PUBLICACIONES DE 4er CURSO

Curso: 4º

Grado: Economía

Asignatura: ECONOMETRÍA III

TRANSPARENCIAS PARTE 3: OTROS TEMAS

TEMA 8: VARIABLES INSTRUMENTALES

Profesores: Antonio Aznar

Departamento de ANÁLISIS ECONÓMICO

Facultad de Economía y Empresa
Universidad Zaragoza

Curso Académico 2015/16

Tema 8. Variables Instrumentales

Índice

1. Definición de Variable Instrumental
2. Criterios para validar la variable instrumental
3. Mínimos cuadrados en dos etapas

1. Definición Variable Instrumental

Tres amenazas importantes para la validez interna son:

- El sesgo de variable omitida por una variable que está correlacionada con X pero no se observa (por lo que no puede incluirse en la regresión) y para la que no hay variables de control adecuadas;
- Sesgo de causalidad simultánea (X causa Y, Y causa X);
- Sesgo de errores en variables (X se mide con error)

Los tres acaban en $E(u|X) \neq 0$.

La regresión de variable instrumental puede eliminar el sesgo utilizando una variable instrumental (IV), Z.

El estimador de VI con un solo regresor y un solo instrumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- La regresión de VI distingue en X dos partes: una parte que puede estar correlacionada con u, y otra que no lo está. Aislando la parte que no está correlacionada con u, es posible estimar consistentemente β_1.
- Esto se hace utilizando una **variable instrumental** , Z_i, que está correlacionada con X_i pero no con u_i.

Terminología**: Endogeneidad y Exogeneidad**

Una variable ***endógena*** es la que está correlacionada con *u*
Una variable ***exógena*** *es la que no está correlacionada con u*

En la regresión de variable instrumental nos centramos en el caso en que *X* es endógena y hay un instrumento, *Z*, que es exógeno.

Nota sobre terminología: "Endógena" literalmente significa "determinada dentro del sistema" Si *X* está determinada conjuntamente con *Y*, entonces una regresión de *Y* sobre *X* nos lleva al sesgo de causalidad simultanea. Pero esta definición de endogeneidad es demasiado estrecha porque la regresión de VI puede usarse para tratar los sesgos de VO y de errores en variables.

Las dos condiciones para validar un instrumento

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Para que una variable instrumental (un "***instrumento***") Z sea válido, debe satisfacer dos condiciones:

1. ***Relevancia instrumental***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Exogeneidad instrumental*** : $\text{corr}(Z_i, u_i) = 0$

Suponer por ahora que tu tienes tal variable Z_i (discutiremos posteriormente como encontrar variables instrumentales).
¿Cómo podemos usar Z_i para estimar β_1?

El estimador de VI con un solo regresor y un solo instrumento

Explicación #1:Mínimos cuadrados en dos etapas (MC2E). Two Stage Least Squares (TSLS)

Como suena, el MC2E tiene dos etapas – dos regresiones:

(1) Aislar la parte de X *que no está correlacionada con* u haciendo la regresión de X sobre Z utilizando MCO:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Como Z_i no está correlacionada con u_i, $\pi_0 + \pi_1 Z_i$ no está correlacionada con u_i. No conocemos π_0 o π_1 pero los podemos estimar, de forma que,…
- Calcular los valores predichos de X_i, $\hat{X}_i$, donde $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

MC2E

(2) Sustituir X_i por $\hat{X}_i$ y hacer la regresión de Y sobre $\hat{X}_i$ utilizando MCO:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- Como $\hat{X}_i$ no está correlacionada con u_i, la primera hipótesis para los mínimos cuadrados se cumple para la regresión (2). (Es necesario que n sea grande de forma que π_0 y π_1 sean estimados con precisión)
- Así, en muestras grandes, β_1 puede ser estimado con los MCO utilizando la regresión (2)
- El estimador resultante es llamado el estimador de los Mínimos Cuadrados en dos Etapas, $\hat{\beta}_1^{MC2E}$ [the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$].

MC2E : Resumen

Suponer que Z_i, cumple las dos condiciones para ser un instrumento válido:

1. ***Relevancia***: corr(Z_i,X_i) $\neq$ 0
2. ***Exogeneidady***: corr(Z_i,u_i) = 0

Mínimos Cuadrados en Dos Etapas:

Etapa 1: Regresar X_i sobre Z_i (incluyendo una constante), y obtener los valores predichos, $\hat{X}_i$

Etapa 2: Regresar Y_i sobre $\hat{X}_i$ (incluyendo una constante); la estimación del coeficiente de $\hat{X}_i$ es el estimador MC2E,

$$\hat{\beta}_1^{MC2E} = \frac{\text{cov}(\hat{X}_i, Y_i)}{Var(\hat{X}_i)}.$$

$\hat{\beta}_1^{MC2E}$ es un estimador consistente de β_1.

