

Clase auxiliar # 9

Endogeneidad y variables instrumentales

¡Recordar!

Causales de endogeneidad: (1) omisión de variables relevantes, (2) existencia de errores de medición no constantes, (3) determinación simultánea de ecuaciones y (4) auto-correlación de los errores (estudiaremos esto cuando lleguemos a series de tiempo).

Identificación: se dice que una ecuación está **sub/sobre/exactamente-identificada** cuando la cantidad de variables presentes en el modelo pero ausentes en la ecuación es **menor/mayor/igual** a la cantidad de variables endógenas del modelo menos 1. $k_{\text{ecuación}} = g_{\text{ecuación}} + e_{\text{ecuación}} (< / > / =) g - 1$

Mínimos Cuadrados en 2 Etapas (MC2E): supongamos que se quiere estimar una ecuación de la forma $Y = X_1\beta_1 + X_2\beta_2 + U$, donde X_1 representa uno o más regresores endógenos. Además se tiene uno o más instrumentos Z que cumplen los supuestos de relevancia y exclusión para la ecuación que se quiere estimar. Para poder usar este método se requiere el la ecuación esté al menos exactamente identificada.

- **Etapla 0:** darse cuenta del problema de endogeneidad dado por X_1 .
- **Etapla 1:** llamaremos “matriz de instrumentos expandida” a $\tilde{Z} = [X_2', Z']$. Regresionar $X_1 = \tilde{Z}\delta + \varepsilon$ para obtener una proyección $\hat{X}_1$ de X_1 ortogonal a U .
- **Etapla 2:** reemplazar en la regresión original X_1 por $\hat{X}_1$, regresionando $Y = \hat{X}_1\tilde{\beta}_1 + X_2\tilde{\beta}_2 + \tilde{U}$, obteniendo el estimador $\hat{\beta}_{MC2E}$.

Supuestos de variables instrumentales: los instrumentos Z deben cumplir los supuestos de relevancia/fortaleza/no-debilidad ($Cov(Z, X_1) \neq 0$) y exclusión/validez ($Cov(Z, U) = 0$).

Estimador de variables instrumentales: Siguiendo la misma notación que en el punto anterior $\hat{\beta}_{VI} = (\tilde{Z}'X)^{-1}\tilde{Z}'Y$. Este estimador solo se puede obtener cuando la ecuación de interés está exactamente identificada¹ En dicho caso $\hat{\beta}_{VI} = \hat{\beta}_{MC2E}$.

1. Endogeneidad

1. Muestre que existe endogeneidad en estas tres situaciones:

- Cuando hay variables omitidas relevantes.

Respuesta:

¹En el caso sobreidentificado no cuadran las dimensiones de las matrices.

- **Modelo verdadero:** $Y_i = \alpha + \beta x_i + \gamma z_i + u_i$.
- **Modelo estimado:** $Y_i = \alpha + \beta x_i + v_i$.

Notemos que existe endogeneidad al omitir variables relevantes:

$$Cov(x_i, v_i) = Cov(x_i, \gamma z_i + u_i) = \gamma Cov(x_i, z_i) + \underbrace{Cov(x_i, u_i)}_{=0} = \gamma Cov(x_i, z_i) \neq 0$$

Notar que el término $Cov(x_i, z_i)$ no tiene por qué ser cero, aunque podría serlo. Cuando es cero, se tiene que no hay correlación entre la variable observable y no observable. En este caso particular, no hay problema de endogeneidad. Como hemos visto al estudiar omisión de variables relevantes la no-existencia de correlación entre la variable observable y no observable anula el sesgo por omisión.

Notar que suponiendo que el modelo “verdadero” está bien especificado, x_i es exógena (no presenta problemas de endogeneidad). Por definición, esto significa que $Cov(x_i, u_i) = 0$.

Es así como la omisión de variables omitidas es un caso particular de endogeneidad.

- Cuando existen errores de medición no constantes. Por simplicidad puede suponer errores de medición exógenos para esta parte.

Respuesta:

- **Modelo verdadero:** $Y_i = \alpha + \beta x_i + \varepsilon_i$.
- **Modelo estimado:** $Y_i = \alpha + \beta x_i^* + u_i$, donde $x_i^* = x_i + \delta_i$.

