ECONOMETRIA

Tema 6: MODELOS CON VARIABLES EXPLICATIVAS ENDÓGENAS

César Alonso

Universidad Carlos III de Madrid

Endogeneidad

Dado el modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- Si se cumple que

$$E(\varepsilon | X_1, X_2, \ldots, X_K) = 0, \ \forall X_1, X_2, \ldots, X_K,$$

  decimos que tenemos **variables explicativas exógenas**.
- **Si** por alguna razón (omisión de variables relevantes, errores de medida, simultaneidad, etc.) X_j **está correlacionada con** ε, decimos que X_j es una **variable explicativa endógena**.

Endogeneidad

- La existencia de variables explicativas endógenas invalida los estimadores MCO de los parámetros del modelo, que serán **inconsistentes**.
- En este tema vamos a estudiar cómo obtener estimadores consistentes de los parámetros del modelo en presencia de variables explicativas endógenas,
  - utilizando **variables instrumentales** y
  - aplicando el método de **mínimos cuadrados bietápicos** o **mínimos cuadrados en dos etapas** (**MC2E**).

Endogeneidad

Ejemplo 1: Error de medida en variables explicativas

- Recordemos en el modelo de regresión simple $Y = \beta_0 + \beta_1 X^* + \varepsilon$ donde se cumplen los supuestos clásicos y por tanto:

$$E(\varepsilon|X^*) = 0 \Rightarrow E(Y|X^*) = L(Y|X^*) = \beta_0 + \beta_1 X^*,$$

de manera que β_0 y β_1 verifican:

$$E(\varepsilon) = 0, \qquad C(X^*, \varepsilon) = 0 \Rightarrow$$

$$\beta_0 = E(Y) - \beta_1 E(X^*) \qquad \beta_1 = C(X^*, Y)/V(X^*).$$

- Sin embargo, X^* se mide con error, de modo que observamos $X = X^* + v_1$, siendo v_1 el error de medida.
- Sustituyendo $X^* = X - v_1$, tenemos:

$$Y = \beta_0 + \beta_1 X + \underbrace{(\varepsilon - \beta_1 v1)}_{u}$$

donde $C(u, X) \neq 0 \Rightarrow X$ es endógena.

Ejemplo 2: Omisión de variables explicativas

- Recordemos el caso de omisión de variables relevantes.
- Sea el modelo $Y = \gamma_0 + \gamma_1 X_1 + u$, donde $u = \varepsilon + \beta_2 X_2$ con $\beta_2 \neq 0 \Rightarrow$ se ha omitido X_2.
- En general, $C(X_1, u) \neq 0 \Rightarrow X_1$ es endógena.

Endogeneidad

Ilustraciones:

1. *Capacidad no observada en una ecuación de salarios.*
   Consideremos la siguiente ecuación de salarios:

$$\log(salario) = \beta_0 + \beta_1 educ + \beta_2\ cap + e.$$

   Como la capacidad *cap* es no observable, nos quedaríamos con el siguiente modelo de regresión simple:

$$\log(salario) = \beta_0 + \beta_1 educ + u,$$

   donde el término de error u contiene *cap*. Si estimamos por MCO, obtendremos un estimador sesgado e inconsistente de β_1 si $C\left(educ, cap\right) \neq 0$.
2. *Efecto del tabaco sobre los salarios (ignorando el nivel de educación).*
3. *Efecto del tabaco sobre el cáncer (ignorando el estado físico).*

Endogeneidad

Ejemplo 3: Simultaneidad

- Es bastante habitual que las realizaciones de distintas variables económicas estén relacionadas entre sí.
- Esto supone que la ecuación de la variable dependiente en que estamos interesados forma parte de un sistema de **ecuaciones simultáneas**:
  - algunas variables que aparecen en el lado derecho de la ecuación de interés aparecen como variables dependientes en otras ecuaciones, y viceversa.

Endogeneidad

Ejemplo 3a: Modelo de equilibrio de mercado

- Consideremos el siguiente sistema:

$$\begin{aligned} Y_1 = &\quad \alpha_1 Y_2 \quad + \alpha_2 X_1 \qquad\qquad\qquad\qquad + u_1 \quad \text{(Demanda)} \\ Y_2 = &\quad \alpha_3 Y_1 \qquad\qquad + \alpha_4 X_2 \quad + \alpha_5 X_3 \quad + u_2 \quad \text{(Oferta)} \end{aligned}$$

- Las variables endógenas $Y_1 =$ cantidad, $Y_2 =$ precio, se determinan por medio de
  - las variables exógenas $X_1 =$ renta, $X_2 =$ salario, $X_3 =$ tipo de interés,
  - y por las perturbaciones $u_1 =$ shock de demanda, $u_2 =$ shock de oferta.
- Las variables Y_1 e Y_2, que aparecen en el lado derecho de las respectivas ecuaciones de oferta y demanda, no son ortogonales a sus respectivas perturbaciones:

$$E(Y_1 | Y_2, X_1) = \alpha_1 Y_2 + \alpha_2 X_1 + \underbrace{E(u_1 | Y_2, X_1)}_{\neq 0}$$

Endogeneidad

Ejemplo 3b: Función de producción

- Si la empresa es maximizadora de beneficios o minimizadora de costes,
  - las cantidades de inputs se determinan simultáneamente con el nivel de producción,
  - la perturbación, que refleja el efecto de shocks tecnológicos, está en general correlacionada con las cantidades de inputs.

Variables instrumentales

- El método de **Variables Instrumentales** (**VI**) permite obtener estimadores consistentes de los parámetros en situaciones en que el estimador MCO es inconsistente (omisión de variables relevantes, errores de medida o simultaneidad).

