Tema 9. Estimación por variables instrumentales

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Generalidades
- Variables instrumentales y validez
- Estimación en el caso de RLS
- VI frente a MCO
- Estimación VI del modelo RLM
- Mínimos cuadrados en dos etapas
- Test de endogeneidad
- Test de sobreidentificación
- Ejercicio aplicado en R: el efecto de la educación sobre el crimen

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 5

- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 8

Generalidades

Recordemos que el modelo clásico de regresión descansa sobre varios supuestos

Sobre las $\mathbf{X}$...
Sobre los $\mathbf{u}$... Uno de estos supuestos era:

$$E(\mathbf{X}'\mathbf{u}) = \mathbf{0} \implies \text{Exogeneidad}$$

¿Qué sucede si se viola este supuesto? Es decir, existencia de **endogeneidad**:

$$E(\mathbf{X}'\mathbf{u}) \neq \mathbf{0}$$

En este caso se dice que X es una **variable explicativa endógena**

Generalidades

Fuentes de endogeneidad:

- Variables omitidas (o heterogeneidad inobservable)
- Errores en las variables
- Simultaneidad (cuando una o más de las variables explicativas se determina conjuntamente con la variable dependiente)

Generalidades

Suponga un modelo de salarios para adultos trabajadores

$$log(wage) = \beta_1 + \beta_2 educ + \beta_3 abil + e$$

Como la habilidad innata $(abil)$ es no observada, entonces se estima el modelo

$$log(wage) = \beta_1 + \beta_2 educ + u$$

donde $u = \beta_3 abil + e$

Se cree que las personas con mayores habilidades innatas suelen alcanzar niveles superiores de educación. Dado que mejores habilidades llevan a salarios más altos, se observa una correlación entre **educación** y factores críticos que afectan el salario

Es posible comprobar que si $\beta_3 \neq 0$, es decir, $abil$ es relevante y $Cov(educ, abil) \neq 0$ el supuesto de exogeneidad en el modelo estimado no se cumple, $E(educ\, u) \neq 0$

Generalidades

Consecuencia:

En ese caso, la estimación por MCO produce estimadores **sesgados e inconsistentes**

Solución??

- Si $abil$ es observable, la solución será incluirla como regresor
- Cuando $abil$ no sea observable, hay que proponer un estimador alternativo...

  ¿Existe un método que permita obtener estimadores consistentes cuando X y u están correlacionadas?

¡Si! Este método reconoce la presencia de endogeneidad y se conoce como **Método de Variables Instrumentales (VI)**

Variables instrumentales y validez

Suponga el modelo de regresión simple

$$Y = \beta_1 + \beta_2 X + u$$

y que X y u están correlacionadas: $Cov(X, u) \neq 0$

Se requiere definir una variable observable, llamémosle z, la cual se denomina **variable instrumental o instrumento** para X que satisface las siguientes condiciones:

- z no está correlacionada con u: $Cov(z, u) = E(zu) = 0 \Longrightarrow$ **Exogeneidad del instrumento** (es una variable exógena)
- z está correlacionada con X: $Cov(z, X) \neq 0 \Longrightarrow$ **Relevancia del instrumento** (está relacionada con la var. explicativa endógena)

Variables instrumentales y validez

- En el caso de una variable endógena y un instrumento, no es posible contrastar si se verifica la condición de exogeneidad $\Longrightarrow$ Usar la intuición y la teoría para decidir si tiene sentido asumir $Cov(z, u) = 0$

- Es posible probar si $Cov(z, X) \neq 0$, mediante una regresión simple entre X y z

$$X = \pi_0 + \pi_1 z + v$$

y probar la hipótesis nula: $H_0 : \pi_1 = 0$ vs. $H_A : \pi_1 \neq 0$

Estimación en el caso de RLS

- Dado $Y = \beta_1 + \beta_2 X + u$ y z un instrumento válido: $Cov(z, u) =$ E$(zu)=0$

$$Cov(z, Y) = \beta_2 Cov(z, X) + Cov(z, u)$$

$$\implies \beta_2 = Cov(z, Y)/Cov(z, X)$$

Así, el estimador de VI para β_2 es:

$$\widehat{\beta}_{2VI} = \frac{\sum(z_i - \bar{z})(Y_i - \bar{Y})}{\sum(z_i - \bar{z})(X_i - \bar{X})}$$

- El supuesto de homocedasticidad en este caso es $E(u^2/z) = \sigma^2 = Var(u)$ y la varianza asintótica estimada está dada por:

$$Var(\widehat{\beta}_{2VI}) = \frac{\widehat{\sigma}^2}{STC_X R^2_{X,z}}$$

$R^2_{X,z}$ mide la fortaleza de la relación lineal entre X y z en la muestra

VI frente a MCO

- La varianza del estimador VI difiere de la MCO, en el R^2 que se obtiene de regresar X sobre z
- Dado que $R^2 < 1$, la varianza de VI siempre es mayor que la varianza de MCO (cuando MCO es válido, este es el costo de realizar la estimación de VI cuando X y u no se correlacionan)
- Cuanto mayor sea la correlación entre z y X, menor será la varianza de VI
- Sin embargo, en el caso en que $cov(X, u) \neq 0$, VI es consistente, mientras MCO es inconsistente