Estimador de VI

Explicación #2: Derivación Algebraica directa

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Por lo que,

$$\begin{aligned}
\text{cov}(Y_i, Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0, Z_i) + \text{cov}(\beta_1 X_i, Z_i) + \text{cov}(u_i, Z_i) \\
&= \quad 0 \quad + \text{cov}(\beta_1 X_i, Z_i) + \quad 0 \\
&= \beta_1 \text{cov}(X_i, Z_i)
\end{aligned}$$

donde $\text{cov}(u_i, Z_i) = 0$ por la exogeneidad del instrumento; así

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

El estimador de VI sustituye las covarianzas poblacionales con las covarianzas muestrales:

$$\hat{\beta}_1^{MC2E} = \frac{s_{YZ}}{s_{XZ}},$$

s_{YZ} y s_{XZ} son las covarianzas muestrales. Este es el mismo MC2E – solo cambia la derivación.

Los Estimadores MC2E y VI coinciden

Basta tener en cuenta que:

$$\hat{\beta}_1^{MC2E} = \frac{s_{\hat{X}Y}}{S_{\hat{X}\hat{X}}} = \frac{\hat{\pi}_1 S_{ZY}}{\hat{\pi}_1^2 S_{ZZ}} = \frac{S_{ZY}}{\frac{S_{ZX}}{S_{ZZ}} S_{ZZ}} = \frac{S_{ZY}}{S_{ZX}}$$

teniendo en cuenta que,

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

Consistencia del Estimador IV o MC2E

$$\hat{\beta}_1^{MC2E} = \frac{s_{YZ}}{s_{XZ}}$$

Las covarianzas muestrales son consistentes: $s_{YZ} \xrightarrow{p} \text{cov}(Y,Z)$ y $s_{XZ} \xrightarrow{p} \text{cov}(X,Z)$. Por lo tanto,

$$\hat{\beta}_1^{MC2E} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)} = \beta_1$$

- La condición de relevancia del instrumento, $\text{cov}(X_i,Z_i) \neq 0$, asegura que no se divida por cero.

Ejemplo : Demanda de cigarrillos

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Porque el estimador MCO de β_1 puede estar sesgado?

- Banco de datos: datos panel sobre el consumo anual de cigarrillos y los precios medios pagados (incluido el impuesto), para cada uno de los 48 estados continentales de USA, 1985-1995.
- Variable instrumental propuesta:
  - Z_i = impuesto general de ventas por paquete en cada estado = $SalesTax_i$
  - ¿Piensas que el instrumento es válido?

    (1) ¿Relevancia? corr($SalesTax_i$, $\ln(P_i^{cigarettes})$) $\neq$ 0?

    (2) ¿Exógena? corr($SalesTax_i$,u_i) = 0?

Ejemplo : Demanda de cigarrillos

Por ahora, utilizar datos solo para 1995.

Primera etapa:

$$\widehat{\ln(P_i^{cigarettes})} = 4.63 + .031 SalesTax_i,\ n = 48$$

Segunda etapa:

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08\ \widehat{\ln(P_i^{cigarettes})},\ n = 48$$

Ejemplo . Salida Gretl

Modelo 4: MCO, usando las observaciones 1-48

Variable dependiente: $\widehat{\ln(P_i^{cigarettes})}$

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	4.61655	0.0291078	158.6	1.20e-064	***
SalesTax	0.0307289	0.00480163	6.400	7.27e-08	***

Media de la vble. dep.	4.781380	D.T. de la vble. dep.	0.127783
Suma de cuad. residuos	0.405979	D.T. de la regresión	0.093945
R-cuadrado	0.470996	R-cuadrado corregido	0.459496
F(1, 46)	40.95588	Valor p (de F)	7.27e-08
Log-verosimilitud	46.43466	Criterio de Akaike	-88.86933
Criterio de Schwarz	-85.12693	Crit. de Hannan-Quinn	-87.45507

Ejemplo . Salida Gretl

Modelo 5: MCO, usando las observaciones 1-48

Variable dependiente: $\widehat{\ln(Q_i^{cigarettes})}$

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	9.71988	1.80120	5.396	2.30e-06	***
yhat4	-1.08359	0.376649	-2.877	0.0061	***

