Monitoria Variables Instrumentales

Juan C. Forero - Jhan Andrade - Germán C. Rodriguéz

7/10/2020

Contents

Introducción **1**
Teoría para el caso univariado . . . 1
MC2E . . . 2
Primera etapa . . . 2
Segunda etapa . . . 2
Paquete para realizar variables instrumentales en R . . . 2

Ejemplo Univariado (un solo instrumento) (Para el salario determiando por la educación) **3**
Mínimos Cuadrados en 2 Etapas . . . 5
Primera etapa . . . 5
Segunda etapa . . . 5
Usando el comando *ivreg* . . . 6
Código *iv_robust* . . . 6

Ejemplo multivariado(Dos instrumentos) (Para el salario determiando por la educación) **8**
Mínimos cuadrados en 2 etapas . . . 8
Primera etapa . . . 8
Segunda etapa . . . 9
Ivreg . . . 9
iv_robust . . . 10

Test de diagnósticos: Hausman, Instrumentos débiles y sobreidentificación de Sargan **10**
Test de Hausman . . . 10
Test de Hausman manual . . . 12
Test de diagnóstico con base en el modelo de errores robustos. . . . 13

Bibliografía **14**

El presente documento fue creado utilizando R markdown. La idea es ilustrar los prinipales conceptos de variables instrumentales mediante el software R

Introducción

Teoría para el caso univariado

El método de estimación por **Variables Instrumentales** nos permite resolver el problema de endogeneidad que pueda presentar algún regresor[1].

[1]En el presente documento, una variable se va a considerar endogena si existe una correlación entre dicha varaible en cuestión y el error ε

Ahora bien, este método nos permite obtener estimadores consistentes[2]. Este método de estimación está basado en encontrar un instrumento (Z) que cumpla las siguientes condiciones:

- **Relevancia**: $cov(Z, X) \neq 0$[3]
- **Exogeneidad**: $cov(Z, \varepsilon)) = 0$

El **supuesto de relevancia** busca que el instrumento empleado Z tenga un alto grado de correlación con el regresor X que se considera endógeno. Entre mayor correlación entre el instrumento que se empleará y el regresor endógeno mejor. Diremos que un instrumento con un alto grado de correlación con el regresor es un instrumento fuerte para éste, de lo contrario diremos que es débil. El **supuesto de exogeneidad** requiere que el instrumento no tenga correlación con el error.

Ahora bien, a pesar de que teóricamente es muy importante que los dos supuestos se cumplan para que las estimaciones por **VI** sean válidas, en la practica se tiene que solo el supuesto de relevancia se puede corroborar. Lo anterior, quiere decir que **no existen** pruebas estadísticas capaces de corroborar el supuesto de **exogeneidad**, la justificación en la práctica de dicho supuesto se hace desde la *teoría económica* o desde la *lógica del modelo* que justifique la exogeneidad del instrumento.

Es importante tener en cuenta, que el método de variables instrumentales es bastante sencillo, sin embargo encontrar buenos instrumentos es la parte compleja del método. En la prática, un investigador que vaya emplear el método de variables instrumentales dedicará la mayor parte del tiempo a buscar y justificar porque su instrumento es adecuado para la estimaciónd dado que el cumplimiento de los dos supuestos mencioandos anteriormente es fundamental para una buena estimación por VI.

MC2E

En la práctica, es muy usual emplear variables instrumentales mediante lo que se conoce como **Mínimos cuadrados en 2 etapas (MC2E)**.

Para el ejemplo ilustrativo de *MC2E*[4], nos basaremos en (Wooldridge 2016)

Suponga que se tiene la siguiente ecuación estructural[5]: $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \nu_1$ y suponga que además y_2 es una variable endógena con dos instrumentos z_2 y z_3. El método consiste en dos etapas en donde en cada etapa se hace una estimación por OLS.

Primera etapa

En la primera etapa, hay que estimar por OLS la ecuación en forma reducida[6]:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \nu_2$$

Dónde para que las variables Z_i sean exógenas se requiere que: $cov(z_1, \nu_2) = cov(z_2, \nu_2) = cov(z_3, \nu_2) = 0$.