Variables instrumentales

- En general, tenemos que dado el modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

donde $C(X, \varepsilon) \neq 0 \Rightarrow$ Los parámetros de interés β_0 y β_1 NO coinciden con los parámetros de la proyección lineal $L(Y|X) \Rightarrow$ los estimadores MCO ($\widehat{\beta}_0$ y $\widehat{\beta}_1$) de la proyección lineal de Y sobre X son estimadores inconsistentes de β_0 y de β_1:

$$p\lim \widehat{\beta}_1 = \frac{p\lim \left(\frac{1}{n}\sum_i x_i y_i\right)}{p\lim \left(\frac{1}{n}\sum_i x_i^2\right)} = \frac{p\lim \left[\frac{1}{n}\sum_i x_i(\beta_1 x_i + \varepsilon_i)\right]}{p\lim \left(\frac{1}{n}\sum_i x_i^2\right)}$$

$$= \beta_1 + \frac{p\lim \left(\frac{1}{n}\sum_i x_i \varepsilon_i\right)}{p\lim \left(\frac{1}{n}\sum_i x_i^2\right)} = \beta_1 + \frac{C(X, \varepsilon)}{V(X)} \neq \beta_1$$

con: $y_i = Y_i - \overline{Y}$, $x_i = X_i - \overline{X}$.

- En el modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

donde $C(X, \varepsilon) \neq 0$,

necesitamos **información adicional** (en forma de variables adicionales) para obtener estimaciones consistentes de β_0 y de β_1.

- Supongamos que disponemos de una variable Z (denominada Variable Instrumental) que:
  - no esté correlacionada con el error del modelo: **(a)** $C(Z, \varepsilon) = 0$;
  - esté correlacionada con la variable endógena X: **(b)** $C(Z, X) \neq 0$

- Empleando Z como instrumento, podremos obtener estimadores consistentes de β_0 y de β_1.
- A partir de (2) podemos escribir:

$$C(Z,Y) = \beta_1 C(Z,X) + C(Z,\varepsilon)$$

lo que, dado **(a)**, implica que en la población se verifica que:

$$\beta_1 = \frac{C(Z,Y)}{C(Z,X)}$$

$$\beta_0 = E(Y) - \beta_1 E(X) = E(Y) - \frac{C(Z,Y)}{C(Z,X)} E(X)$$

- Suponiendo que disponemos de una muestra aleatoria de la población de tamaño n, y sustituyendo momentos poblacionales por muestrales (*principio de analogía*) en las expresiones anteriores, se obtiene el Estimador de Variables Instrumentales (VI):

$$\begin{aligned} \widetilde{\beta}_1 &= \frac{S_{YZ}}{S_{XZ}} = \frac{\sum_i z_i y_i}{\sum_i z_i x_i} \\ \widetilde{\beta}_0 &= \overline{Y} - \widetilde{\beta}_1 \overline{X} \end{aligned}$$

con: $y_i = Y_i - \overline{Y}$, $x_i = X_i - \overline{X}$, $z_i = Z_i - \overline{Z}$.

- Siempre que se cumplan **(a)** y **(b)**, el estimador de VI será un estimador consistente:

$$\begin{aligned} p\lim \widetilde{\beta}_1 &= \frac{p\lim\left(\frac{1}{n}\sum_i z_i y_i\right)}{p\lim\left(\frac{1}{n}\sum_i z_i x_i\right)} = \frac{p\lim\left[\frac{1}{n}\sum_i z_i(\beta_1 x_i + \varepsilon_i)\right]}{p\lim\left(\frac{1}{n}\sum_i z_i x_i\right)} \\ &= \beta_1 + \frac{p\lim\left(\frac{1}{n}\sum_i z_i \varepsilon_i\right)}{p\lim\left(\frac{1}{n}\sum_i z_i x_i\right)} = \beta_1 + \frac{C(Z,\varepsilon)}{C(Z,X)} = \beta_1 \end{aligned}$$

- Toda variable instrumental o instrumento debe cumplir las dos propiedades, **(a)** y **(b)**. A este respecto:
  - La condición **(a)** de que $C(Z, \varepsilon) = 0$, no puede contrastarse. Debemos suponer que es así mediante argumentos basados en el comportamiento económico o en alguna conjetura.
    $\Rightarrow$ Hay que ser muy cuidadoso en la elección de Z.

- La condición **(b)** de que $C(Z, X) \neq 0$ sí puede contrastarse en la muestra.
  Lo más sencillo es considerar la proyección lineal de X sobre Z:

$$X = \pi_0 + \pi_1 Z + v,$$

  estimarla por MCO y contrastar:

$$H_0 : \pi_1 = 0 \text{ frente a } H_1 : \pi_1 \neq 0$$

- **Nota:** Si $Z = X$, obtenemos la estimación de MCO.
  - Es decir, cuando X es exógena, puede utilizarse como su propio instrumento, y el estimador de VI es entonces idéntico al estimador MCO.

- En general, el estimador de VI tendrá una varianza mayor que el de MCO.
  - Para verlo, nótese que la varianza estimada del estimador de VI de $\widetilde{\beta}_1$, $s^2_{\widetilde{\beta}_1}$, puede escribirse como:

$$s^2_{\widetilde{\beta}_1} \equiv \widehat{V}(\widetilde{\beta}_1) = \frac{\widetilde{\sigma}^2 S_z^2}{n S_{ZX}^2} = \frac{\widetilde{\sigma}^2}{n\ r_{ZX}^2\ S_X^2},$$

donde

$$r_{ZX} = \frac{S_{ZX}}{S_Z S_X}$$

es el coeficiente de correlación muestral entre Z y X (que mide el grado de relación lineal entre X y Z en la muestra).

