Econometría I - EAE- 250-A

Estimación con Variables Instrumentales

Ezequiel Garcia-Lembergman

Instituto de Economía - Pontificia Universidad Católica de Chile

2023

Introducción

- En la clase anterior vimos que cuando teníamos un regresor **endógeno** precisabamos un **instrumento**
- Para que el instrumento z sea válido precisamos:
    1. Exógeneidad: $Cov(u, z) = 0$
    2. Relevancia: $Cov(x, z) \neq 0$
- Hoy vamos a ver como podemos usar ese instrumento para estimar consistentemente
- Vamos a ver dos métodos:
    1. Estimación por Variables Instrumentales (VI) (clase pasada y un poco de hoy)
    2. Mínimos cuadrados en dos etapas (MC2E) (hoy)

Estimación por Variables Instrumentales (VI)

- La clase pasada vimos que el estimador por VI lo podemos calcular como:

$$\hat{\beta}_{VI} = \frac{S_{y_i,z_i}}{S_{x_i,z_i}} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}$$

- Si se cumplen los supuestos del instrumento, es un estimador insesgado de la relacion entre X e Y.

Inferencia y estimación por variables Instrumentales

- Al igual que MCO, VI es un método de estimación

  $\rightarrow$ Un método para obtener un estimador de nuestro parámetro de interés β

- Tal como con MCO, una vez obtenidas las estimaciones $\hat{\beta}_{VI}$ queremos hacer inferencia estadística sobre las mismas. Necesitamos errores estandar!

  $\rightarrow$ Ej. pruebas t de significación de los parámetros

- Tal como en MCO, vamos a asumir homoscedasticidad, ahora condicional en el instrumento z: $Var(u|z) = \sigma^2$
- Se puede demostrar que se puede estimar la varianza de β^{VI} como:
- $\hat{\beta}_{VI}$:

$$Var(\hat{\beta}_{VI}) = \frac{\hat{\sigma}^2}{SCT_x R^2_{x,z}}$$

- Se puede demostrar que se puede estimar la varianza de β^{VI} como:
- $\hat{\beta}_{VI}$:

$$Var(\hat{\beta}_{VI}) = \frac{\hat{\sigma}^2}{SCT_x R^2_{x,z}} \quad donde$$

1. $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i{}^2$

- $R^2_{x,z}$ es el R^2 de la regresion: $X_i = \pi_0 + \pi_1 Z_i + v$
- $SCT_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ es la suma de cuadrados explicada de X.

Estimador para la varianza de las estimaciones por VI

- Noten que:

$$\hat{Var}(\hat{\beta}_{VI}) = \frac{\hat{\sigma}^2}{SCT_x R^2_{x,z}}$$

- **Noto:** Mientras mas débil la relación entre x y su instrumento $z \rightarrow$ menor $R^2_{x,z}$ $\rightarrow$ Mayor $\hat{Var}(\hat{\beta}_{VI})$
- Recuerdo, en el Modelo de Regresión Simple:

$$\hat{Var}(\hat{\beta}_{MCO}) = \frac{\hat{\sigma}^2}{SCT_x}$$

- Dado que $R^2_{x,z} \leq 1$, la varianza de las estimaciones por VI será mayor o igual a la varianza de los estimadores MCO
  - Mientras mas fuerte la relación entre x y z, mas se acercarán las varianzas
  - Caso extremo: x es exógena $\rightarrow$ x es el instrumento de $x \rightarrow R^2_{x,z} = 1 \rightarrow$ $\hat{Var}(\hat{\beta}_{VI}) = \hat{Var}(\hat{\beta}_{MCO})$

Estimación por mínimos cuadrados en dos etapas

- Otro método para estimar consistentemente los parámetros con instrumentos es el **Mínimos Cuadrados en 2 etapas (MC2E)**. El coeficiente que obtenemos sera el mismo.
- Comenzamos con el siguiente modelo donde estamos interesados en el efecto de x_1 en y_1.

$$y_i = \beta_0 + \beta_1 \underbrace{x_{1i}}_{Endo} + \beta_2 \underbrace{x_{2i}}_{Exog} + u_i \quad \forall i = 1 ... N$$

Con: **(i)** $Cov(x_{1i}, u_i) \neq 0$; **(ii)** $Cov(x_{2i}, u_i) = 0$; y **(iii)** $Cov(z_i, u_i) = 0$

- Por ejemplo:

$$ln(salario_i) = \beta_0 + \beta_1 educ_i + \beta_2 mujer + u_i$$

- Y podemos usar como instrumento de *educ* la distancia al establecimiento educativo mas cercano.