Se busca que los instrumentos sean relevantes, es decir, que $\pi_2 \neq 0$ o $\pi_3 \neq 0$. Por tanto, de las siguientes hipótesis nulas:

[2]Fijense que en el curso se hace un mayor enfásis en la consistencia de los estiamdores que en la insesgadez. Ello se debe a que requerir que los estiamdores sean insesgados puede ser una condición muy fuerte. No obstante, la consistencia es una propiedad que no es tan restrictiva y que es fundamental que tengan cualquier estimador que se desea utilizar en la práctica

[3]X sería el regresor que se considera endógeno y por ende es necesario el uso de un instrumento para corregir por ese problema de endogeneidad

[4]Para justificar el método de MC2E se debe satisfacer las condiciones de orden y rango. La de orden requiere que haya al menos tantos instrumentos como regresores endógenas y la condición de rango es una condición más elaborada que se encuentra por ejemplo en el libro avanado de Wooldridge.

[5]Recuerden que una ecuación estructural es aquella en la que la variable de interés se escribe tanto en los regresores endógenos como exógenos

[6]Recuerden que una ecuación en forma reducida es aquella en la que una variable endógena se escribe a partir de variables exógenas exclusivamente

$$H_0 : \pi_2 = \pi_3 = 0$$
$$H_a : \pi_2 \neq 0 \quad o \quad \pi_3 \neq 0$$

Se emplea un *estadístico F* para realizar las pruebas de hipótesis anteriores y lo que se busca es rechazar la hipótesis nula a favor de la alternativa para corroborar relevancia.

Luego, de mirar la relevancia de los instrumentos y justificar que los instrumentos son exógenos se procede a realizar la segunda etapa.

Segunda etapa

Dados los resultados de la estimación, es decir $\hat{y_2} = \hat{\pi_0} + \hat{\pi_1} z_1 + \hat{\pi_2} z_2 + \hat{\pi_3} z_3$, se reemplaza y_2 por $\hat{y_2}$[7] y se estima la ecuación estructural:

$$y_1 = \beta_0 + \beta_1 \hat{y_2} + \beta_2 z_1 + \nu_1$$

por OLS[8]. Los parámetros estimados por el procedimiento anterior, de dos etapas se le conoce como los *parámetros estimados por MC2E*.

Paquete para realizar variables instrumentales en R

En R Studio instalaremos los siguientes paquetes para proceder a realizar el ejercicio de Variables Instrumentales.

```
#install.packages("AER") #Applied Econometrics with R for Instrumental Variables
#install.packages("foreing") #Para cargar datos con formato Stata
#install.packages("stargazer") #Para una presentación más estética de los resultados
#install.packages("estimatr") #Para hacer MC2E con errores robustos
#install.packages("arm") #Análisis de datos utilizando regresiones
#install.packages("lmtest")
```

Cargamos los paquetes:

```
library(AER);library(foreign); library(stargazer);
library(arm);library(lmtest);library(estimatr);library(tidyverse)
```

- **AER**: Applied Econometrics with R for Instrumental Variables. Contiene la función *ivreg*
- **foreing**: Para importar bases de datos tipo stata *dta*
- **stargazer**: Para prestnación de resultados
- **estimatr**: Para hacer MC2E con errores robustos. Contiene la función *iv_robust*
- **arm**: Para análisis de datos utilizando regresiones
- **lmtest**: Conjutntos de test para modelos de estimación lineales

Ejemplo Univariado (un solo instrumento) (Para el salario determiando por la educación)

```
#Cargamos la base de datos.
data=read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.dta")
```

La estructura de datos es de corte transversal y corresponde a un data.frame de 753 para 22 variables:

