
Universidad de
SanAndrés

Econometría Avanzada

Walter Sosa Escudero
Gastón García Zavaleta

# Trabajo Práctico 2

García Vassallo, Hedemann, Iolster, Sury

**2022**

# Ejercicio 1

Para estimar $\delta_2$ usando 2SLS seguimos los siguientes pasos. En primer lugar, usaremos como instrumento de $x_1$ una combinación lineal de las variables contenidas en z, descartando $z_1$. Esto es así ya que $z_1$ no puede ser instrumento de $x_1$ dado que viola el supuesto de exogeneidad al ser un determinante directo de $y_1$. De este modo, al combinar todas las variables que pertenecen a z, logramos un instrumento válido.
Luego, en una primera etapa, regresamos la variable endógena $x_1$ en el instrumento y en $z_1$ ya que esta es una variable exógena en el modelo para $y$. De esta forma obtenemos $X_1^*$, la mejor predicción que se puede hacer de $X_1$ en base al instrumento y $z_1$. Así, $X_1^*$ cae en el Span de Z y como el Span de Z no esta correlacionado con $\mu_1$, $X_1^*$ tampoco correlaciona con $\mu_1$ tampoco. Por lo tanto, obtenemos una proyección de $X_1$ que es exógena. Además, trivialmente de la primera etapa para $z_1$ se obtiene $z_1^* = z_1$ ya que $z_1$ es exógena.
Por último, la segunda etapa consiste en regresar el modelo original por MCO pero usando $X_1^*$ en vez de $X_1$. De este modo, el estimador de $\delta_2$ es $\hat{\delta}_2 = (X_1^{*\prime} X_1^*)^{-1} X_1^{*\prime} Y^*$

# Ejercicio 2

### A)

Si, podemos construir un estimador consistente para $\beta_1$ con los datos que tenemos. La forma de encontrar $\beta_1$ sería dividiendo $\hat{\omega}_{yz} \div \hat{\omega}_{xz}$.

$$\frac{\frac{\widehat{Cov(y,z)}}{\widehat{V(z)}}}{\frac{\widehat{Cov(x,z)}}{\widehat{V(z)}}}$$

Donde las varianzas se cancelan y nos queda:

$$\frac{\widehat{Cov(y,z)}}{\widehat{Cov(x,z)}}$$

Se puede escribir $\beta$ en términos de los momentos poblacionales que se pueden estimar con los datos de la muestra:

$$Cov(z,y) = \beta Cov(z,x) + Cov(z,u)$$

Y dado que nos dijeron que se cumple $Cov(u,z) = 0$ y $Cov(x,z) \neq 0$:

$$\beta = \frac{Cov(y,z)}{Cov(x,z)}$$

Que retomando y desarrollando lo anterior es igual a:

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})}$$

Y por lo tanto vale:

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{n} z_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} z_i y_i\right) = (Z'X)^{-1} Z'Y$$

Para que el estimador sea consistente queremos ver que:

$$\hat{\beta}_{IV} = (z'x)^{-1} z'y \xrightarrow{p} \beta$$

$$\hat{\beta}_{IV} = (z'x)^{-1} z'(x\beta + u)$$

$$\hat{\beta}_{IV} = \beta + (z'x)^{-1} z'\mu$$

Sabiendo que:

$$\beta \xrightarrow{p} \beta$$

Por exogeneidad sabemos que:

$$z'\mu \xrightarrow{p} 0$$

Por otro lado:

$$(z'x)^{-1} \xrightarrow{p} \Sigma_{zx} < \infty$$

Esto es, $(z'x)^{-1}$ tiende en probabilidad a algo finito dado que:

$$(z'x)z'\mu$$

$$\left(\frac{z'x}{n}\right)^{-1} \frac{z'\mu}{n}$$

$$\left(\frac{z'x}{n}\right)^{-1} = \frac{\sum_{i=1}^{n} z_i x_i}{n}$$

Esto es un vector y para poder aplicar ley de grandes números tiene que ser un número.

