Variables Instrumentales

Maestría en Economía Uninorte

Carlos Andrés Yanes

2024-08-24

Preambulo

Los modelos de **variables instrumentales** (IV) son una técnica econométrica crucial cuando se enfrentan problemas de endogeneidad en la estimación de relaciones causales. En la evaluación de impacto, la endogeneidad surge cuando una o más variables explicativas están correlacionadas con el término de error, lo que sesga los estimadores y dificulta identificar efectos causales verdaderos.

¿Qué es la endogeneidad?

La endogeneidad puede surgir por varias razones, como la presencia de variables omitidas, la simultaneidad, o el error de medición. Por ejemplo, al evaluar el impacto de la educación en los ingresos, es posible que factores no observados como la habilidad innata o el entorno familiar influyan tanto en la educación como en los ingresos, generando una correlación entre la educación y el término de error.

¿Qué son las Variables Instrumentales?

Las variables instrumentales se utilizan para resolver este problema. Una variable instrumental es una variable que está correlacionada con la variable explicativa endógena pero que no está correlacionada con el término de error. En otras palabras,

el instrumento afecta el resultado solo a través de su impacto en la variable endógena.

Supongamos que estamos interesados en medir el impacto de una intervención educativa en los salarios futuros. Si los individuos que eligen participar en la intervención son aquellos más motivados o con más recursos, la simple estimación de una regresión lineal podría sobreestimar el impacto de la intervención debido a la selección no aleatoria. Aquí es donde un buen instrumento es vital.

Si encontramos un instrumento válido, como una política de expansión educativa que no está directamente relacionada con los salarios futuros, podemos utilizarlo para estimar el impacto causal de la educación en los ingresos, aislando el efecto de la endogeneidad.

> Cuidado
>
> La base de datos para este modulo solo será enviada por correo electronico a los estudiantes del curso de Econometría de la Universidad del Norte

Datos

- Fuente original es de Jeffrey R. Kling (2001).
- El artículo se llama: "Interpreting Instrumental Variables Estimates of the Returns to Schooling". Journal of Business and Economic Statistics, 19, 358-364.
- Tiene que ver con los retornos de la educación[1]

[1] Hemos insistido en los temas de educación ya que existe una amplia literatura en la implementación de este tipo de **instrumentos** para una correcta estimación

Limpiando el Environment de R

```
rm(list = ls())
```

Preparación del entorno para ejecución

Primero que nada preparar los paquetes que se van a usar para realizar el ejercicio. Estos permitiran usar las `funciones` para los cálculos pertinentes

```
library(pacman)
p_load(lmtest, foreign, haven, tidyverse, stargazer, dplyr, estimatr, ggplot2, sandwich)
```

Cargar la base de datos

Del archivo proporcionado y utilizando a `haven` procedemos a importar la base

```
base.ed <- read_dta("Returnseducational.dta")
head(base.ed)
```

```
# A tibble: 6 x 101
     id black imigrnt hhead mag_14 news_14 lib_14 num_sib fgrade mgrade    iq
  <dbl> <dbl>   <dbl> <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>  <dbl> <dbl>
1     2     1       0     1      0       1      0       7     NA     NA    NA
2     3     0       0     1      1       1      1       1      8      8    93
3     4     0       0     1      1       1      1       2     14     12   103
4     5     0       0     1      0       1      1      NA     11     12    88
5     6     0       0     1      0       1      0       0      8      7   108
6     7     0       0     1      1       1      1      NA      9     12    85
# i 90 more variables: bdate <dbl>, gfill76 <dbl>, wt76 <dbl>, grade76 <dbl>,
#   grade66 <dbl>, age66 <dbl>, smsa66 <dbl>, region <dbl>, smsa76 <dbl>,
#   col4 <dbl>, mcol4 <dbl>, col4pub <dbl>, south76 <dbl>, wage76 <dbl>,
#   exp76 <dbl>, expsq76 <dbl>, age76 <dbl>, agesq76 <dbl>, reg1 <dbl>,
#   reg2 <dbl>, reg3 <dbl>, reg4 <dbl>, reg5 <dbl>, reg6 <dbl>, reg7 <dbl>,
#   reg8 <dbl>, reg9 <dbl>, momdad14 <dbl>, sinmom14 <dbl>, nodaded <dbl>,
#   nomomed <dbl>, daded <dbl>, momed <dbl>, famed <dbl>, famed1 <dbl>, ...
```