[7] Recordar que en el ejemplo y_2 es el regresor endógeno
[8] Recordar que en el ejemplo z_1 no es un instrumento

```
glimpse(data)
```

```
## Rows: 753
## Columns: 22
## $ inlf     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ hours    <dbl> 1610, 1656, 1980, 456, 1568, 2032, 1440, 1020, 1458, 1600,...
## $ kidslt6  <dbl> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
## $ kidsge6  <dbl> 0, 2, 3, 3, 2, 0, 2, 0, 2, 2, 1, 1, 2, 2, 1, 3, 2, 5, 0, 4...
## $ age      <dbl> 32, 30, 35, 34, 31, 54, 37, 54, 48, 39, 33, 42, 30, 43, 43...
## $ educ     <dbl> 12, 12, 12, 12, 14, 12, 16, 12, 12, 12, 12, 11, 12, 12, 10...
## $ wage     <dbl> 3.3540, 1.3889, 4.5455, 1.0965, 4.5918, 4.7421, 8.3333, 7....
## $ repwage  <dbl> 2.65, 2.65, 4.04, 3.25, 3.60, 4.70, 5.95, 9.98, 0.00, 4.15...
## $ hushrs   <dbl> 2708, 2310, 3072, 1920, 2000, 1040, 2670, 4120, 1995, 2100...
## $ husage   <dbl> 34, 30, 40, 53, 32, 57, 37, 53, 52, 43, 34, 47, 33, 46, 45...
## $ huseduc  <dbl> 12, 9, 12, 10, 12, 11, 12, 8, 4, 12, 12, 14, 16, 12, 17, 1...
## $ huswage  <dbl> 4.0288, 8.4416, 3.5807, 3.5417, 10.0000, 6.7106, 3.4277, 2...
## $ faminc   <dbl> 16310, 21800, 21040, 7300, 27300, 19495, 21152, 18900, 204...
## $ mtr      <dbl> 0.7215, 0.6615, 0.6915, 0.7815, 0.6215, 0.6915, 0.6915, 0....
## $ motheduc <dbl> 12, 7, 12, 7, 12, 14, 14, 3, 7, 7, 12, 14, 16, 10, 7, 16, ...
## $ fatheduc <dbl> 7, 7, 7, 7, 14, 7, 7, 3, 7, 7, 3, 7, 16, 10, 7, 10, 7, 12,...
## $ unem     <dbl> 5.0, 11.0, 5.0, 5.0, 9.5, 7.5, 5.0, 5.0, 3.0, 5.0, 5.0, 5....
## $ city     <dbl> 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0...
## $ exper    <dbl> 14, 5, 15, 6, 7, 33, 11, 35, 24, 21, 15, 14, 0, 14, 6, 9, ...
## $ nwifeinc <dbl> 10.910060, 19.499981, 12.039910, 6.799996, 20.100060, 9.85...
## $ lwage    <dbl> 1.2101541, 0.3285121, 1.5141380, 0.0921233, 1.5242720, 1.5...
## $ expersq  <dbl> 196, 25, 225, 36, 49, 1089, 121, 1225, 576, 441, 225, 196,...
```

Las principales variables son:

- inlf: =1 if in lab frce, 1975;
- hours: hours worked, 1975;
- kidslt6: kids < 6 years kidsge6: # kids 6-18;
- age: woman's age in yrs;
- educ: years of schooling;
- wage: est. wage from earn, hrs;
- repwage: rep. wage at interview in 1976;
- hushrs: hours worked by husband, 1975;
- husage: husband's age;
- huseduc: husband's years of schooling;
- huswage: husband's hourly wage, 1975;
- faminc: family income, 1975;
- mtr: fed. marg. tax rte facing woman;
- motheduc: mother's years of schooling;
- fatheduc: father's years of schooling;
- unem: unem. rate in county of resid;
- city: =1 if live in SMSA;
- exper: actual labor mkt exper;
- nwifeinc: (faminc - wage*hours)/1000;
- lwage: log(wage);
- expersq: exper^2

Ahora vamos a eliminar las observaciones que no tienen salario: !is.na(wage)

```
data.VI <- subset(data,!is.na(wage))
attach(data.VI)
```

Se busca explicar la variación de salario en función de la educación de las personas. No obstante, es de esperarse que la educación sea una variable endógena dado que dependerá de factores inobservables de las personas como su habilidad innata u otros factores no observables que se recogen en el error.