- Recordemos que la varianza estimada del estimador MCO de β_1, $\widehat{\beta}_1$, es

$$s^2_{\widehat{\beta}_1} = \frac{\widehat{\sigma}^2}{nS_X^2},$$

donde:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\varepsilon}_i^2,$$

siendo $\widehat{\varepsilon}_i$ el residuo de la estimación de MCO.

- Si X es en realidad exógena, los estimadores MCO son consistentes, y en tal caso

$$p\lim \widehat{\sigma}^2 = p\lim \widetilde{\sigma}^2 = \sigma^2.$$

Como $0 < |r_{ZX}| < 1$, esto implica que: $s^2_{\widetilde{\beta}_1} > s^2_{\widehat{\beta}_1}$ (y la diferencia será tanto mayor cuanto menor sea r_{ZX} en valor absoluto).

- Por tanto, **si X es exógena**, realizar la estimación por VI en vez de por MCO tiene un coste en términos de eficiencia.
  Cuanto menor sea la correlación entre Z y X, mayor será la varianza de VI respecto a la de MCO.
  - En ese caso, tanto $\widehat{\beta}_1$ como $\widetilde{\beta}_1$ son consistentes, y asintóticamente la varianza del estimador de VI relativa al de MCO depende inversamente de ρ_{ZX},
    $$p\lim\left(s^2_{\widetilde{\beta}_1} \Big/ s^2_{\widehat{\beta}_1}\right) = 1/\rho^2_{ZX}$$
  - Es decir, en el límite,
  - Si $\rho_{ZX} = 1\% = 0.01$, $V(\widetilde{\beta}_1) \simeq 10000 V\left(\widehat{\beta}_1\right)$, y por tanto $s_{\widetilde{\beta}_1} \simeq 100 s_{\widehat{\beta}_1}$.
  - Si $\rho_{ZX} = 10\% = 0.01$, $V(\widetilde{\beta}_1) \simeq 100 V\left(\widehat{\beta}_1\right)$, y por tanto $s_{\widetilde{\beta}_1} \simeq 10 s_{\widehat{\beta}_1}$.
  - Incluso con una correlación relativamente alta, $\rho_{ZX} = 50\% = 0.5$, $V(\widetilde{\beta}_1) \simeq 4 V\left(\widehat{\beta}_1\right)$, y por tanto $s_{\widetilde{\beta}_1} \simeq 2 s_{\widehat{\beta}_1}$.

- PERO **si X es endógena**, la comparación entre el estimador MCO y el de VI en términos de eficiencia NO tiene sentido, porque el estimador MCO es inconsistente.

- Consideremos el modelo simple

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- Suponiendo **homocedasticidad** condicional:

$$V(\varepsilon|Z) = \sigma^2 = V(\varepsilon),$$

se puede demostrar que

$$\frac{\widetilde{\beta}_1 - \beta_1}{s_{\widetilde{\beta}_1}} \stackrel{\cdot}{\sim} N(0,1)$$

- donde $s_{\widetilde{\beta}_1}$ es el error estándar del estimador de variables instrumentales:

$$s^2_{\widetilde{\beta}_1} \equiv \widehat{V}(\widetilde{\beta}_1) = \frac{\widetilde{\sigma}^2 S_z^2}{n S_{ZX}^2}$$

$$\Rightarrow \quad s_{\widetilde{\beta}_1} = \frac{\widetilde{\sigma} S_z}{\sqrt{n} S_{ZX}}$$

- y donde:

$$\widetilde{\sigma}^2 = \frac{1}{n} \sum_i \widetilde{\varepsilon}_i^2,$$

siendo $\widetilde{\varepsilon}_i$ el residuo de la estimación de VI:

$$\widetilde{\varepsilon}_i = Y_i - \left(\widetilde{\beta}_0 + \widetilde{\beta}_1 X_i\right) = y_i - \widetilde{\beta}_1 x_i$$

- Esto permite construir intervalos de confianza y realizar contrastes de hipótesis.

- La mayor parte de los programas econométricos calculan el R^2 con la estimación de VI mediante la fórmula convencional:

$$R^2 = 1 - \frac{\sum_i \widetilde{\varepsilon}_i^2}{\sum_i y_i^2},$$

  donde $\widetilde{\varepsilon}_i$ son los residuos de VI.
- Sin embargo, cuando X y ε están correlacionadas esta fórmula del R^2 no es correcta. El R^2 de la estimación de VI:
  - puede ser negativo si $\sum_i \widetilde{\varepsilon}_i^2 > \sum_i y_i^2$.
  - no tiene una interpretación natural, porque si $C(X, \varepsilon) \neq 0$, no podemos descomponer la varianza de Y como $\beta_1^2 V(X) + V(\varepsilon)$.
  - no puede utilizarse para construir el estadístico de contraste W^0 (hay que usar la Suma de cuadrados de los residuos).

VI vs MCO

- Si nuestro objetivo fuese maximizar el R^2, siempre utilizaríamos MCO.
- Pero si nuestro objetivo es estimar apropiadamente el efecto causal de X sobre Y, β_1:
  - Si $C(X, \varepsilon) = 0$, deberíamos utilizar MCO
    (será más eficiente que cualquier estimador de VI que utilice $Z \neq X$).
  - Si $C(X, \varepsilon) \neq 0$, MCO no será consistente,
    mientras que sí lo será un estimador de VI con $Z \neq X$ apropiado
    (La bondad del ajuste, en este contexto, no es el aspecto de interés).