$$Y_i = \alpha + \beta x_i^* + u_i = \alpha + \beta(x_i + \delta_i) + u_i = \alpha + \beta x_i + \underbrace{(\beta \delta_i + u_i)}_{\varepsilon_i}$$

$$\begin{aligned} Cov(x_i^*, u_i) &= Cov(x_i + \delta_i, \varepsilon_i - \beta \delta_i) \\ &= Cov(x_i + \delta_i, \varepsilon_i) + Cov(x_i + \delta_i, -\beta \delta_i) \\ &= \underbrace{Cov(x_i, \varepsilon_i)}_{=0} + \underbrace{Cov(\delta_i, \varepsilon_i)}_{=0} + Cov(x_i + \delta_i, -\beta \delta_i) \\ &= Cov(x_i + \delta_i, -\beta \delta_i) \\ &= -\beta Cov(x_i + \delta_i, \delta_i) \\ &= -\beta [Cov(x_i, \delta_i) + \mathbb{V}(\delta_i)] \neq 0 \end{aligned}$$

Notar que es posible que $Cov(x_i, \delta_i) + \mathbb{V}(\delta_i) = 0$. Este sería un caso puntual en el que no habría problemas de endogeneidad en el modelo estimado.

Suponiendo que el modelo “verdadero” está bien planteado $Cov(x_i, \varepsilon_i) = 0$. Por exogeneidad del error de medición (supuesto simplificador), $Cov(\delta_i, \varepsilon_i) = 0$.

- Cuando dos variables se determinan simultáneamente.

Respuesta:

- **Modelo 1:** $y_i = \alpha_0 + \alpha_1 z_i + \alpha_2 x_{1i} + u_i$.
- **Modelo 2:** $z_i = \beta_0 + \beta_1 y_i + \beta_2 x_{2i} + v_i$.

$$\begin{aligned}
 Cov(z_i, u_i) &= Cov(\beta_0 + \beta_1 y_i + \beta_2 x_{2i} + v_i, u_i) \\
 &= \beta_1 Cov(y_i, u_i) \\
 &= \beta_1 Cov(\alpha_0 + \alpha_1 z_i + \alpha_2 x_{1i} + u_i, u_i) \\
 &= \beta_1 [\alpha_1 Cov(z_i, u_i) + \mathbb{V}(u_i)] \\
 \Rightarrow Cov(z_i, u_i) &= \frac{\beta_1 \mathbb{V}(u_i)}{1 - \alpha_1 \beta_1} \neq 0
 \end{aligned}$$

2. Muestre que en presencia de endogeneidad $\hat{\beta}$ es sesgado.

Respuesta:

$$\begin{aligned}
 \hat{\beta} &= (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + U) = \beta + (X^T X)^{-1} X^T U \\
 \Rightarrow \mathbb{E}(\hat{\beta}|X) &= \beta + \underbrace{(X^T X)^{-1} X^T \mathbb{E}(U|X)}_{\text{sesgo}}
 \end{aligned}$$

Ahora debemos demostrar que el sesgo calculado es distinto de cero.

Demostremos que $\mathbb{E}(X^T U) \neq 0 \Rightarrow \mathbb{E}(U|X) \neq 0$.

Por contrarecíproca demostraremos que $\mathbb{E}(U|X) = 0 \Rightarrow \mathbb{E}(X^T U) = 0$.

Utilizando la ley de las esperanzas iteradas, se obtiene esta propiedad.

$$\begin{aligned}
 \mathbb{E}(U|X) = 0 &\Rightarrow \mathbb{E}(X^T U) = \mathbb{E}[\mathbb{E}(X^T U|X)] = \mathbb{E}[X^T \underbrace{\mathbb{E}(U|X)}_{=0}] = 0 \\
 \mathbb{E}(U|X) \neq 0 &\Rightarrow \underbrace{(X^T X)^{-1} X^T \mathbb{E}(U|X)}_{\text{sesgo}} \neq 0
 \end{aligned}$$

3. Mostrar que en presencia de endogeneidad $\hat{\beta}$ es inconsistente.

Respuesta:

$$\hat{\beta} = \beta + \left(\frac{1}{N} X^T X \right)^{-1} \left(\frac{1}{N} X^T U \right)$$

Utilizando los teoremas de Slutsky y de mapeo continuo...

$$\hat{\beta} \xrightarrow{p} \beta + \underbrace{\mathbb{E}(X^T X)^{-1} \mathbb{E}(X^T U)}_{\neq 0} \neq \beta$$

2. Identificación y mínimos cuadrados indirectos: oferta y demanda

Considere el mercado de las papas, descrito por las siguientes propiedades:

- **Curva de oferta:** $Q_t^s = \alpha_0 + \alpha_1 P_t + \alpha_2 L_t + U_t$.
 - **Curva de demanda:** $Q_t^d = \beta_0 + \beta_1 P_t + \beta_2 I_t + V_t$.
 - **Variables exógenas:** L_t es la lluvia caída e I_t el ingreso nominal.
 - Asuma que el mercado de las papas se rige por los supuestos de competencia perfecta y que las muestras que se tienen fueron tomadas en equilibrio ($Q_t^d = Q_t^s$).
- a. Muestre que cantidades y precios son variables endógenas.