Por tanto, utilizando Mínimos Cuadrados Ordinarios (MCO), sin ningun consideración adicional, se espera que en este procedimiento existan problemas de endogeneidad:

```
R.MCO = lm(log(wage)~ educ, data=data.VI)
summary(R.MCO)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ, data = data.VI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10256 -0.31473  0.06434  0.40081  2.10029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1852     0.1852  -1.000    0.318
## educ          0.1086     0.0144   7.545 2.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.68 on 426 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.1158
## F-statistic: 56.93 on 1 and 426 DF,  p-value: 2.761e-13
```

Usaremos el método de variables instrumentales, de tal manera que la educación del padre *fatheduc* será un instrumento para *educ*. Pues, es de esperarse que la educación de los padres determine la educación de los hijos. Entre más años de escolaridad tengan los padres de las personas, los hijos tengan un mayor grado de escolaridad. Es decir, una persona cuyos padres tienen un nivel de educación básica-primaria posiblemente tenga un nivel de escolaridad menor que el de una persona cuyos padres tienen doctorado.

Empezaremos calculando el coeficiente β_1 de forma manual [9]:

$$\frac{cov(y,z)}{cov(x,z)}$$

```
Beta = with(data.VI,cov(log(wage),fatheduc)/cov(educ,fatheduc)); Beta
```

```
## [1] 0.05917348
```

Mínimos Cuadrados en 2 Etapas

Primera etapa

En esta etapa se deben incluir como controles todas las variables exógenas del modelo. También debe evaluarse la relevancia del instrumento.

```
Reg.aux=lm(educ ~ fatheduc, data = data.VI); summary(Reg.aux)
```

```
##
## Call:
## lm(formula = educ ~ fatheduc, data = data.VI)
```

[9] Acá se está empleando la fórmula de variable instrumental cuándo solo hay un instrumento

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4704 -1.1231 -0.1231  0.9546  5.9546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.23705    0.27594  37.099   <2e-16 ***
## fatheduc     0.26944    0.02859   9.426   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.081 on 426 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:  0.1706
## F-statistic: 88.84 on 1 and 426 DF,  p-value: < 2.2e-16
```

De lo anterior, vemos que **nuestro instrumento *fatheduc* es estadísticamente significativo y por lo tanto se puede concluir que se cumple el supuesto de relevancia**

Calculamos los valores ajustados de la regresión de la primera etapa para la variable *educ*, que harán de instrumento en la segunda etapa.

```
educ.fitted = fitted(Reg.aux)
```

Segunda etapa

Incluimos los valores ajustados de educ que se obtuvieron de la primera etapa:

```
Reg.VI = lm(log(wage)~educ.fitted, data=data.VI)
summary(Reg.VI)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ.fitted, data = data.VI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2126 -0.3763  0.0563  0.4173  2.0604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.44110    0.46711   0.944    0.346
## educ.fitted  0.05917    0.03680   1.608    0.109
##
## Residual standard error: 0.7219 on 426 degrees of freedom
## Multiple R-squared:  0.006034,   Adjusted R-squared:  0.003701
## F-statistic: 2.586 on 1 and 426 DF,  p-value: 0.1086
```

Usando el comando *ivreg*

El comando ***ivreg*** del paquete **AER** nos permite realizar el procedimiento de Variables instrumentales sin necesidad de hacer etapa por etapa. En otras palabras, este comando nos ahorra lineas de código.

```
R.VI = ivreg(log(wage) ~ educ| fatheduc , data=data.VI); summary(R.VI)
```

```
##
## Call:
```

```
## ivreg(formula = log(wage) ~ educ | fatheduc, data = data.VI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0870 -0.3393  0.0525  0.4042  2.0677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.44110    0.44610   0.989   0.3233
## educ         0.05917    0.03514   1.684   0.0929 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6894 on 426 degrees of freedom
## Multiple R-Squared: 0.09344, Adjusted R-squared: 0.09131
## Wald test: 2.835 on 1 and 426 DF,  p-value: 0.09294
```

La estructura del código consiste en:

ivreg(Variable Explicada ~ regresores del modelo (incluyendo regresores endógenos y exógenos) | Variables exógenas + instrumentos)

Código *iv_robust*

El comando ***iv_robust*** del paquete *estimatr* incorpora los errores estándar robustos. La estructura de código es la misma que la de *ivreg*. La principal diferencia es que con *iv_robust* el R^2 se corrige y es posible hacer una adecuada interpretación de este[10].