- El estimador de VI es consistente si **(a)** $C(Z, \varepsilon) = 0$, **(b)** $C(Z, X) \neq 0$.
- Si no se cumplen estas condiciones, el estimador de VI puede tener un sesgo asintótico mayor que el de MCO, especialmente si $|\rho_{XZ}|$ es pequeña..
- Podemos ver esto comparando los $p\lim$ de ambos estimadores, VI (cuando existe la posibilidad de que Z y ε estén correlacionadas) y MCO (cuando X es endógena).

$$p\lim \widetilde{\beta}_1 = \beta_1 + \frac{C(Z, \varepsilon)}{C(Z, X)}$$

$$p\lim \widehat{\beta}_1 = \beta_1 + \frac{C(X, \varepsilon)}{V(X)}$$

- Expresado en términos de las correlaciones y desviaciones estándar poblacionales de ε y X respectivamente:

$$p\lim \widetilde{\beta}_1 = \beta_1 + \frac{\rho_{Z\varepsilon}}{\rho_{ZX}} \frac{\sigma_\varepsilon}{\sigma_X}$$

$$p\lim \widehat{\beta}_1 = \beta_1 + \rho_{X\varepsilon} \frac{\sigma_\varepsilon}{\sigma_X}$$

- Por tanto, preferiríamos el estimador de VI al MCO si

$$\frac{\rho_{Z\varepsilon}}{\rho_{ZX}} < \rho_{X\varepsilon}.$$

- Cuando Z y X no están correlacionadas en absoluto, la situación es especialmente mala, esté o no Z correlacionada con ε.
- Cuando Z y X presentan una correlación muestral r_{ZX} muy pequeña, el problema será muy parecido:
  - Puede estar reflejando que $C(Z, X) = 0$.
  - Las estimaciones serán muy imprecisas, pudiendo presentar valores implausibles.

Ejemplo: Efecto del consumo de tabaco sobre el peso del niño al nacer

- El siguiente ejemplo ilustra por qué siempre deberíamos comprobar si la variable explicativa endógena está correlacionada con el instrumento potencial.
- Al estimar el efecto de varias variables, entre ellas el consumo de tabaco por parte de la madre, en el peso de los recién nacidos, se han obtenido los siguientes resultados:

Model 1: OLS, using observations 1-1388

Dependent Variable: LBWGHT

Variable	Coefficient	Std. Error	t-Statistic	p-value
PACKS	−0.0837	0.0175	−4.80	0.000
MALE	0.0262	0.0100	2.62	0.009
PARITY	0.0147	0.0054	2.72	0.007
LFAMINC	0.0180	0.0053	3.40	0.001
const	4.6756	0.0205	228.53	
R^2	0.0350			

- Donde
  - `LBWGHT` = logaritmo del peso del bebé al nacer,
  - `MALE` = variable binaria que vale 1 si el bebé es varón y 0 en otro caso,
  - `PARITY` = orden de nacimiento (entre sus hermanos) del bebé,
  - `LFAMINC` = logaritmo de la renta familiar en miles de dólares,
  - `PACKS` = número medio de cajetillas diarias fumadas por la madre durante el embarazo.

- Puede que PACKS esté correlacionado con otros hábitos de salud y/o con un buen cuidado prenatal $\Rightarrow$ PACKS y el término de error podrían estar correlacionados.
- Posible variable instrumental para PACKS: precio medio de los cigarrillos en el Estado de residencia (CIGPRICE).
  - Supondremos que CIGPRICE no está correlacionado con el término de error (aunque las ayudas estatales a la salud podrían estar correlacionadas con los impuestos al tabaco).
  - La teoría económica sugiere que $C(\texttt{PACKS},\texttt{CIGPRICE}) < 0$.

- Proyección lineal de `PACKS` sobre `CIGPRICE` y resto de las variables exógenas (forma reducida):

Model 1: OLS, using observations 1-1388

Dependent Variable: `PACKS`

Variable	Coefficient	Std. Error	t-Statistic	p-value
`CIGPRICE`	0.0008	0.0008	1.00	0.317
`MALE`	−0.0047	0.0158	−0.30	0.766
`PARITY`	0.018	0.0089	2.04	0.041
`LFAMINC`	−0.0526	0.0087	−6.05	0.000
const	0.1374	0.1040	1.32	0.187
R^2	0.0305			

- Los resultados de la forma reducida indican que no hay relación entre el consumo de cigarrillos durante el embarazo y el precio de los cigarrillos (es decir, que la elasticidad precio del consumo de tabaco, que es un bien adictivo, no es estadísticamente distinta de cero).
- Dado que `PACKS` y `CIGPRICE` no están correlacionadas, `CIGPRICE` no cumple la condición **(b)** $\Rightarrow$ no deberíamos utilizar `CIGPRICE` como instrumento.
- Pero, ¿qué sucede si utilizamos `CIGPRICE` como instrumento? Los resultados de la estimación VI son:

Model 1: TSLS, using observations 1-1388
Dependent Variable: LBWGHT
Instrumented: PACKS
Instruments: CIGPRICE

Variable	Coefficient	Std. Error	t-Statistic	P-VALUE
PACKS	0.7971	1.1132	0.72	0.474
MALE	0.0298	0.0172	1.73	0.084
PARITY	−0.0012	0.0254	−0.05	0.961
LFAMINC	0.0636	0.0571	1.12	0.265
C	4.4679	0.2563	17.43	0.000
R^2		.	F-statistic	2.50

Variables instrumentales

Instrumentos no adecuados: Ejemplo

- El coeficiente de `PACKS` es muy grande y tiene un signo opuesto al esperado. El error estándar es también muy grande.
- Pero las estimaciones carecen de sentido, ya que `CIGPRICE` no cumple uno de los requisitos para ser un instrumento válido.