Media de la vble. dep.	4.538837	D.T. de la vble. dep.	0.243346
Suma de cuad. residuos	2.358809	D.T. de la regresión	0.226447
R-cuadrado	0.152490	R-cuadrado corregido	0.134066
F(1, 46)	8.276648	Valor p (de F)	0.006069
Log-verosimilitud	4.204011	Criterio de Akaike	-4.408022
Criterio de Schwarz	-0.665620	Crit. de Hannan-Quinn	-2.993763

- Estas son las estimaciones MC2E
- Los errores estándar no son exactos porque no tienen en cuenta las dos etapas

Resumen

- Para que un instrumento Z sea válido debe satisfacer dos condiciones:

  (1) *relevancia*: $\text{corr}(Z_i, X_i) \neq 0$

  (2) *exogeneidad*: $\text{corr}(Z_i, u_i) = 0$

- Los MC2E primero hacen la regresión de X sobre Z para obtener $\hat{X}$, a continuación se hace la regresión de Y sobre $\hat{X}$
- La idea clave es que en la primera etapa se aisla la parte de X que no está correlacionada con u.
- Si el instrumento es válido entonces la distribución en muestras grandes de los MC2E siguen una distribución normal y se puede hacer la inferencia como ya se ha visto.

EXTENSIONES. MODELO GENERAL

El modelo general de VI es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{ri} + u_i \quad (1)$$

$X_{1i}...X_{ki}$ son k regresores endógenos potencialmente correlacionados con u_i

$W_{1i}...W_{ri}$ son r regresores exógenos incluidos que no están correlacionados con u_i

$Z_{1i}...Z_{mi}$: son m variables instrumentalEs

Los coeficientes $\beta_1 ... \beta_k$ están sobreidentificados si existen más instrumentos que regresores endógenos y están exactamente identificados si m = k. Si no están identificados los parámetros no se pueden estimar.

Ejemplo : Demanda de cigarrillos (extensiones)

Suponer que la renta es exógena (esto es plausible – ¿por qué?), y queremos también estimar la elasticidad de renta:

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

Tenemos dos instrumentos:

Z_{1i} = impuesto general de ventas

Z_{2i} = impuesto específico sobre el tabaco

- Variable endógena: $\ln(P_i^{cigarettes})$ ("una X")
- Variable exógena incluida: $\ln(Income_i)$ ("una W")
- Instrumentos (excluidas variables endógenas): impuesto general sobre ventas, impuesto específico sobre tabaco ("dos Zs")

- ¿Está β_1 sobre–, *sub–, o Exactamente identificado?*

Un instrumento: Impuesto General sobre las venta**s.**

La salida de Gretl es:

Modelo 2: MCO combinados, utilizando 48 observaciones

Variable dependiente: $\widehat{\ln(Q_i^{cigarettes})}$

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	9.43066	1.62471	5.805	6.09e-07	***
linco	0.214515	0.321247	0.6678	0.5077	
yhat1	-1.14338	0.429972	-2.659	0.0108	**

Media de la vble. dep.	4.538837	D.T. de la vble. dep.	0.243346
Suma de cuad. residuos	2.313601	D.T. de la regresión	0.226745
R-cuadrado	0.168733	R-cuadrado corregido	0.131788
F(2, 45)	4.567115	Valor p (de F)	0.015637
Log-verosimilitud	4.668448	Criterio de Akaike	-3.336895
Criterio de Schwarz	2.276708	Crit. de Hannan-Quinn	-1.215506

Linco: log de la renta real per-cápita de cada estado
Yhat1: valores ajustados en la primera etapa del log del precio de cigarrillos en términos reales utilizando en la regresión la variable instrumental (impuesto general sobre ventas) y linco.

Ejemplo: Demanda de Cigarrillos

Dos instrumentos:

- Impuesto general sobre ventas
- Impuesto específico sobre el tabaco

Modelo 4: MCO combinados, utilizando 48 observaciones
Se han incluido 48 unidades de sección cruzada
Largura de la serie temporal = 1

Variable dependiente: $\widehat{\ln(Q_i^{cigarettes})}$

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	9.89496	1.14126	8.670	3.72e-011	***
linco	0.280405	0.257203	1.090	0.2814	
yhat3	-1.27742	0.283761	-4.502	4.73e-05	***