```
R.VI.Robust = iv_robust(log(wage) ~ educ| fatheduc , data=data.VI);summary(R.VI.Robust)
```

```
##
## Call:
## iv_robust(formula = log(wage) ~ educ | fatheduc, data = data.VI)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  0.44110    0.46651  0.9455   0.3449 -0.47584   1.3581 426
## educ         0.05917    0.03712  1.5940   0.1117 -0.01379   0.1321 426
##
## Multiple R-squared:  0.09344 ,   Adjusted R-squared:  0.09131
## F-statistic: 2.541 on 1 and 426 DF,  p-value: 0.1117
```

Presentando los resultados mediante el comando stargazer[11]. La tabla generada en latex por el paquete stargazer es:

Llegados a este punto, se pueden dar cuenta que para el caso univariado la estimación manual por covarianzas, Mínimos Cuadrados Ordinarios en 2 etapas, IVREG y IV_ROBUST nos llevan al mismo coeficiente. (0.059)

[10] Por lo anterior, se recomienda que utilizen **iv_robuts** a la hora de realizar estiamciones por *MC2E*

[11] Fijense que una de las principales débilidades del comando **iv_robust** es que no se puede usar directamente en el stargazer.

Table 1: Tabla de regresiones para el ejemplo univariado

	Dependent variable:		
	log(wage)		
	MCO	MC2E	IVREG
	(1)	(2)	(3)
educ	0.109***		
	(0.014)		
educ.fitted		0.059	0.059
		(0.037)	(0.037)
Constant	−0.185	0.441	0.441
	(0.185)	(0.467)	(0.467)
Observations	428	428	428
R^2	0.118	0.006	0.006
Adjusted R^2	0.116	0.004	0.004
Residual Std. Error (df = 426)	0.680	0.722	0.722
F Statistic (df = 1; 426)	56.929***	2.586	2.586
Note:		*p<0.1; **p<0.05; ***p<0.01	

Ejemplo multivariado(Dos instrumentos) (Para el salario determiando por la educación)

Estimaremos la regresión MCO que originalmente debe tener problemas de endogenidad[12].

```
Reg.MCO = lm(log(wage)~educ+exper+I(exper^2)+city, data=data.VI)
summary(Reg.MCO)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + I(exper^2) + city, data = data.VI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10084 -0.32453  0.05292  0.36261  2.34806
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5308476  0.1990253  -2.667  0.00794 **
## educ         0.1057097  0.0143280   7.378 8.58e-13 ***
## exper        0.0410584  0.0131963   3.111  0.00199 **
## I(exper^2)  -0.0007973  0.0003938  -2.025  0.04352 *
## city         0.0542225  0.0680903   0.796  0.42629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6667 on 423 degrees of freedom
```

[12]Por la misma razón que en el ejemplo anterio, muy seguramente la variable educación está correlacionada con el error (i.e. con factores inobservables)

```
## Multiple R-squared:  0.1581, Adjusted R-squared:  0.1501
## F-statistic: 19.86 on 4 and 423 DF,  p-value: 5.389e-15
```

Al igual que con el ejemplo univariado, es de esperarse que la educación sea endógena en la regresión. Para resolver este problema tomaremos la ecuación del padre (*fatheduc*) y la educación de la madre (*motheduc*) como instrumentos para *educ*[13].

Vamos a realizar un análisis preliliminar la relevancia de los instrumentos, de tal manera que la covarianza entre *educ* y el instrumento debe ser diferente de cero.

```
cor(educ,motheduc)
```

```
## [1] 0.3870198
```

```
cor(educ,fatheduc)
```

```
## [1] 0.415403
```

Mínimos cuadrados en 2 etapas

Primera etapa

Recuerden que en esta etapa se hace la regresión de la variable endógena (*educ*) en función de las exógenas y los instrumentos.

