-14 ***
## as.factor(YOB)38 -0.036890   0.004866  -7.582 3.42e-14 ***
## as.factor(YOB)39 -0.048164   0.004847  -9.937  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6376 on 329498 degrees of freedom
## Multiple R-squared:  0.1177, Adjusted R-squared:  0.1177
## F-statistic:  4397 on 10 and 329498 DF,  p-value: < 2.2e-16
```

A continuación, vamos utilizar la lógica de Mínimos Cuadrados Dos Etapas para reproducir la columna 2. Primero vamos hacer una regresión de la variable endógena, EDUC, sobre YOB YOB interactuando con QOB. Todavía es necesario usar as.factor para crear varias dummies.

```
# Primera etapa:
formula <- as.formula("EDUC ~ as.factor(YOB):as.factor(QOB)")

PE <- lm(formula, data=ak_data)
```

Ahora generamos el valor predicho de EDUC:

```
ak_data$EDUC_hat <- predict(PE)
```

Su tarea es ejecutar la segunda etapa. El coeficiente estimado debería ser 0.0891, como en la coluna 2 del paper.

Usando IVREG

Cuando lo logre, observe que el coeficiente es el mismo, pero el eror estándar es un poco diferente. Esto es porque cuando hacemos 2SLS de esa manera, estamos ignorando la incertidumbre de en la primera etapa. En este caso, la diferencia es mínima, porque tenemos muchos datos, pero a veces puede ser grande.

Cuando estimamos los efectos con el comando ivreg del paquete AER, los errores son correctos.

Es similar a una regresión con lm, pero tenemos que especificar las dos etapas. Para esto, utilizamos el símbolo "|". Por ejemplo: ivreg(y~x1+c1+c2 | z1+z2+c1+c2, data=data).

Utilice este comando para replicar la columna 2 y verifique que los errores estándar son corectos.

Para finalizar, reproduzca las otras columnas de la Tabla V. Para ayudarle: las columnas 5 y 6 incluyen "8-region of residence dummies". Ellas son: MIDATL, MT, NEWENG, SOATL, WNOCENT, WSOCENT, ENOCENT y ESOCENT.