El h-ésimo elemento es:

$$\frac{\sum_{i=1}^{n} z_{hi} x_{ji}}{n} \xrightarrow{p} E(zx) = \Sigma_{zx}$$

Esto es así si se cumple que $z_i x_i$ son iid por muestra aleatorio y también si se cumple que $E(z_{hi}x_{ji}) = \Sigma_{zx_{hj}} < \infty$

Entonces

$$(z'x)^{-1}z'\mu \xrightarrow{p} 0$$

Por lo que:

$$\beta + (z'x)^{-1}z'\mu \xrightarrow{p} \beta$$

**B)**

Si nos facilitaran la base de datos lo que haríamos seria estimar MCO en dos etapas. Para esto, necesitamos un instrumento que sea relevante $E(Z'X) \neq 0$ y exógeno $E(Z'\mu) = 0$. Entonces, en la primera etapa lo que se hace es regresar la X, la variable endógena, contra el instrumento al cual llamaremos Z y las variables control. Esperamos que el estimador sea distinto de cero para que haya una relación relevante entre el instrumento y la variable a instrumentar. De este resultado de la regresión de la primera etapa nosotros lo que queremos es quedarnos con la proyección ortogonal la cual llamaremos $\hat{X}$ que es la mejor predicción de X dado Z, es toda la variabilidad de X contenida en Z. Es decir, parte de la variabilidad de X no esta correlacionada con el error, y es esta variabilidad la que queremos capturar con el instrumento. En esta instancia es cuando podemos chequear el supuesto de relevancia dado que lo que necesitamos es que los coeficientes asociados a Z sean significativos. Luego de esto, hacemos una segunda etapa donde regresamos Y, la variable explicada, contra $\hat{X}$ para obtener los $\hat{\beta}$.

La demostración de consistencia del estimador de variables instrumentales se encuentra en el punto anterior.

# Ejercicio 3

**A)**

| | **Promedio** | **Mínimo** | **máximo** |
|---|---|---|---|
| age | 38.48743 | 0 | 50 |
| ageq | 38.86205 | 30.25 | 50 |
| ageqsq | 1544.237 | 915.0625 | 2500 |
| educ | 13.24814 | 0 | 20 |
| lwage | 5.838823 | -2.341806 | 11.22524 |
| married | 0.8277217 | 0 | 1 |
| qob | 2.524654 | 1 | 4 |
| race | 0.08165745 | 0 | 1 |
| smsa | 0.1858311 | 0 | 1 |
| yob | 40.75679 | 30 | 49 |

En el cuadro anterior podemos analizar las estadísticas descriptivas de distintas variables. En primer lugar, en la primera columna observamos los promedios de las variables. Podemos observar que la edad promedio es 38 años, la edad medida en trimestres (ageq) es casi 39 años, la edad medida en trimestres al cuadrado (ageqsq) es 1544 años, en cuanto a la educación la cantidad de años promedio completos (educ) es 13, el logaritmo del salario mensual (lwage) promedio es 5.838823, el año promedio de nacimiento (yob) es el 40.

Por otro lado, podemos observar que la edad tiene un mínimo en 0 y un máximo en 50, la edad media en trimestres tiene un mínimo en 30 y un máximo en 50, la edad medida en trimestres al cuadrado tiene un mínimo en 915 y un máximo en 2500, el número de años completos de educación tiene un mínimo en 0 y un máximo en 20, el logaritmo del salario mensual tiene un mínimo en -2.342806 y un máximo en 11.22524, el año de nacimiento más bajo es el 30 y el máximo es el 49.

Por otro lado, en la siguiente tabla podemos observar el número promedio de años de escolaridad completa para cada "*quarter of birth*" para los hombres nacidos en las décadas del 30 y 40.

| **qob** | **educ** |
|---|---|
| 1 | 13.17287 |
| 2 | 13.22884 |
| 3 | 13.27137 |
| 4 | 13.31496 |

Como primera observación se puede apreciar poca diferencia entre los distintos quarters, aunque si se ve un aumento escalonado a medida que se avanza entre los trimestres. Podemos decir que los que nacieron en el cuarto trimestre del año son los que tienen un mayor promedio de años de escolaridad completa. En segundo lugar, se encuentran los del tercer trimestre, luego los del segundo trimestre y luego los del primer trimestre. Entonces, podemos concluir que nacer a fines de año tiene un impacto positivo en los años de escolaridad completa.