Etiquetas

La base de datos de acuerdo a las columnas de datos podemos decir de cada una de ellas lo siguiente:

Variable	Tipo	Etiqueta de la Variable
wage76	float	Salario en el '76
grade76	float	Nivel educativo en el '76
col4	float	Si hay alguna universidad de 4 años cerca
age76	float	Edad en el '76 (edad66 + 10)

Estadistica

```
summary(base.ed[c("wage76", "grade76", "col4", "age76")])
```

```
     wage76          grade76           col4            age76      
 Min.   :0.000   Min.   : 1.00   Min.   :0.0000   Min.   :24.00  
 1st Qu.:1.372   1st Qu.:12.00   1st Qu.:0.0000   1st Qu.:25.00  
 Median :1.682   Median :13.00   Median :1.0000   Median :28.00  
 Mean   :1.657   Mean   :13.26   Mean   :0.6821   Mean   :28.12  
 3rd Qu.:1.958   3rd Qu.:16.00   3rd Qu.:1.0000   3rd Qu.:31.00  
 Max.   :3.180   Max.   :18.00   Max.   :1.0000   Max.   :34.00  
```

Correlación

Por un momento miremos la correlación que existe entre este par de variables. El **objetivo** es mirar si existe relación entre ellas, es una especie de tener presente que no vayamos a tener Este es un texto normal y multicolinealidad en el modelo.

```
# Correlación
cor(base.ed$grade76, base.ed$col4)
```

```
[1] 0.1442402
```

Primer modelo de estimación

Vamos a mirar el resultado de la estimación

```r
ols_model <- lm(wage76 ~ grade76 + age76, base.ed)
coeftest(ols_model, vcov = vcovHC(ols_model, type = "HC"))
```

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1830865  0.0772381 -2.3704  0.01783 *
grade76      0.0525110  0.0027812 18.8808  < 2e-16 ***
age76        0.0406575  0.0023938 16.9848  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Denotando que el Número de años que llevaba aprobados hasta el año 76 genera un impacto en el salario bastante significativo. A medida que esto aumenta en un año adicional el salario tambien aumenta.

Guardar el modelo MCO

A continuación con el objetivo de hacer una comparación adecuada de modelos vamos a ir guardando las salidas correspondientes en objetos para darle una forma mas adecuada y con ellos tener una mejor interpretabilidad

```r
ols_model <- lm(wage76 ~ grade76 + age76, data = base.ed)
```

Modelo IV usando la función ivreg del paquete AER

Empecemos a instrumentar, la idea parte de decir que el hecho de que una universidad este cerca a un individuo incide en su escolaridad o tomar una elección de hacerlo pero no tiene nada que ver con el salario que percibe. La edad tambien se convierte en otro instrumento clave. La edad es un factor decisorio para de alguna manera tomar la decisión de escolarizarse y solo

en algunas veces puede incidir algo en el salario debido a politicas que puedan tener las empresas pero no deberia darnos problemas adicionales.

```
library(AER)
iv_model <- ivreg(wage76 ~ grade76 + age76 | col4 + age76, data = base.ed)
summary(iv_model, vcov = sandwich)
```

```
Call:
ivreg(formula = wage76 ~ grade76 + age76 | col4 + age76, data = base.ed)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06376 -0.33914  0.01597  0.34869  2.04965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.819567   0.334451  -5.440 5.74e-08 ***
grade76      0.173973   0.024236   7.178 8.85e-13 ***
age76        0.041563   0.003019  13.767  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5168 on 3007 degrees of freedom
Multiple R-Squared: -0.3554,    Adjusted R-squared: -0.3563
Wald test: 118.3 on 2 and 3007 DF,  p-value: < 2.2e-16
```

Los resultados son homocedasticos al aplicar la corrección tipo `sandwich` para corregir problemas de varianza no constante.