```
stage1 <- lm(educ~exper+I(exper^2)+city+motheduc+fatheduc, data=data.VI)
```

Segunda etapa

```
stage2<-lm(log(wage)~fitted(stage1)+exper+I(exper^2)+city, data=data.VI)
summary(stage2)
```

```
##
## Call:
## lm(formula = log(wage) ~ fitted(stage1) + exper + I(exper^2) +
##     city, data = data.VI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08238 -0.34886  0.03594  0.37694  2.31829
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.0723140  0.4221131   0.171   0.8641
## fitted(stage1)  0.0552272  0.0341642   1.617   0.1067
## exper           0.0434902  0.0140541   3.094   0.0021 **
## I(exper^2)     -0.0008816  0.0004202  -2.098   0.0365 *
## city            0.0916476  0.0756022   1.212   0.2261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 423 degrees of freedom
## Multiple R-squared:  0.05558,    Adjusted R-squared:  0.04665
## F-statistic: 6.223 on 4 and 423 DF,  p-value: 7.135e-05
```

[13]Fijense que se puede tener más de un instrumento para la misma variable endógena pero al menos se necesita un instrumento, lo anterior dado por la condción de orden para *MC2E*

Ivreg

ivreg(Variable Explicada ~ regresores del modelo (incluyendo regresores endógenos y exógenos) | Variables exógenas + instrumentos, data="...")

OJO: recuerden que la forma en cómo funciona el comando *ivreg* es: ivreg(Variable Explicada ~ regresores del modelo (incluyendo regresores endógenos y exógenos) | Variables exógenas + instrumentos, data="...")[14]

```
aut.MC2E<-ivreg(log(wage)~educ+exper+I(exper^2)+
                  city|motheduc+fatheduc+exper+I(exper^2) + city , data=data.VI)
summary(aut.MC2E)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + I(exper^2) + city |
##     motheduc + fatheduc + exper + I(exper^2) + city, data = data.VI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05056 -0.33418  0.04927  0.35824  2.30873
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0723140  0.4043536   0.179  0.85815
## educ         0.0552272  0.0327268   1.688  0.09224 .
## exper        0.0434902  0.0134629   3.230  0.00133 **
## I(exper^2)  -0.0008816  0.0004025  -2.190  0.02906 *
## city         0.0916476  0.0724214   1.265  0.20640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6764 on 423 degrees of freedom
## Multiple R-Squared: 0.1334,  Adjusted R-squared: 0.1252
## Wald test: 6.782 on 4 and 423 DF,  p-value: 2.684e-05
```

iv_robust

El comando *iv_robust*[15] funciona con la misma sintaxis que *ivreg*, tienen la misma estructura de código.

```
MC2E.Robusto <- iv_robust(log(wage)~educ+exper+I(exper^2)+
                            city|motheduc+fatheduc+exper+I(exper^2) + city , data=data.VI)
summary(MC2E.Robusto)
```

```
##
## Call:
## iv_robust(formula = log(wage) ~ educ + exper + I(exper^2) + city |
##     motheduc + fatheduc + exper + I(exper^2) + city, data = data.VI)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower   CI Upper  DF
```

[14] Adicionalmente es importante que tengan en cuenta que "|" no es una L ni una I, es una raya vertical que por lo general se encuentra en la esquina superior izquierda de los teclados (debajo del esc). La barra vertical nos ayudará a indicarle al código cuales son nuestras variables exógenas y cuál es la endógena.

[15] El método que utiliza por default iv_robuts para corregir por errores robustos es el método que se conoce como **HC2**, como lo pueden observar en el summary