**B)**

| | *Dependent variable:* |
|---|---|
| | Sales |
| educ | $0.064^{***}$ |
| | (0.0003) |
| ageq | $-0.005$ |
| | (0.013) |
| ageqsq | 0.0001 |
| | (0.0001) |
| race | $-0.269^{***}$ |
| | (0.004) |
| smsa | $-0.195^{***}$ |
| | (0.003) |
| married | $0.243^{***}$ |
| | (0.003) |
| Constant | $4.968^{***}$ |
| | (0.293) |
| Observations | 329,509 |
| $R^2$ | 0.157 |
| Adjusted $R^2$ | 0.157 |
| Residual Std. Error | 0.623 (df = 329502) |
| F Statistic | $10{,}211.540^{***}$ (df = 6; 329502) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Estimamos el modelo lineal bajo los supuestos clásicos y obtuvimos los resultados que se pueden observar en el cuadro anterior.

Encontramos que hay varias variables que son estadísticamente significativas sobre el logaritmo del salario mensual. Tanto la educación como la raza que es una dummy que toma valor 1 si la persona es negra, la dummy de ciudad central que toma valor 1 si la persona reside en una ciudad central y la dummy de matrimonio que toma valor 1 si la persona esta casada son significativas al 1 %.

Entonces, analizando las variables podemos decir que cuando la variable educación y la dummy de matrimonio aumentan, la variable explicada también aumenta en promedio y ceteris paribus dado que los coeficientes asociados a la variables son 0.064 y 0.243 respectivamente. Por otro lado, tanto la raza como dummy de ciudad central toman valores negativos por lo que cuando estas toman valor 1, el logaritmo del salario mensual disminuye en promedio y ceteris paribus. Los coeficientes asociados a estas variables son -0.269 y -0.195 respectivamente.

Por último, el $R^2$ toma un valor de 0.157, esto implica que la variabilidad de las X explican un 15.7 % la variabilidad del logaritmo del salario mensual.

**C)**

La variable *quarter of birth* podría ser una buena variable instrumental para la educación dado que no está directamente relacionada con salarios pero correlaciona positivamente con la edad. La covarianza entre los residuos del modelo lineal y la variable instrumental es 0.00, por lo que se cumple exogeneidad. Por otro lado, la covarianza entre educacion y el trimestre de nacimiento es 0.06. Esta relación positiva podría darse por distintas razones, una de ellas podría ser que las madres hayan pensado estratégicamente el trimestre en el que quisieran que sus hijos nazcan para extender la licencia de maternidad y estar más presentes en la crianza. Esto tiene

distintos análisis, por un lado podemos decir que la madre al pensarlo estratégicamente es más inteligente en promedio por lo que esta inteligencia también la heredarían sus hijos, por otro lado, podemos decir que la madre al estar más presente en la crianza de su hijo en promedio, este va a tener un mayor desarrollo cognitivo.

**D)**

La resolución de este punto se encuentra en el script.

**E)**

| | *Dependent variable:* |
|---|---|
| | lmm |
| z1 | $-0.151^{***}$ |
| | (0.016) |
| z2 | $-0.095^{***}$ |
| | (0.016) |
| z3 | $-0.034^{**}$ |
| | (0.016) |
| Constant | $12.839^{***}$ |
| | (0.012) |
| Observations | 329,509 |
| $R^2$ | 0.0003 |
| Adjusted $R^2$ | 0.0003 |
| Residual Std. Error | 3.281 (df = 329505) |
| F Statistic | $34.009^{***}$ (df = 3; 329505) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Podemos observar que las 3 variables ($Z_1,\ Z_2,\ Z_3$ son significativas al 1 % por lo que podemos decir que tienen efecto sobre la educación.