Tabla de comparaciones entre MCO y IV

Vamos hacer la comparación entre modelos para ver que diferencias tienen los resultados. El paquete `stargazer` ayuda a darle visibilidad a los datos

```r
library(stargazer)
stargazer(ols_model, iv_model, type = "text",
          keep = c("grade76"), digits = 4,
          se = list(coeftest(ols_model, vcov = vcovHC(ols_model, type = "HC"))[, 2],
                    sqrt(diag(vcovHC(iv_model, type = "HC")))),
          dep.var.labels = "Salario en el 76'", covariate.labels = "Años aprobados 76'")
```

```
================================================================
                                        Dependent variable:
                                    -----------------------------
                                           Salario en el 76'
                                        OLS            instrumental
                                                         variable
                                        (1)                (2)
----------------------------------------------------------------
Años aprobados 76'                   0.0525***          0.1740***
                                     (0.0028)           (0.0242)

----------------------------------------------------------------
Observations                           3,010              3,010
R2                                     0.1813            -0.3554
Adjusted R2                            0.1808            -0.3563
Residual Std. Error (df = 3007)        0.4017             0.5168
F Statistic                  333.0018*** (df = 2; 3007)
================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Note la diferencia que existe en el estimador IV. Al parecer estábamos **sobreestimando** el efecto que tiene la educación en el salario de los individuos.

Estimación del estimador IV en dos etapas

Miremos esto por fases o etapas[2] que consideramos importante en la estimación del modelo

[2] Son los comunmente conocidos Modelos 2LS (Two least Square) o Bi-etapicos.

Primera etapa: regresión OLS para obtener la predicción

Corremos un modelo donde se intenta explicar la educación con los controles sugeridos

```
first_stage <- lm(grade76 ~ col4 + age76, data = base.ed )
base.ed$grade76hat <- predict(first_stage)
```

Segunda etapa: regresión OLS con el valor predicho

Estimamos el modelo pero con la educación estimada ($\hat{educ}$)

```
ols_two_stage <- lm(wage76 ~ grade76hat + age76, data = base.ed)
coeftest(ols_two_stage, vcov = vcovHC(ols_two_stage, type = "HC"))
```

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -1.8195667  0.2742206 -6.6354 3.823e-11 ***
grade76hat   0.1739731  0.0200007  8.6983 < 2.2e-16 ***
age76        0.0415634  0.0024976 16.6415 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Repetir el análisis con más controles

Recordemos que podemos adherir un número mayor de controles[3] sobre nuestra estimación. Para eso, primero crearemos una lista o `list`, eso con el objeto de no tener una formula muy larga en la línea de comandos. Luego usamos la función de `as.formula` y luego se corre.

[3] El conjunto controles son muchas variables "categoricas" que contienen valores de 1 o 0, para aquellas caracteristicas de si vive en cierta region (reg), si vive el el sur (south76), si vive en área metropolitana. Con la opcion `View` estan mejor clarificadas

```
# Crear una lista de variables independientes
regresores <- c("south76", "smsa76", "reg2", "reg3", "reg4", "reg5", "reg6", "reg7", "reg8", "r