```
## (Intercept)  0.0723140  0.4337123  0.1667 0.867660 -0.78019  0.9248137 423
## educ         0.0552272  0.0343161  1.6094 0.108282 -0.01222  0.1226785 423
## exper        0.0434902  0.0157510  2.7611 0.006011  0.01253  0.0744503 423
## I(exper^2)  -0.0008816  0.0004369 -2.0180 0.044221 -0.00174 -0.0000229 423
## city         0.0916476  0.0691717  1.3249 0.185910 -0.04432  0.2276106 423
##
## Multiple R-squared:  0.1334 ,    Adjusted R-squared:  0.1252
## F-statistic:  5.66 on 4 and 423 DF,  p-value: 0.0001909
```

Presentación de resultados[16]:

Table 2: Resultados de la estimación para el caso de dos instrumentos

	Dependent variable:			
	log(wage)	educ	log(wage)	
	OLS	*OLS*	*OLS*	*instrumental variable*
	MCO	Etapa 1	Etapa 2	IVREG
	(1)	(2)	(3)	(4)
educ	0.106^{***}			0.055^{*}
	(0.014)			(0.033)
fitted(stage1)			0.055	
			(0.034)	
exper	0.041^{***}	0.040	0.043^{***}	0.043^{***}
	(0.013)	(0.040)	(0.014)	(0.013)
I(exper^2)	-0.001^{**}	-0.001	-0.001^{**}	-0.001^{**}
	(0.0004)	(0.001)	(0.0004)	(0.0004)
city	0.054	0.467^{**}	0.092	0.092
	(0.068)	(0.209)	(0.076)	(0.072)
motheduc		0.164^{***}		
		(0.036)		
fatheduc		0.174^{***}		
		(0.034)		
Constant	-0.531^{***}	8.914^{***}	0.072	0.072
	(0.199)	(0.433)	(0.422)	(0.404)
Observations	428	428	428	428
R^2	0.158	0.221	0.056	0.133
Adjusted R^2	0.150	0.211	0.047	0.125
Residual Std. Error	0.667 (df = 423)	2.029 (df = 422)	0.706 (df = 423)	0.676 (df = 423)
F Statistic	19.856^{***} (df = 4; 423)	23.901^{***} (df = 5; 422)	6.223^{***} (df = 4; 423)	
Note:				*p<0.1; **p<0.05; ***p<0.01

[16]Nuevamente, se resalta que a pesar de que se sugiere emplear iv_robuts para hacer las estiamciones de MC2E se tiene que no es posible incluir los resultados de dicho estimación directamente en stargazer

Test de diagnósticos: Hausman, Instrumentos débiles y sobreidentificación de Sargan

Test de Hausman

Algo que es muy impotante es saber si un regresor es endógeno o no lo es. Lo anterior se debe a que si el regresor no fuera endógeno entonces la estimación habiltual por MCO sería muchisimo más eficiente que la estimación por VI. Es decir, uno debería realizar una estiamción por VI exclusivamente cuando hay presencia de regresores endógenos, de lo contrario, si todos los regresores fueran exógenoes entonces variables instrumentales sería infeciente frente a MCO.

El problema de la endogeneidad de un regresor es un problema más teórico que práctica. Es decir, en el ejemplo anterior uno esperaría desde la teoría que la educación se encuentre correlacionada con factores inobservables que hagan parte del error como lo es la habilidad innata de los individuos dado que se espera teóricamente que personas más habilidosas generlamente o en promedio tengan mayores niveles de educación que personas menos habilidosas.

No obstante, a pesar de que la pregunta es de carácter teórica, existe una prueba formal que podría ayudar a iluminar un poco si algún regresor presenta endógenidad. Dicha prueba consiste como el test de Hausman y busca determinar si un regresor es endógeno o no.

El test de Hausman dice que:

$$H_0 = \text{ exogeneidad de los regresores} H_a \qquad = \text{ al menos un regresor es endógeno}$$

Esta prueba nos indicará si es necesario o no el uso de variables instrumentales de tal manera que la **hipotesis nula** es Exógeneidad. De tal manera, que si rechazamos la hipotesis nula, a favor de la alternativa significa que debemos hacer uso de Variables instrumentales.

Test de Hausman manual

La primera forma de aplicar el test de Hausman es realizar el test de manera manual. Este test consiste en dos etapas:

Etapa 1 Para aplicar esta prueba, es necesario calcular los errores o residuos de la **primera etapa**

```
res.stage1<-resid(stage1)
```

Ello se debe a que si el regresor fuera endógeno, entonces la *parte endógena* del regresor quedaría *almacenada* en los *residuales de la primera etapa.*

Etapa 2 Luego, estimaremos la regresión aumentada de la Variable dependiente contra los *regresores exógenas originales*, *la variable regresora que se cree endógena* y los *residuales de la primera etapa.* Si encontramos que el coeficiente que acompaña a los residuales de la primera etapa es *estadísticamente significativos* concluiremos que el modelo *sufre de endogeneidad.*