**F)**

Estimamos el first stage del modelo MCO en 2 etapas y obtuvimos los resultados que se puede observar en el siguiente cuadro.

| | *Dependent variable:* |
|---|---|
| | First |
| ageq | 0.051 |
| | (0.067) |
| ageqsq | −0.001* |
| | (0.001) |
| race | −1.709*** |
| | (0.021) |
| married | 0.157*** |
| | (0.016) |
| smsa | −1.169*** |
| | (0.014) |
| z1 | −0.093*** |
| | (0.016) |
| z2 | −0.056*** |
| | (0.016) |
| z3 | −0.024 |
| | (0.016) |
| Constant | 13.327*** |
| | (1.510) |
| Observations | 329,509 |
| $R^2$ | 0.042 |
| Adjusted $R^2$ | 0.042 |
| Residual Std. Error | 3.212 (df = 329500) |
| F Statistic | 1,785.255*** (df = 8; 329500) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Encontramos que hay varias variables con un efecto estadísticamente significativo en niveles habituales sobre el logaritmo del salario mensual. Tanto la raza (dummy que toma valor 1 cuando la persona es de color), como la variable married (dummy que toma valor 1 cuando la persona esta casada), la dummy de ciudad central (toma valor 1 cuando la persona vive en una ciudad central), como $Z_1$ (dummy que representa a los que nacieron en el primer trimestre) y $Z_2$ (dummy que representa a los que nacieron en el segundo trimestre) son significativas al 1 %. Por otro lado, la edad medida en trimestres al cuadrado es significativa al 10 %.

Entonces, analizando las variables significativas podemos decir que todas menos la dummy de matrimonio tienen un efecto negativo sobre el logaritmo del salario mensual. Esto es, cuando las vairbles ageqsq, raza, smsa, $Z_1$ y $Z_2$ toma valor 1, la variable dependiente disminuye en promedio y ceteris paribus dado que los coeficientes asociado son -1.709, -0.001, -1.169, -0.093 y -0.056 respectivamente. Luego, cuando la variable matrimonio toma valor 1 el logaritmo del salario mensual aumenta en promedio y ceteris paribus dado que el coeficiente asociado a esta variables es 0.157. -

Por último, el $R^2$ toma un valor de 0.042, esto implica que la variabilidad de las variables explicativas del modelo explican un 4.2 % la variabilidad del logaritmo del salario mensual.

**G)**

Una variable instrumental es débil cuando no está suficientemente correlacionada con la variable endógena, imposibilitando la corrección de endogeneidad. Esto puede resultar en una mayor varianza de los coeficientes y un mayor sesgo para muestras finitas. Para analizar si nuestros instrumentos son débiles o no, lo que hacemos es observar el estadístico F de la primea etapa de la regresión. Como F = 1785,2 concluimos que no son instrumentos débiles.

**H)**

Estimamos el modelo MCO en 2 etapas y obtuvimos los resultados que se puede observar en el siguiente cuadro.

| | *Dependent variable:* |
|---|---|
| | iv |
| educ | $0.138^{***}$ |
| | (0.034) |
| ageq | $-0.009$ |
| | (0.014) |
| ageqsq | 0.0002 |
| | (0.0002) |
| race | $-0.143^{**}$ |
| | (0.057) |
| smsa | $-0.109^{***}$ |
| | (0.039) |
| married | $0.231^{***}$ |
| | (0.006) |
| Constant | $3.983^{***}$ |
| | (0.547) |
| Observations | 329,509 |
| $R^2$ | 0.035 |
| Adjusted $R^2$ | 0.035 |
| Residual Std. Error | 0.667 (df = 329502) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Al observar los retornos a la educación por MC2E encontramos que hay variables estadísticamente significativas tanto al 1 % como al 5 %. Las variables estadísticamente significativas al 1 % son el número de años completos de educación (educ), la dummy de ciudad central y la dummy de matrimonio (married). De la variable educación podemos interpretar que cuando esta sube en 1 año, el logaritmo del salario mensual aumenta aproximadamente en 14.8 % ($100[e^{0,138}-1]$) en promedio y ceteris paribus. Por otro lado, en cuanto a la dummy de ciudad central podemos interpretar que cuando toma valor 1, si la persona reside en una ciudad central, el logaritmo del salario mensual disminuye en promedio y ceteris paribus dado que el coeficiente asociado a esta variable es -0.109. Luego, en cuanto a la variable dummy podemos interpretar que cuando ésta toma valor 1, cuando la persona está casada, el logaritmo del salario mensual aumenta aproximadamente 26 % ($100[e^{0,231}-1]$) en promedio y ceteris paribus.