Media de la vble. dep.	4.538837	D.T. de la vble. dep.	0.243346
Suma de cuad. residuos	1.845868	D.T. de la regresión	0.202532
R-cuadrado	0.336787	R-cuadrado corregido	0.307311
F(2, 45)	11.42578	Valor p (de F)	0.000097
Log-verosimilitud	10.08898	Criterio de Akaike	-14.17796
Criterio de Schwarz	-8.564362	Crit. de Hannan-Quinn	-12.05658

Estimaciones MC2E, Z = Impuesto ventas (m = 1)

$$\widehat{\ln(Q_i^{cigarettes})} = 9.43 - 1.14\widehat{\ln(P_i^{cigarettes})} + 0.21\ln(Income_i)$$
$$(1.26)\quad (0.37)\qquad\qquad (0.31)$$

Estimaciones MC2E , Z = Impuesto Ventas & impuesto tabaco (m = 2)

$$\widehat{\ln(Q_i^{cigarettes})} = 9.89 - 1.28\widehat{\ln(P_i^{cigarettes})} + 0.28\ln(Income_i)$$
$$(0.96)\quad (0.25)\qquad\qquad (0.25)$$

- Menor *SE* para m = 2. Si se utilizan dos instrumentos se tiene más información.

- La elasticidad de la renta es baja(no es un bien de lujo) y, estadísticamente, no puede aceptarse que sea diferente de cero
- Sorprende la alta elasticidad precio.

2. Criterios para validar las variables instrumentales

Recordar que los dos requisitos para ser instrumentos válidos

1. *Relevancia* (caso especial de un X)

   Al menos un instrumento debe entrar en la regresión de la primera etapa.

2. *Exogeneidad*

   Ningún instrumento está correlacionado con el error:

   $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_{mi},u_i) = 0$

¿Qué ocurre si uno de estos requisitos no se satisface?
¿Cómo se puede chequear? ¿Qué se hace?
Si tienes varios instrumentos, ¿cuáles de ellos se utilizan?

Consideremos el caso de una sola X. Los dos requisitos para la validez de los instrumentos son:

1. ***Relevancia***
   - Al menos disponer de un instrumento para la regresión de la primera etapa.
   - Si los instrumentos son débiles entonces el estimador MC2E es sesgado y el estadístico t sigue una distribución diferente a la normal.
   - Para chequear la posible debilidad de los instrumentos con un solo regresor endógeno incluido utilizamos el estadístico-F,
     - Si $F>10$, los instrumentos son fuertes – usar MC2E
     - Si $F<10$, los instrumentos son débiles – replantear.

2. *Exogeneidad de los instrumentos*

Sólo se puede contrastar cuando hay sobreidentificación porque el contraste se basa en comparar diferentes estimadores en dos etapas.

El contrate de exogeneidad es como un contraste de sobreidentificación de restricciones y lo llamaremos el contraste J.

Sea $\hat{u}_i^{MC2E}$ el residuo de la estimación MC2E de la ecuación (1). Se utiliza MCO para la estimación de los coeficientes de la regresión en:

$$\hat{u}_i^{MC2E} = \delta_0 + \delta_1 Z_{1i} + \ldots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \ldots + \delta_{m+r} W_{ri} + e_i$$

Sea F la expresión del estadístico F válido con homocedasticidad para el contraste de la hipótesis de que $\delta_1 = ... = \delta_m = 0$. El estadístico para el contraste de sobreidentificación de restricciones es $J = mF$. Bajo la hipótesis nula de que todos los instrumentos son exógenos, si e_i es homocedástico, en muestras grandes J se distribuye como $\chi^2(m-k)$, donde *m-k* es el "grado de sobreidentificación", es decir, el número de instrumentos menos el número de regresores endógenos.

CONTRASTE DE EXOGENEIDAD

Todo lo anterior depende del diagnóstico que se haga sobre la exogeneidad de la variable de interés.

Una forma de contrastar la exogeneidad es mediante el contraste de Hausman.

Este contraste está basado en el estadístico:

$$H = \frac{\left(\hat{\beta}_1^{MCO} - \hat{\beta}_1^{MC2E}\right)^2}{\hat{Var}(\hat{\beta}_1^{MCO} - \hat{\beta}_1^{MC2E})}$$

El estimador $\hat{\beta}_1^{MCO}$ es consistente y asintóticamente eficiente si se cumple la hipótesis nula, pero si no se cumple es inconsistente. $\hat{\beta}_1^{MC2E}$ es consistente tanto si se cumple H_0 como si no se cumple. Pero no es eficiente.

El estadístico H bajo la hipótesis nula sigue una distribución $\chi^2(1)$.

La región crítica viene dada por:

$$H > \chi^2(1)$$