MC2E - Etapa 1

Primera etapa: $Z \rightarrow X$ Regresamos la variable **endógena** sobre **todas las exógenas** incluido el instrumento

- Paso 1: Estimamos la siguiente ecuación por MCO.

$$\underbrace{x_{1i}}_{End} = \pi_o + \pi_1 \underbrace{z_i}_{\substack{\text{Instr}\\ \text{(Exog)}}} + \pi_2 \underbrace{x_{2i}}_{Exog} + v_i \tag{1}$$

- Paso 2: Obtenemos los $\hat{\pi}$ y construimos el valor predicho de x_1:

$$\hat{x}_{1i} = \hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{\pi}_2 x_{2i} \tag{2}$$

- Noten que $\hat{x}$ es una funcion de z y x_2 que son variables exogenas. Por lo tanto, no correlaciona c/ el error (u). El instrumento permite quedarnos con la parte exogena de x.
- intuicion: $x_i = \hat{x}_i + \hat{v}_i$. Solo usamos la parte exogena: $\hat{x}_i$ y nos deshacemos de la enodgena .
- Paso 3: Noten que durante este paso, se aprovecha para testear la relevancia del instrumento. Haciendo un test de $\hat{\pi}_1 \neq 0$. Si $F > 10$, relevante.

MC2E - Etapa 2

Segunda etapa: $\hat{x} \rightarrow y$

- En la ecuacion original, remplazamos la variable endógena x_1 por la predicción de esta variable que obtuvimos en la primera etapa $\hat{x}_1$ (Ec.2)
- Nuestra ecuación original (Ec.1) la podemos escribir entonces:

$$y_i = \beta_0 + \beta_1 \hat{x}_{1i} + \beta_2 x_{2i} + u_i \tag{3}$$

- $\hat{x}_{1i}$ es exógena, ya que es una combinación lineal de variables exógenas (Ec.2)
- Ahora, podemos estimar Ec.3 por MCO!

$$\hat{\beta}_{MC2E} = (\hat{X}'\hat{X}')^{-1}\hat{X}'Y$$

con $\hat{X} = \begin{pmatrix} 1 & \hat{x}_{1,1} & x_{2,1} \\ 1 & \hat{x}_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & \hat{x}_{1,n} & x_{2,n} \end{pmatrix}$

- El $\hat{\beta}_1$ hallado de esta manera se conoce como $\hat{\beta}_{1,MC2E}$

MC2E- Varianza e inferencia

- Si bien $\hat{\beta}_{MC2E}$ son la estimación por MCO de la segunda etapa su varianza no es tal cuál la varianza de un estimador MCO ya que debe recoger las correlaciones entre la primera y la segunda etapa
- La varianza de los estimadores MC2E se puede aproximar asintóticamente a:

$$Var(\hat{\beta}_{1,MC2E}) \approx \frac{\sigma^2}{\hat{SCT}_{\hat{x}}(1-\hat{R}_1^2)} \; ; Var(\hat{\beta}_{1,MCO}) = \frac{\sigma^2}{SCT_x(1-R_1^2)}$$

- Donde:
  - $\hat{SCT}_{\hat{x}} = \sum_{i=1}^{i=n}(\hat{x}-\bar{\hat{x}})^2$ es la variación total de $\hat{x}$
  - $\hat{R}_1^2$ es el R^2 de regresar $\hat{x}_1$ sobre los demás regresores (x_2): e.g: obtener el R^2 de $\hat{x}_1 = \alpha + \alpha_1 x_2 + c$
- Siempre $SCT_x \geq \hat{SCT}_{\hat{x}}$. Intuicion? acuerdense que de toda la variabilidad de X, agarramos solo una parte (la parte exogena). Ademas, generalmente, $\hat{R}_1^2 \geq R_1^2$
- Entonces: $Var(\hat{\beta}_{1,MC2E}) \geq Var(\hat{\beta}_{1,MCO})$
- Al igual que con VI, una relación débil entre la variable endógena y su instrumento incrementa la varianza de la estimación (porque achica el $SCT_{\hat{x}}$).

R cuadrado y pruebas F

- Tanto cuando estimamos por VI como por MC2E el R^2 calculado carecen de interpretación.
- Estimar usando instrumentos nos permite tener una **estimación consistente de la relación causal entre x y y** cuando x es endógena, y su objetivo no es mejorar la prediccion del modelo.

MC2E -Ejemplo 1 etapa

- Queremos saber los efectos del turismo en Mexico en los ingresos de la ciudad.

$$ln(income_i) = \beta_0 + \beta_1 HotelSales_i + \beta_2 log(temp)_i + u_i$$

- Ciudades mas ricas pueden atraer mas turismo, entonces hay endogeneidad. Usamos: distancia de la ciudad a USA como instrumento y estimamos por MC2E.