```
Reg.aum <- lm(log(wage)~educ+exper+I(exper^2)+city+res.stage1, data=data.VI)
summary(Reg.aum)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + I(exper^2) + city + res.stage1,
##     data = data.VI)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -3.02457 -0.31160  0.03785  0.39279  2.31036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0723140  0.3975779   0.182   0.8558
## educ         0.0552272  0.0321784   1.716   0.0868 .
## exper        0.0434902  0.0132373   3.285   0.0011 **
## I(exper^2)  -0.0008816  0.0003958  -2.227   0.0264 *
## city         0.0916476  0.0712079   1.287   0.1988
## res.stage1   0.0628909  0.0359160   1.751   0.0807 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6651 on 422 degrees of freedom
## Multiple R-squared:  0.1642, Adjusted R-squared:  0.1543
## F-statistic: 16.58 on 5 and 422 DF,  p-value: 5.936e-15
```

De lo anterior, concluimos que a un nivel de significancia del 10 % el test confirma nuestra intuición de que la variable educación es endógena.

Test de diagnóstico con base en el modelo de errores robustos.

De igual forma, cuando se tiene un sistema sobreidentificado, también es conveniente realizar algunos test de diagnóstico para corroborar ciertas propiedades de los instrumentos y de los regresores del modelo. Para realizar dichos test de diagnóstico es necesario agregar *diagnostics=TRUE* al comando iv_robust.

```
MC2E.Robusto <- iv_robust(log(wage)~educ+exper+
                            I(exper^2)+city|motheduc+fatheduc+exper+I(exper^2) + city ,
                          data=data.VI, diagnostics=TRUE)
summary(MC2E.Robusto, diagnostics = TRUE)
```

```
##
## Call:
## iv_robust(formula = log(wage) ~ educ + exper + I(exper^2) + city |
##     motheduc + fatheduc + exper + I(exper^2) + city, data = data.VI,
##     diagnostics = TRUE)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower   CI Upper  DF
## (Intercept)  0.0723140  0.4337123  0.1667 0.867660 -0.78019  0.9248137 423
## educ         0.0552272  0.0343161  1.6094 0.108282 -0.01222  0.1226785 423
## exper        0.0434902  0.0157510  2.7611 0.006011  0.01253  0.0744503 423
## I(exper^2)  -0.0008816  0.0004369 -2.0180 0.044221 -0.00174 -0.0000229 423
## city         0.0916476  0.0691717  1.3249 0.185910 -0.04432  0.2276106 423
##
## Multiple R-squared:  0.1334 ,    Adjusted R-squared:  0.1252
## F-statistic:  5.66 on 4 and 423 DF,  p-value: 0.0001909
##
## Diagnostics:
##                   numdf dendf  value p.value
## Weak instruments      2   422 46.184  <2e-16 ***
## Wu-Hausman            1   422  2.782  0.0961 .
```

```
## Overidentifying      1    NA  0.200  0.6547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con este comando se podrá determinar si los instrumento son débiles, si al menos un regresor es endógeno en el modelo y si los instrumentos son exógenos. En ese orden los test de diagnóstico sería, respectivamente:

- **Weak instrument:** Ho= Los instrumentos son débiles Ha= al menos un instrumento es débil (no se correlaciona con la variable endógena)
- **Hausman:** Ho = No hay endogeneidad (si se rechaza hay que usar variables instrumentales (i.e. MC2E))
- **Sargan:** Ho= Instrumentos exógenos - Ha= al menos uno es endógeno (sistema sobre-identificado)[17]

De los resultados del código anterior, es posible concluir que el modelo instrumentado:

- Del test de *Weak instruments* no se rechaza por lo que se podría decir que ambos instrumentos son fuertes (i.e. tienen fuera correlación con la variable que se supone endógena (la educación)). Con eso se estaría corroborando el supuesto de relvancia para los dos instrumentos
- El test de Hausman es el mismo test que se realizó en la sección anterior solo que acá se hace de manera automática y se corrige por errores robustos. Como se rechaza la hipótesis nula a un nivel de significancia del 10 % entonces se puede decir que hay endógenidad en el regresor.
- No rechazo la hipótesis nula de que ambos instrumentos son exógenos

Por los resultados anterioes, en la práctica podría utilizar ambos instrumentos *fatheduc* y *motheduc* para la variable regresora endógena *educ*

Bibliografía

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics: A Modern Approach.* CENGAGE Learning.

[17] El test de sobreidentificación solo aparece en sistemas sobreidentificados, es decir, cuando hayan más instrumentos que regresores endógenos