Estimación VI del modelo RLM

- La estimación VI puede ser extendida con facilidad al caso de regresión múltiple
- En este caso podemos tener una o más variables que son endógenas
- Necesitamos variables instrumentales (tantas como variables endógenas entre los regresores)
- Es posible tener múltiples instrumentos

Mínimos cuadrados en dos etapas (MC2E)

- Considere el modelo de interés

$$y_1 = \beta_1 + \beta_2 y_2 + \beta_3 z_1 + u_1$$

donde y_2 es endógena y z_1 es exógena

- Asumimos que z_2 y z_3 son instrumentos válidos. En este caso, el mejor instrumento es una combinación lineal de todas las variables exógenas

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

donde $\pi_2 \neq 0$ o $\pi_3 \neq 0$

Mínimos cuadrados en dos etapas (MC2E)

El método consiste en dos etapas:

- Primera etapa: Estimar y_2^*, regresando y_2 sobre z_1, z_2 y z_3 y obtenemos los valores predichos $\hat{y}_2$
- Segunda etapa: Sustituimos y_2 por $\hat{y}_2$ en el modelo de interés y estimamos por MCO, los estimadores asi obtenidos se conocen como estimadores de MC2E
- El método se extiende a múltiples variables endógenas. Es necesario asegurarse de que tenemos al menos tantas variables exógenas excluidas (instrumentos) como variables endógenas en la ecuación de interés

Test de endogeneidad

- Si no existe endogeneidad en el modelo original, MCO es preferido a VI, por tanto, es necesario probar la existencia de endogeneidad
- Si no existe endogeneidad, MCO y VI son consistentes, pero MCO es más eficiente (óptimo)
- Si existe endogeneidad, solamente VI es consistente
- Por tanto, es importante realizar un contraste de endogeneidad para evitar usar VI cuando no es necesario
- Se utiliza el test de Hausman (H_0: Exogeneidad)

Test de endogeneidad

H_0: y_2 es exógena, H_A: y_2 es endógena

Pasos:

- Guarde los residuos de la primera etapa: $\hat{v}_i$
- Incluya $\hat{v}_i$ en la ecuación principal (contiene y_2):

$$y_{1i} = \beta_1 + \beta_2 y_{2i} + \beta_3 z_{1i} + \delta \hat{v}_i + u_{1i}$$

- Si el coeficiente asociado al residuo es estadísticamente diferente de cero, rechace la hipótesis nula de exogeneidad

  Contrastamos: H_0: $\delta = 0$ frente a H_A: $\delta \neq 0$

  Si no rechazamos H_0, no rechazaremos que y_2 es exógena

Test de sobreidentificación o test de Sargan

- Si solo hay un instrumento para nuestra variable endógena, no podemos probar si el instrumento no está correlacionado con el error. Decimos que el modelo está identificado
- Si tenemos varios instrumentos, es posible probar las restricciones de sobreidentificación, para ver si algunos de los instrumentos están correlacionados con el error
- Pasos:
  - Estime el modelo de interés usando MC2E y obtenga los residuos: $\widehat{u}_i^{MC2E}$
  - Regrese los $\widehat{u}_i^{MC2E}$ sobre todas las variables exógenas (incluido los instrumentos) y obtenga el R^2 para calcular nR^2
  - Bajo la hipótesis nula que todos los instrumentos no están correlacionados con el error, $LM \sim \chi^2_q$, donde q es el número de instrumentos adicionales

Ejercicio aplicado en R: el efecto de la educación sobre el crimen

La educacion reduce el crimen? Si es así, gastar más en educacion podría ser una herramienta de largo plazo para combatir el crimen. Este ejercio aplicado se basa en el *paper* de Lochner, L. y Moretti, E. (2004). "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports", American Economic Review, 94(1):155-189.

Las principales variables para el análisis son:
$prision$: variable binaria igual a 1 si la persona esta en prision, 0 no
$educ$: años de escolaridad
age: edad
$AfAm$: variable binaria igual a 1 para afroamericano, 0 no
$ca8$: variable binaria igual a 1 si la escolaridad obligatoria estatal es de 8 o menos años
$ca9$: variable binaria igual a 1 si la escolaridad obligatoria estatal es de 9 años
$ca10$: variable binaria igual a 1 si la escolaridad obligatoria estatal es de 10 años
$ca11$: variable binaria igual a 1 si la escolaridad obligatoria estatal es de 11 o mas años

La idea es estimar el siguiente modelo

$$prison = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 AfAm + u$$

En el siguiente link se encuentra la base de datos y el código en R utilizado:

- Datos
- Código en R