Respuesta:

Se sabe que las muestras fueron tomadas en equilibrio, por lo que oferta y demanda se igualan. A continuación se despeja P_t en función de I_t , L_t , U_t y V_t . Esto nos permitirá calcular las covarianzas que nos permitirán demostrar que las variables señaladas son endógenas.

$$\begin{aligned} Q_t^d &= Q_t^s \\ \Rightarrow \alpha_0 + \alpha_1 P_t + \alpha_2 L_t + U_t &= \beta_0 + \beta_1 P_t + \beta_2 I_t + V_t \\ \Rightarrow P_t &= \frac{\beta_0 + \beta_2 I_t + V_t - \alpha_0 - \alpha_2 L_t - U_t}{\alpha_1 - \beta_1} \end{aligned}$$

A continuación se calculan las covarianzas²:

$$\bullet \text{Cov}(P_t, U_t | L, I) = \text{Cov} \left(\frac{\beta_0 + \beta_2 I_t + V_t - \alpha_0 - \alpha_2 L_t - U_t}{\alpha_1 - \beta_1}, U_t \middle| L, I \right) = \frac{-\sigma_U^2}{\alpha_1 - \beta_1} \neq 0$$

²Notar que usaremos para los cálculos el hecho de que la lluvia y los ingresos son variables exógenas. Esto implica que no están correlacionadas con los errores correspondientes a las ecuaciones en las que están presentes.

- $Cov(P_t, V_t|L, I) = Cov(\frac{\beta_0 + \beta_2 I_t + V_t - \alpha_0 - \alpha_2 L_t - U_t}{\alpha_1 - \beta_1}, V_t|L, I) = \frac{\sigma_V^2}{\alpha_1 - \beta_1} \neq 0$
- $Cov(Q_t^s, U_t|L, I) = Cov(\alpha_0 + \alpha_1 P_t + \alpha_2 L_t + U_t, U_t|L, I) = \alpha_1 Cov(P_t, U_t|L, I) + Cov(U_t, U_t|L, I) = \frac{-\alpha_1 \sigma_U^2}{\alpha_1 - \beta_1} + \sigma_U^2 = \frac{-\beta_1 \sigma_U^2}{\alpha_1 - \beta_1} \neq 0$
 Notar que puede darse el caso en el que esta covarianza sea cero (si $\beta_1 = 0$) y por tanto Q_t^s no sea endógena, sin embargo es muy poco probable.
- $Cov(Q_t^d, V_t|L, I) = Cov(\beta_0 + \beta_1 P_t + \beta_2 I_t + V_t, V_t|L, I) = \beta_1 Cov(P_t, V_t|L, I) + Cov(V_t, V_t|L, I) = \frac{\beta_1 \sigma_V^2}{\alpha_1 - \beta_1} + \sigma_V^2 = \frac{\alpha_1 \sigma_V^2}{\alpha_1 - \beta_1} \neq 0$

b. ¿El modelo está sobre, sub o exactamente identificado?

Respuesta:

Para responder a esta pregunta, en primer lugar se calcula la cantidad de variables endógenas totales g , las cantidades de variables endógenas presentes g_E , las cantidades de variables endógenas ausentes g_{-E} , las cantidades de variables exógenas presentes e_E , las cantidades de variables exógenas ausentes e_{-E} y la cantidad de variables totales ausentes k_E en cada ecuación $E \in \{\text{Oferta, Demanda}\}$.

g	$= 3$	(Q_t^d, Q_t^s, P_t)
g_{oferta}	$= 2$	(Q_t^s, P_t)
$g_{-oferta}$	$= 1$	(Q_t^d)
$g_{demanda}$	$= 2$	(Q_t^d, P_t)
$g_{-demanda}$	$= 1$	(Q_t^s)
e_{oferta}	$= 1$	(L_t)
$e_{-oferta}$	$= 1$	(I_t)
$e_{demanda}$	$= 1$	(I_t)
$e_{-demanda}$	$= 1$	(L_t)

A partir de esta información se puede calcular la cantidad de variables totales que están presentes en el modelo, pero ausentes en cada ecuación.

$$k_{demanda} = g_{-demanda} + e_{-demanda} = 1 + 1 = 2$$

$$k_{oferta} = g_{-oferta} + e_{-oferta} = 1 + 1 = 2$$

Por otro lado, $g - 1 = 3 - 1 = 2$. De esta manera, se concluye que ambas ecuaciones están exactamente identificadas, pues $k_{demanda} = g - 1$ y $k_{oferta} = g - 1$.

c. Dada su respuesta en la parte (b), ¿es posible estimar los parámetros (α, β) sin usar variables instrumentales?

Respuesta:

En este caso sí, pues las ecuaciones están exactamente identificadas. Se puede utilizar el método de mínimos cuadrados indirectos, método que se explica a continuación.