Generalización: el estimador de MC2E

Modelo simple

- Sea el modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{donde } C(X, \varepsilon) \neq 0,$$

- Disponemos de dos posibles Variables Instrumentales Z_1 y Z_2 que cumplen:

$$C(Z_1, \varepsilon) = 0, \quad C(Z_2, \varepsilon) = 0,$$
$$C(Z_1, X) \neq 0, \quad C(Z_2, X) \neq 0.$$

- Podríamos obtener dos estimadores simples de VI, uno con Z_1 y otro con Z_2, que serán numéricamente distintos.
- PERO también podemos obtener un estimador de VI que use como instrumento una combinación lineal de Z_1 y Z_2:
  - Obtendríamos el estimador de **Mínimos Cuadrados en 2 Etapas** (MC2E)

- **1^a Etapa**: Se estima por MCO la proyección lineal de la variable endógena X sobre los instrumentos Z_1 y Z_2 (conocida como **forma reducida**):

$$X = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + v. \tag{3}$$

  - Sean $\widehat{\pi}_0$, $\widehat{\pi}_1$, $\widehat{\pi}_2$ los estimadores MCO de dicha forma reducida.
  - Los valores ajustados de X a partir de dichas estimaciones son:

$$\widehat{X} = \widehat{\pi}_0 + \widehat{\pi}_1 Z_1 + \widehat{\pi}_2 Z_2.$$

- **2^a Etapa**: Se estima por MCO la regresión de Y sobre $\widehat{X}$ (de ahí el nombre de MC2E):

$$Y = \beta_0 + \beta_1 \widehat{X} + u$$

  - Dicho estimador equivale a estimar β_0 y β_1 por VI usando $Z = \widehat{X}$

- Aunque en ambos casos los coeficientes son los mismos, los errores estándar al obtener MC2E secuencialmente son incorrectos.
  - La razón es que el término de error de la segunda etapa, u, incluye v, pero los errores estándar comprenden la varianza de ε solamente.
- La mayoría de los paquetes econométricos tienen instrucciones especiales para llevar a cabo MC2E, por lo que no es preciso realizar las dos etapas secuencialmente.

- La forma reducida (3) descompone de forma aditiva la variable explicativa endógena en dos partes:
  - La parte exógena de X, explicada linealmente por los instrumentos, $\pi_0 + \pi_1 Z_1 + \pi_2 Z_2$.
  - La parte endógena de X, que es lo que queda sin explicar por los instrumentos, es decir, el error de la forma reducida v.

- Si los instrumentos son válidos y $V(\varepsilon|\, Z_1, Z_2)$ es homocedástica, se demuestra que los estimadores de MC2E son **consistentes** y **asintóticamente normales.**
- Por tanto, puede hacerse inferencia usando como estimador de la varianza poblacional

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_i \widetilde{u}_i^2,$$

  donde $\widetilde{u}_i^2$ son los residuos basados en la estimación MC2E.
- Al igual que ocurre con el estimador simple de VI, cuando los instrumentos no son apropiados (porque están correlacionados con el término de error o poco correlacionados con la variable endógena) los estimadores de MC2E pueden ser peores que los de MCO.

- Consideremos para simplificar el modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

donde:

$$E(\varepsilon) = 0, \quad C(X_1, \varepsilon) = 0, \quad C(X_2, \varepsilon) \neq 0.$$

- Es decir: X_1 es una variable exógena pero X_2 es una variable endógena.

- Supongamos que disponemos de una variable instrumental Z tal que

$$C(Z, \varepsilon) = 0.$$

- La forma reducida será:

$$X_2 = \pi_0 + \pi_1 X_1 + \pi_2 Z + v.$$

  Para que Z sea un instrumento válido será necesario que $\pi_2 \neq 0$ (es decir, que $C(Z, X_2) \neq 0$).

- **Muy importante**: Nótese que la forma reducida para la variable explicativa endógena incluye los instrumentos y **todas** las variables explicativas exógenas del modelo.

- **¿Qué pasa si tenemos más de una variable endógena?**
  Supongamos que

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

donde X_1 y X_2 son endógenas, mientras que X_3 es exógena.

$$E(\varepsilon) = 0, \quad C(X_1, \varepsilon) \neq 0, \quad C(X_2, \varepsilon) \neq 0, \quad C(X_3, \varepsilon) = 0.$$

  - En ese caso, necesitaremos, al menos tantas variables exógenas adicionales como variables explicativas endógenas haya para poder utilizar como instrumentos.

Generalización: el estimador de MC2E

Modelo múltiple con varias variables explicativas endógenas

- En este caso, sean Z_1 y Z_2 tales que $C(Z_1, \varepsilon) = C(Z_2, \varepsilon) = 0$.
- Tendremos una ecuación de forma reducida para cada variable explicativa endógena,
  donde aparecerán todas las variables explicativas exógenas y todos los instrumentos:

$$\begin{aligned} X_1 &= \pi_{10} + \pi_{11} X_3 + \delta_{11} Z_1 + \delta_{12} Z_2 + v_1, \\ X_2 &= \pi_{20} + \pi_{21} X_3 + \delta_{21} Z_1 + \delta_{22} Z_2 + v_2, \end{aligned}$$

  donde debe cumplirse al menos que $\delta_{11} \neq 0$ y $\delta_{22} \neq 0$ o que $\delta_{12} \neq 0$ y $\delta_{21} \neq 0$.

- En general, todos los instrumentos estarán presentes en las ecuaciones de forma reducida de cada uno de las variables explicativas endógenas.