Primera etapa

- Estimamos la primera etapa por MCO (instrumentos en X)

$$log(HotelSales)_i = \pi_0 + \pi_1 distUS_i + \pi_2 log(temp)_i + v_i$$

Primera Etapa:

1) Regresion

```
. reg log_hotelsales dist_us_km logtemp
```

Source	SS	df	MS
Model	98.796289	2	49.3981445
Residual	8306.26629	1,150	7.22284025
Total	8405.06258	1,152	7.29606126

Number of obs	=	1,153
F(2, 1150)	=	6.84
Prob > F	=	0.0011
R-squared	=	0.0118
Adj R-squared	=	0.0100
Root MSE	=	2.6875

log_hotels~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist_us_km	-.0010099	.0002958	-3.41	0.001	-.0015902	-.0004296
logtemp	.8949208	.3890481	2.30	0.022	.1315972	1.658244
_cons	2.195632	2.010381	1.09	0.275	-1.748794	6.140057

```
. test dist_us_km

 ( 1)  dist_us_km = 0

       F(  1,  1150) =   11.66
            Prob > F =    0.0007
```

2) Testeo que el instrumento sea relevante. F>10.

```
. predict log_hotelsales_HAT
(option xb assumed; fitted values)
```

Genero una variable que sea XHAT (la prediccion de X basada en Z y temp)

MC2E -Ejemplo 2 etapa: prediccion de X en Y

- Usamos los valores predichos $Hot\hat{e}lSales$ con la primera etapa y estimamos la segunda etapa:

$$ln(Income_i) = \beta_0 + \beta_1 log(Hot\hat{e}lSales)_i + \beta_2 log(temp)_i + u_i$$

Segunda etapa: B_MC2E=0.45

```
. reg log_income log_hotelsales_HAT logtemp
```

Source	SS	df	MS
Model	22.3238319	2	11.161916
Residual	155.178897	1,150	.134938171
Total	177.502728	1,152	.15408223

Number of obs	=	1,153
F(2, 1150)	=	82.72
Prob > F	=	0.0000
R-squared	=	0.1258
Adj R-squared	=	0.1242
Root MSE	=	.36734

log_income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
$\boxed{\text{log\_hotelsales\_HAT}}$	.4544345	.040031	11.35	0.000	.3758924	.5329766
logtemp	-.5510402	.0554067	-9.95	0.000	-.6597497	-.4423307
_cons	7.85789	.3029513	25.94	0.000	7.263491	8.452289

MC2E -Comparamos con MCO

- 0.45 con una estimamos por MCO:

```
. reg log_income log_hotelsales

      Source |       SS           df       MS      Number of obs   =     1,153
-------------+----------------------------------   F(1, 1151)      =    205.29
       Model |  26.8673156         1  26.8673156   Prob > F        =    0.0000
    Residual |  150.635413     1,151  .130873512   R-squared       =    0.1514
-------------+----------------------------------   Adj R-squared   =    0.1506
       Total |  177.502728     1,152   .15408223   Root MSE        =    .36176

------------------------------------------------------------------------------
  log_income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
log_hotelsales |   .0565382    .003946    14.33   0.000      .048796    .0642803
       _cons |   7.421847   .0267578   277.37   0.000     7.369347    7.474347
------------------------------------------------------------------------------
```

- Estimación por MC2E: un incremento de 1% en las ventas de hoteles aumenta 0.45% el ingreso de la ciudad.
- Estimación por MCO: un incremento de 1% de las ventas en hoteles aumenta 0.05% el ingreso en la ciudad. MCO estaba sesgado, subestimando el verdadero valor.

Conclusiones

- Vimos dos métodos para estimar consistentemente los parámetros cuando el regresor es endógeno y contamos con un instrumento válido
  - VI
  - MC2E
- En el MRS, con un sólo instrumento: ambos métodos nos darán las mismas estimaciones puntuales y los mismos errores estándar
- Los resultados serán distintos cuando tengamos más de un instrumento para la variable endógena
- Estimar por estos métodos nos asegura la consistencia (insesgadez) de las estimaciones siempre que los instrumentos sean válidos
- Pero pagaremos el precio de menos precisión en esas estimaciones (errores estándar más grandes que en MCO)
- La próxima clase vamos a ver que pasa cuando tenemos instrumentos débiles, como podemos testear si hay endogeneidad y en algunos casos incluso la restricción de exogeneidad.