En primer lugar, es importante notar que se puede estimar P_t y Q_t , en función de las variables exógenas I_t y L_t , de la siguiente forma:

$$\begin{aligned}P_t &= \gamma_0 + \gamma_1 I_t + \gamma_2 L_t + \tilde{U}_t \\Q_t &= \delta_0 + \delta_1 I_t + \delta_2 L_t + \tilde{V}_t\end{aligned}$$

Donde $\gamma_0 = \frac{\alpha_0 - \beta_0}{\beta_1 - \gamma_1}$, $\gamma_1 = \frac{-\beta_2}{\beta_1 - \gamma_1}$, $\gamma_2 = \frac{\alpha_2}{\beta_1 - \gamma_1}$, $\delta_0 = \frac{\alpha_0 \beta_1 - \alpha_1 \beta_0}{\beta_1 - \gamma_1}$, $\delta_1 = \frac{-\alpha_1 \beta_2}{\beta_1 - \gamma_1}$, $\delta_2 = \frac{\alpha_2 \beta_1}{\beta_1 - \gamma_1}$ (*).

Al estimar las ecuaciones para P_t y Q_t , se obtienen los parámetros (γ, δ) . Sin embargo, nuestros parámetros de interés son (α, β) . La pregunta clave es: ¿podemos estimar (α, β) a partir de las estimaciones que obtengamos de (γ, δ) ? La respuesta es: depende.

Una vez estimados (γ, δ) , tenemos un sistema de 6 ecuaciones y 6 incógnitas (α, β) [ver (*)]. Siguiendo la notación del profesor, a partir de este sistema tenemos una función $H : \mathbb{R}^6 \rightarrow \mathbb{R}^6$ tal que $(\gamma, \delta) = H(\alpha, \beta)$. ¿Existe H^{-1} tal que $(\alpha, \beta) = H^{-1}(\gamma, \delta)$?

- Si la respuesta es NO, el modelo estructural está subidentificado. En este caso no es posible estimar los parámetros de interés (α, β) .
- Si la respuesta es SÍ y $\forall(\gamma, \delta) \exists! (\alpha, \beta)$, el modelo estructural está exactamente identificado. En este caso sí es posible estimar los parámetros de interés (α, β) .
- Si la respuesta es SÍ y $\forall(\gamma, \delta)$ existen varios posibles valores de (α, β) , el modelo estructural está sobreidentificado. En este caso no es posible estimar precisamente los parámetros de interés (α, β) .

3. Variables instrumentales: computacional

En el archivo Auxiliar 09 - BD Variables Instrumentales.dta contenido en Material Docente se encuentra una base de datos que contiene información sobre aspectos económicos y de salud de distintos ciudadanos estadounidenses. En una revista de salud, de dudosa procedencia, se estima mediante MCO el siguiente modelo de que busca explicar el gasto médico de los ciudadanos estadounidenses:

$$\ln(gastomed)_i = \beta_0 + \beta_1 \text{segurosalud}_i + \beta_2 \text{enfermedad}_i + \beta_3 \text{edad}_i + \beta_4 \ln(\text{ingreso})_i + u_i$$

1. Usted le comenta esto a una amiga, quien le dice “hay algo que me hace ruido: una persona puede contratar un seguro de salud porque sus gastos médicos son altos, así como puede tener altos gastos médicos gracias a que contrata un seguro de salud”. ¿A qué posible fuente de endogeneidad se puede estar refiriendo su amiga?

Respuesta: La amiga en cuestión se está refiriendo a un problema de simultaneidad, que –como vimos en la pregunta 1– genera endogeneidad en el modelo.

2. Realice la estimación de Mínimos Cuadrados en 2 Etapas en *Stata* realizando las regresiones de manera secuencial ocupando como instrumento la variable ssi^3 como instrumento para la variable endógena $seguro_salud$.

Respuesta: Ver archivo Auxiliar 09 P3 Código (MC2E).do.

3. Realice la estimación de Mínimos Cuadrados en 2 Etapas en *Stata* mediante el comando *ivregress* y compare sus resultados con los de la parte anterior.

Respuesta: Ver archivo Auxiliar 09 P3 Código (MC2E).do.

4. Obtenga el estimador $\hat{\beta}_{VI} = (\bar{Z}^T X)^{-1} \bar{Z}^T Y$ en *Matlab*.

Respuesta: Ver archivo Auxiliar 09 P3 Código (VI).m.

³Es el Seguro Suplementario al Ingreso, un programa del gobierno de EEUU que proporcion pagos a personas de bajos ingresos que tienen 65 años o más, que sean ciegos o discapacitados. Esta variable contiene el cuociente entre dicho aporte y el ingreso de la persona.