Contraste de endogeneidad (contraste de Hausman)

- En la práctica, existen muchas situaciones en las que no sabemos si una variable explicativa es o no endógena. Por ello se han propuesto contrastes de endogeneidad.
- En el contexto del modelo

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

podemos considerar las hipótesis alternativas:

$$\begin{aligned} H_0 &: C(X, \varepsilon) = 0 \text{ (exogeneidad)} \\ H_1 &: C(X, \varepsilon) \neq 0 \text{ (endogeneidad)} \end{aligned}$$

- ¿Cómo puedo realizar el contraste de la hipótesis nula de exogeneidad?
- Supongamos que disponemos de un instrumento válido Z (de manera que $C(Z, \varepsilon) = 0$ y $C(Z, X) \neq 0$)

Contraste de endogeneidad (contraste de Hausman)

- Entonces, a partir de la forma reducida

$$X = \pi_0 + \pi_1 Z + v,$$

es fácil obtener que

$$\begin{aligned} C(X, \varepsilon) &= C(\pi_0 + \pi_1 Z + v, \varepsilon) = C(v, \varepsilon) \Rightarrow \\ C(X, \varepsilon) &= 0 \Leftrightarrow C(v, \varepsilon) = 0 \end{aligned}$$

Por tanto, si $H_0 : C(X, \varepsilon) = 0$ es cierta, el coeficiente α en la regresión:

$$\varepsilon = \alpha v + \xi$$

verifica que $\alpha = 0$, o de manera equivalente el coeficiente α en la regresión,

$$Y = \beta_0 + \beta_1 X + \alpha v + \xi \tag{4}$$

verifica que $\alpha = 0$.

Contraste de endogeneidad (contraste de Hausman)

- Por tanto, si pudiera estimar (4) podría contrastar $H_0 : \alpha = 0$, que es equivalente a $H_0 : C(X, \varepsilon) = 0$.
- En la práctica, como v no es observable, se sustituye por el residuo de MCO $\widehat{v}$ de la forma reducida, lo que no tiene consecuencias.
- Por tanto, el modelo

$$Y = \beta_0 + \beta_1 X + \alpha \widehat{v} + \xi' \tag{5}$$

con $\widehat{v} = X - (\widehat{\pi}_0 + \widehat{\pi}_1 Z)$ (residuo MCO de la forma reducida), se estima por MCO.

- La hipótesis nula es que X es exógena, es decir: $H_0 : \alpha = 0$.
- Por tanto, si rechazamos que α es cero en el modelo (5), concluiremos que X es endógena.

Contraste de endogeneidad (contraste de Hausman)

- **Generalización**:
  El contraste de Hausman para el caso de r variables potencialmente endógenas consistiría en:
  - estimar las r formas reducidas correspondientes para cada una de dichas variables,
  - obtener los residuos de cada forma reducida,
  - incluir como r regresores adicionales cada uno de estos residuos en el modelo de interés,
  - y contrastar la significación conjunta de dichos residuos mediante el estadístico W^0:

$$W^0 = n \times \frac{SRR - SRS}{SRS} \tilde{\cdot} \chi^2_r$$

Contraste de endogeneidad (contraste de Hausman)

- donde
  - *SRR* es la suma de los cuadrados de los residuos del modelo original sin los residuos de las formas reducidas,
  - *SRS* es la suma de los cuadrados de los residuos del modelo ampliado que incluye los residuos de cada una de las formas reducidas como regresores adicionales con dichos residuos
  - r es el número de variables potencialmente endógenas.
- Si se concluye que los residuos de las formas reducidas son conjuntamente significativos,
  ello indica que al menos una de las variables explicativas potencialmente endógenas lo es en realidad.

Contraste de endogeneidad (contraste de Hausman)

- *Ejemplo*
  Como ilustración, supongamos que tenemos el modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  donde X_1 y X_2 son potencialmente endógenas, y X_3 es exógena.

  - Necesitaremos, al menos, dos instrumentos Z_1 y Z_2 tales que $C(Z_1, \varepsilon) = C(Z_2, \varepsilon) = 0$.
  - Tendremos dos ecuaciones de forma reducida:

$$\begin{aligned} X_1 &= \pi_{10} + \pi_{11} X_3 + \delta_{11} Z_1 + \delta_{12} Z_2 + v_1, \\ X_2 &= \pi_{20} + \pi_{21} X_3 + \delta_{21} Z_1 + \delta_{22} Z_2 + v_2. \end{aligned}$$

Contraste de endogeneidad (contraste de Hausman)

- La hipótesis nula de exogeneidad es ahora
  $H_0 : C(X_1, \varepsilon) = 0,\ C(X_2, \varepsilon) = 0.$
- De forma equivalente, podemos usar la regresión ampliada

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_1 \widehat{v}_1 + \alpha_2 \widehat{v}_2 + \xi',$$

  donde $\widehat{v}_1$, $\widehat{v}_2$ son los residuos de las formas reducidas de X_1, X_2, respectivamente,
  la hipótesis nula se puede escribir como $H_0 : \alpha_1 = \alpha_2 = 0$.
- Para contrastar dicha hipótesis (que se compone de dos restricciones), deberíamos estimar dicha regresión ampliada y calcular su suma de cuadrados de los residuos SRS así como el modelo bajo H_0

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

  y calcular la suma de cuadrados de los residuos SRR para construir el contraste W^0, cuya distribución aproximada es una χ^2_2.

Contraste de restricciones de sobreidentificación (contraste de Sargan)

- Si tenemos solamente una variable instrumental para cada variable explicativa endógena, decimos que el modelo está "exactamente identificado".
  - en ese caso, no podemos contrastar la condición de no correlación de los instrumentos con el error.
- PERO si tenemos más variables instrumentales que variables explicativas potencialmente endógenas, decimos que el modelo está "sobreidentificado".
  - en ese caso, podemos contrastar si alguna de ellas no está correlacionada con el término de error.
- Supongamos que tenemos r variables explicativas potencialmente endógenas y q instrumentos, donde $q > r$
  - $(q - r)$ es el número de restricciones de sobreidentificación (número de instrumentos "extra").