## Acá los juntamos todos
formula <- as.formula(paste("wage76 ~", paste(regresores, collapse = " + ")))
```

```r
# Estimamos el modelo con todos los controles
ols_model_controls <- lm(formula, data = base.ed)
coeftest(ols_model_controls, vcov = vcovHC(ols_model_controls, type = "HC"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.2428234  0.0857897 14.4869 < 2.2e-16 ***
south76     -0.1445395  0.0308281 -4.6886 2.875e-06 ***
smsa76       0.1386612  0.0212900  6.5130 8.615e-11 ***
reg2         0.0776287  0.0384625  2.0183 0.0436497 *
reg3         0.1197300  0.0367253  3.2601 0.0011261 **
reg4         0.0529788  0.0429885  1.2324 0.2178982
reg5         0.0321178  0.0458129  0.7011 0.4833165
reg6         0.0256598  0.0486610  0.5273 0.5980128
reg7         0.0446355  0.0493165  0.9051 0.3654948
reg8        -0.0190857  0.0542837 -0.3516 0.7251694
reg9         0.1353619  0.0425367  3.1822 0.0014764 **
smsa66       0.0424718  0.0207877  2.0431 0.0411282 *
momdad14     0.1519061  0.0276634  5.4912 4.326e-08 ***
sinmom14     0.0192168  0.0395725  0.4856 0.6272796
nodaded     -0.0663376  0.0542202 -1.2235 0.2212431
nomomed      0.0035360  0.0378875  0.0933 0.9256489
daded        0.0080566  0.0048607  1.6575 0.0975265 .
momed        0.0156137  0.0045005  3.4693 0.0005292 ***
famed1      -0.1584073  0.0868107 -1.8247 0.0681396 .
famed2      -0.0734555  0.0756583 -0.9709 0.3316846
famed3      -0.1117697  0.0696830 -1.6040 0.1088258
famed4       0.0677970  0.0477179  1.4208 0.1554828
famed5      -0.0754945  0.0672963 -1.1218 0.2620280
famed6      -0.0809125  0.0654690 -1.2359 0.2165971
famed7      -0.1272778  0.0695159 -1.8309 0.0672128 .
famed8      -0.0854991  0.0593577 -1.4404 0.1498580
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo IV con más controles

De la misma forma como se hizo anteriormente vamos a instrumentar nuestro modelo, haciendo uso de los instrumentos dispuestos para ello $z = \{col4, age76\}$.

```
formula_iv <- as.formula(paste("wage76 ~ grade76 +",
                               paste(regresores, collapse = " + "),
                               "| col4 +",
                               paste(regresores, collapse = " + ")))

# Ajustar el modelo IV
iv_model_controls  <- ivreg(formula_iv, data = base.ed)
summary(iv_model_controls, vcov = sandwich)
```

```
Call:
ivreg(formula = formula_iv, data = base.ed)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82689 -0.27226  0.02088  0.28000  1.42205

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.408944   0.443146   0.923  0.35618
grade76      0.102552   0.053324   1.923  0.05455 .
south76     -0.131382   0.031430  -4.180    3e-05 ***
smsa76       0.060411   0.046635   1.295  0.19528
reg2         0.037715   0.045200   0.834  0.40411
reg3         0.082576   0.043468   1.900  0.05757 .
reg4        -0.005886   0.055586  -0.106  0.91568
reg5         0.031876   0.047325   0.674  0.50064
reg6         0.030389   0.049921   0.609  0.54274
reg7         0.015680   0.052746   0.297  0.76627
reg8        -0.091339   0.067261  -1.358  0.17457
reg9         0.073415   0.054004   1.359  0.17411
smsa66       0.066245   0.024507   2.703  0.00691 **
momdad14     0.104967   0.038218   2.747  0.00606 **
sinmom14     0.008422   0.041196   0.204  0.83802
nodaded     -0.064221   0.057085  -1.125  0.26068
```

```
nomomed      0.029217   0.040729   0.717  0.47321
daded       -0.009247   0.010283  -0.899  0.36857
momed       -0.002514   0.010504  -0.239  0.81087
famed1      -0.287469   0.112147  -2.563  0.01042 *
famed2      -0.225740   0.111728  -2.020  0.04343 *
famed3      -0.202399   0.087057  -2.325  0.02014 *
famed4      -0.026049   0.068230  -0.382  0.70265
famed5      -0.139581   0.077837  -1.793  0.07304 .
famed6      -0.150518   0.077045  -1.954  0.05084 .
famed7      -0.173131   0.076272  -2.270  0.02328 *
famed8      -0.159091   0.071819  -2.215  0.02682 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4274 on 2983 degrees of freedom
Multiple R-Squared: 0.08057,    Adjusted R-squared: 0.07255
Wald test: 21.83 on 26 and 2983 DF,  p-value: < 2.2e-16
```

Tabla de comparaciones entre OLS y IV con más controles

Una salida general a todo esto que hemos realizado nos da como resultado[4] un modelo en un formato de estilo mas científico

```
stargazer(ols_model_controls, iv_model_controls, type = "text",
          keep = c("grade76"), digits = 4,
          dep.var.labels = "wage76", covariate.labels = "grade76")
```