Luego, la única variable significativa al 5 % es la raza. Esta es una dummy que toma valor 1 cuando la persona es negra, cuando esto sucede el logaritmo del salario mensual disminuye en promedio y ceteris paribus dado que el coeficiente asociado a la raza es -0.143.

Por último, podemos observar que el $R^2$ es 0.035. Esto implica que la variabilidad de las variables explicativas del modelo explican un 3,5 % la variabilidad del logaritmo del salario mensual.

En comparación con los resultados obtenidos en el punto **b)** podemos decir que el modelo lineal ajusta mejor dado que el $R^2$ es 0,122 puntos más grande. Por lo que en el modelo lineal la variabilidad de las variables explicativas explican mejor la variabilidad del logaritmo del salario mensual. Por otro lado, podemos observar que las variables que son estadísticamente significativas en el modelo lineal, también lo son en el modelo de variables instrumentales. La única diferencia en cuanto a significatividad es que la dummy raza en el modelo lineal es significativa al 1 % mientras que en este caso es significativa al 5 %. Cabe resaltar que los signos de los coeficientes asociados a las variables se mantienen en ambos modelos.

```
#         TP3
setwd("~/Desktop/Econometria Tutorial /TPs/TP3")

library(haven) # para abrir bases de datos en formato .dta
library(stargazer) # para tablas
library(tidyverse) # librería muy usada para manipular los datos y
hacer gráficos
library(AER) # para usar el comando de variables instrumentales

data <- read_dta("qob.dta")

#A)

promedios <- data %>% summarise_all(mean)
promedios
minimos <- data %>% summarise_all(min)
minimos
maximos <- data %>% summarise_all(max)
maximos

#Como el minimo año de nacimiento es 30 y el maximo es 49 no
eliminamos nada
aggregate(data$educ, by=list(data$qob), mean)


#B)

data2 <- subset(data, yob>= 30 & yob <= 39)


attach(data2)

lm.fit <- lm(lwage ~ educ + ageq + ageqsq + race +smsa + married)
summary(lm.fit)
stargazer(lm.fit,
          type="latex",
          dep.var.labels=c("Sales"),
          out="lm.fit")

#C)

residuos = lm.fit$residuals
cov(residuos, data2$qob)
cov(data2$educ, data2$qob)

#D)

data2$z1 <- ifelse(data2$qob==1,1,0)
data2$z2 <- ifelse(data2$qob==2,1,0)
data2$z3 <- ifelse(data2$qob==3,1,0)

#E)

cor(data2$educ, data2$z1)
```

```
cor(data2$educ, data2$z2)
cor(data2$educ, data2$z3)

attach(data2)
lm.fit2 = lm(educ ~ z1 + z2 + z3)
summary(lm.fit2)
stargazer(lm.fit2,
          type="latex",
          dep.var.labels=c("lmm"),
          out="lm.fit2")


#F)
#First stage:
#regresion de la variable endogena en los instrumentos y en las
variables del modelo:
attach(data2)
first.stage <- lm(educ ~ ageq + ageqsq + race + married + smsa + z1
+ z2 + z3)
summary(first.stage)
stargazer(first.stage,
          type="latex",
          dep.var.labels=c("First"),
          out="first.stage")
data2$lwage_hat <- first.stage$fitted.values #me quedo con la
proyeccion ortogonal

#G)
var.test(educ,z1, alternative="two.sided")
var.test(educ,z2, alternative="two.sided")
var.test(educ,z3, alternative="two.sided")

#H)
iv.fit <- ivreg(lwage ~ educ + ageq + ageqsq + race +smsa + married
| #barra y del lado derecho pongo los instrumentos
                  ageq + ageqsq + race +smsa + married  + z1 + z2 +
z3, # instrumentos (cada variable exógena es su propio instrumento)
                data = data2)
summary(iv.fit)
stargazer(iv.fit,
          type="latex",
          dep.var.labels=c("iv"),
          out="iv.fit")
```