Contraste de restricciones de sobreidentificación (contraste de Sargan)

- No observamos los errores de la ecuación de interés u
- Pero podemos implementar un contraste basado en los residuos MC2E, $\widetilde{u}$ (análogos muestrales de u).
- Procedimiento del contraste:
  - Estimar la ecuación de interés por MC2E y obtener los residuos MC2E, $\widetilde{u}$.
  - Regresar $\widetilde{u}$ sobre todas las variable exógenas del modelo y sobre todos los instrumentos. Obtener el R^2 de dicha regresión, $R^2_{\widetilde{u}}$.
  - Bajo la hipótesis nula de que ninguna de las VI está correlacionada con $\widetilde{u}$, tenemos que

$$nR^2_{\widetilde{u}} \overset{\sim}{\cdot} \chi^2_{q-r},$$

Contraste de restricciones de sobreidentificación (contraste de Sargan)

- Intuición del contraste: los valores ajustados de esta regresión auxiliar, $\widehat{\widetilde{u}}_i$, tienen media cero y varianza $\sigma^2_{\widetilde{u}}$. Bajo homocedasticidad condicional, asintóticamente,
  $_i \left(\widehat{\widetilde{u}}_i^2 \Big/ \sigma^2_{\widetilde{u}} \right) \equiv$ Suma de $N(0,1)$ al cuadrado, de las cuales solamente $(q-r)$ son independientes.
  Por tanto, dicha expresión se distribuye asintóticamente como una χ^2_{q-r}.
- En la práctica, sustituiremos $\sigma^2_{\widetilde{u}}$ por un estimador consistente $s^2_{\widetilde{u}} = \dfrac{1}{n} \, _i \widetilde{u}_i^2$,
- Por tanto, nuestro estadístico será

$$_i \frac{\widehat{\widetilde{u}}_i^2}{\frac{1}{n} \, _i \widetilde{u}_i^2} \equiv n \frac{_i \widehat{\widetilde{u}}_i^2}{_i \widetilde{u}_i^2} = n \times R^2_{\widetilde{u}}.$$

Contraste de restricciones de sobreidentificación (contraste de Sargan)

- Si $nR^2_{\widetilde{u}}$ excede el valor crítico de la distribución χ^2_{q-r} al nivel de significación prefijado, rechazaremos la hipótesis nula a dicho nivel de significación y concluiremos que al menos alguna de las VI no es exógena.
- Otra cosa es que este contraste no establece qué variable es la responsable de rechazar la hipótesis nula de no correlación. (No obstante, en la medida en que $q - r$ sea grande, podríamos aplicar el proceso secuencialmente para averiguar qué instrumentos son responsables del rechazo).
- Este contraste también se conoce como contraste de Hansen-Sargan.

Ejemplo: ecuación de salarios

- Sea la ecuación:

$$\ln(salario) = \beta_0 + \beta_1 educ + \beta_2 cap + \varepsilon$$

  donde $\beta_2 \neq 0$ (es decir, la variable *cap*, capacidad, que es inobservable, es una variable relevante).

- Si estimamos por MCO:

$$\ln(salario) = \gamma_0 + \gamma_1 educ + u$$

  con $u = \beta_2 cap + \varepsilon$
  $\Rightarrow \widehat{\gamma}_1$ será un estimador inconsistente de β_1.

- Si disponemos de una variable instrumental para *educ* podremos estimar por VI.

Ejemplo: ecuación de salarios

- ¿Qué condiciones debe cumplir el instrumento para que nuestro estimador de VI sea consistente?
  - $C(Z, u) = 0$: No estar correlacionado con la capacidad u otros inobservables que afecten al salario.
  - $C(Z, educ) \neq 0$: Estar correlacionado con la educación.
- Algunos ejemplos de posibles instrumentos (Z) para *educ*: Educación de la madre, educación del padre, número de hermanos, distancia al colegio, etc.
- Disponemos de una muestra **de 336 mujeres casadas.**

Ejemplo: ecuación de salarios

Estimación MCO

- Los resultados de la estimación MCO son:

$$\widehat{\ln(salario)} = \underset{(0.120)}{0.286} + \underset{(0.009)}{0.083}\ educ$$

- La interpretación es que un año adicional de edcuación incrementa el salario en promedio en un 8.3%.

Ejemplo: ecuación de salarios

Estimación por VI (un único instrumento)

- Posible Instrumento: Educación del padre (*educp*)
- **Forma reducida**:

$$\begin{array}{lccc} \widehat{educ} & = & 9.799 & + & 0.282\ educp \\ & & (0.198) & & (0.021) \\ R^2 = 0.196 & & & & \end{array}$$

  - El estadístico t para el instrumento en esta forma reducida es

$$t = 0.282/0.021 \simeq 13.52,$$

    es decir, se rechaza $H_0 : \pi_1 = 0$.
  - Por tanto, la educación de la mujer (*educ*) está significativamente correlacionada con la educación del padre (*educp*).

- **Estimación de VI**:

$$\widehat{\ln(salario)} = \underset{(0.289)}{0.363} + \underset{(0.023)}{0.076}\ educ$$

- Al comparar la estimación MCO con la de VI, sugiere que la estimación MCO es demasiado elevada y está en consonancia con un sesgo positivo del estimador MCO al omitir la capacidad.
- Nótese que los errores estándar de la estimación VI son sustancialmente mayores que los de la estimación MCO, tal y como sugiere la teoría (aunque en todo caso la educación sigue siendo claramente significativa).