[4] La sobreidentificación se hace con una prueba F, comparando dos tipos de modelos. Uno con todos los regresores y otro sin ellos, al ver la significancia podemos entonces concluir que no tenemos sobreidentificación.

```
==========================================================
                            Dependent variable:
                    ---------------------------------------
                                    wage76
                              OLS               instrumental
                                                  variable
                              (1)                   (2)
----------------------------------------------------------
grade76                                           0.1026*
                                                  (0.0550)
```

```
---------------------------------------------------------------
Observations                    3,010                    3,010
R2                              0.1570                   0.0806
Adjusted R2                     0.1499                   0.0726
Residual Std. Error     0.4092 (df = 2984)     0.4274 (df = 2983)
F Statistic         22.2239*** (df = 25; 2984)
===============================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Diagnóstico de instrumentos débiles: correlación y regresión

Para verificar relevancia de los instrumentos podemos estimar entonces

```r
cor(base.ed$grade76, base.ed$col4)
```

```
[1] 0.1442402
```

```r
first_stage_diagnostic <- lm(grade76 ~ col4 + age76, data = base.ed)
summary(first_stage_diagnostic)
```

```
Call:
lm(formula = grade76 ~ col4 + age76, data = base.ed)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6845  -1.5549  -0.5801   2.4325   5.3786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.05037    0.43882  29.740  < 2e-16 ***
col4         0.83256    0.10379   8.021 1.49e-15 ***
age76       -0.01262    0.01541  -0.819    0.413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 3007 degrees of freedom
```

```
Multiple R-squared:  0.02102,   Adjusted R-squared:  0.02037
F-statistic: 32.29 on 2 and 3007 DF,  p-value: 1.337e-14
```

La edad en el año 76 parece no ser relevante, sin embargo en conjunto y mirando la prueba F, podemos entonces decir que si es útil

Test de instrumentos débiles (Anderson-Rubin Wald test)

El test de Anderson-Rubin evalúa si los coeficientes de las variables instrumentales en la regresión auxiliar (donde las variables instrumentales se utilizan para predecir la variable endógena) son estadísticamente diferentes de cero. Si se rechaza la hipótesis nula, esto indica que los instrumentos son relevantes y, por lo tanto, no son débiles.

```r
# Del Paquete de AER
waldtest(iv_model_controls)
```

```
Wald test

Model 1: wage76 ~ grade76 + south76 + smsa76 + reg2 + reg3 + reg4 + reg5 +
    reg6 + reg7 + reg8 + reg9 + smsa66 + momdad14 + sinmom14 +
    nodaded + nomomed + daded + momed + famed1 + famed2 + famed3 +
    famed4 + famed5 + famed6 + famed7 + famed8 | col4 + south76 +
    smsa76 + reg2 + reg3 + reg4 + reg5 + reg6 + reg7 + reg8 +
    reg9 + smsa66 + momdad14 + sinmom14 + nodaded + nomomed +
    daded + momed + famed1 + famed2 + famed3 + famed4 + famed5 +
    famed6 + famed7 + famed8
Model 2: wage76 ~ 1 | col4 + south76 + smsa76 + reg2 + reg3 + reg4 + reg5 +
    reg6 + reg7 + reg8 + reg9 + smsa66 + momdad14 + sinmom14 +
    nodaded + nomomed + daded + momed + famed1 + famed2 + famed3 +
    famed4 + famed5 + famed6 + famed7 + famed8
  Res.Df  Df  Chisq Pr(>Chisq)
1   2983
2   3009 -26 512.74  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Constatando que los instrumentos utilizados deben estar presente[5].

[5] La sobreidentificación se hace con una prueba F, comparando dos tipos de modelos. Uno con todos los regresores y otro sin ellos, al ver la significancia podemos entonces concluir que no tenemos sobreidentificación.

Agradecimientos

Mucho de este trabajo se debe a los artículos, recursos (free commons) y material del Profesor Colin Cameron, autor del libro: Microeconometrics Using Stata.

Carlos Yanes Guerra | Departamento de Economía | Universidad del Norte