- **Contraste de Hausman**
  - A partir de la forma reducida, generamos la variable $\widehat{v}$ como el residuo de dicha ecuación estimada:

$$\widehat{v} = educ - (9.799 + 0.282\ educp),$$

y realizamos la regresión por MCO del modelo

$$\ln(salario) = \beta_0 + \beta_1 educ + \alpha\widehat{v} + e,$$

obteniendo:

$$\widehat{\ln(salario)} = \underset{}{\widehat{\beta}_0 + \widehat{\beta}_1 educ +} \underset{(0.024)}{0.007\ \widehat{v}} + \widehat{e}$$

  - Contrastamos $H_0 : \alpha = 0$ (*educ* es exógena).
    $t = 0.007/0.024 \simeq 0.3$.
    $\Rightarrow$ No se rechaza la exogeneidad de *educ*.

Ejemplo: ecuación de salarios

Estimación por VI (varios instrumentos)

- Supongamos que, además de la educación del padre *educp* disponemos de la educación de la madre *educm* como instrumento. Ahora, la forma reducida sería

$$\begin{array}{lccc} \widehat{educ} = & 8.976 + & 0.183\ educp + & 0.183\ educm \\ & (0.226) & (0.025) & (0.026) \end{array}$$

$$R^2 = 0.245$$

  El estadístico para el contraste de significación conjunta de *educp* y *educm* en esta forma reducida es $W^0 \simeq 243.3$, que se distribuye aproximadamente como una χ^2_2.
- La estimación de MC2E utilizando *educp* y *educm* como instrumentos es ahora

$$\begin{array}{lcc} \widehat{\ln(salario)} = & 0.396 + & 0.074\ educ \\ & (0.272) & (0.022) \end{array}$$

Ejemplo: ecuación de salarios

Estimación por VI (varios instrumentos)

- Para implementar ahora el contraste de Hausman, tomamos el residuo $\widehat{v}$ de la forma reducida

$$\widehat{v} = educ - (8.976 + 0.183\ educp + 0.183\ educm),$$

y realizamos la regresión por MCO del modelo

$$\ln(salario) = \beta_0 + \beta_1 educ + \alpha\widehat{v} + e,$$

obteniendo:

$$\widehat{\ln(salario)} = \underset{}{\widehat{\beta}_0 + \widehat{\beta}_1 educ +} \underset{(0.022)}{0.0107\ \widehat{v}} + \widehat{e}$$

Contrastamos $H_0 : \alpha = 0$ ($educ$ es exógena).
$t = 0.0107/0.022 \simeq 0.5$.
$\Rightarrow$ No se rechaza la exogeneidad de $educ$.

Ejemplo: ecuación de salarios

Contraste de Sargan

- Continuando con el último caso, teníamos dos instrumentos (*educp* y *educm*) para una variable potencialmente endógena (*educ*), con lo que tenemos $q - r = 1$ restricción de sobreidentificación.
- Podemos por tanto evaluar parcialmente la validez de los instrumentos (es decir, la hipótesis nula de exogeneidad) contrastando la no correlación de los instrumentos con el término de error de la ecuación de interés utilizando un contraste de Sargan.

Ejemplo: ecuación de salarios

Contraste de Sargan

- Para ello, calculamos los residuos de la estimación MC2E

$$\widetilde{u} = \ln(\textit{salario}) - (0.396 + 0.074\ \textit{educ})$$

y realizamos la regresión auxiliar de dichos residuos tanto sobre:

- las variables exógenas que haya y
- sobre los instrumentos utilizados,

$$\begin{array}{lccc} \widehat{\widetilde{u}} = & 0.0054 + & 0.0020\ \textit{educp} - & 0.0025\ \textit{educm} \\ & (0.0703) & (0.0075) & (0.0081) \end{array}$$

$$R^2 = 0.0003$$

- Por tanto, el estadístico de contraste es igual a

$$nR^2_{\widetilde{u}} = 0.1008,$$

que tiene un valor muy bajo para una distribución aproximada χ^2_1, con lo que no rechazamos la hipótesis nula de no correlación de los instrumentos con el término de error del modelo.

- En consecuencia, no hay evidencia en contra de la validez de los instrumentos.

Consideraciones finales

- En la práctica, **en muchas situaciones es difícil encontrar instrumentos válidos**, es decir, variables no incluidas en la ecuación de interés que, estando muy correlacionadas con las variables explicativas potencialmente endógenas, no estén correlacionadas con el término de error de la ecuación de interés.
- El problema es que en el contexto de variables económicas, **la mayoría de las variables disponibles son resultado de las decisiones de los agentes**, y por tanto su exogeneidad es muy cuestionable.

Consideraciones finales

- Idealmente, nos gustaría poder contar como variables instrumentales con **variables cuyas realizaciones vinieran dadas** a los agentes económicos objeto de estudio (y fueran por tanto exógenas). Hemos visto como ejemplo el precio de los cigarrillos como instrumento para el número de cajetillas de tabaco consumidas.

- El problema es que, en muchos contextos (como el de dicho ejemplo), la **calidad del instrumento** se ve mermada por la **débil correlación con la variable explicativa endógena** que se desea instrumentar.

Consideraciones finales

- EJEMPLO: La existencia de información pasada de las variables de interés abre posibilidades para encontrar instrumentos adicionales. Así, variables explicativas endógenas podrían instrumentarse mediante los valores que dichas variables tomaron en períodos pasados (dado que los valores pasados de dichas variables están dadas antes de que se realicen los valores corrientes).
  - Por ejemplo, en el contexto de una ecuación de consumo y renta permanente (inobservable) en el que se utiliza en lugar de ésta la renta disponible, lo que induce un problema de endogeneidad por error de medida, si se dispone de la renta disponible del año anterior podría emplearse como instrumento.
  - Si se analiza dicha relación con datos agregados de series temporales, se podría usar la renta disponible desfasada como instrumento.
  - Si se analiza dicha relación con datos de familias y se dispone de datos longitudinales (datos de panel), la renta disponible desfasada de cada familia podría emplearse como instrumento.