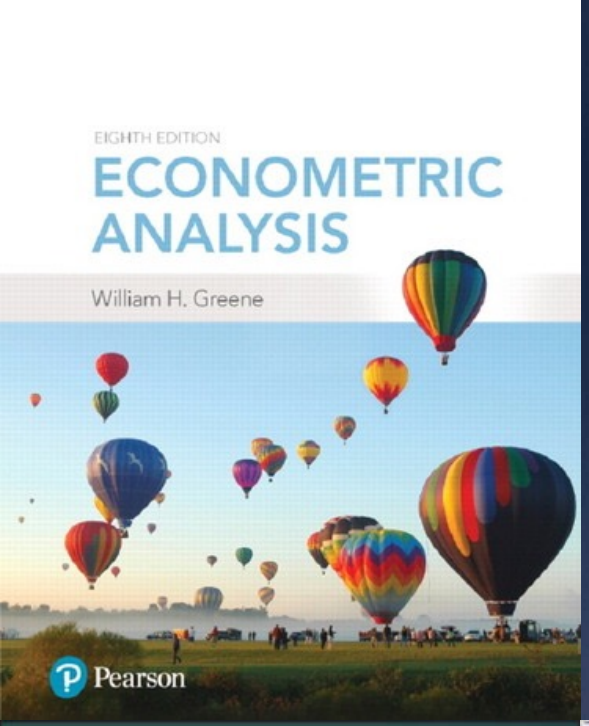
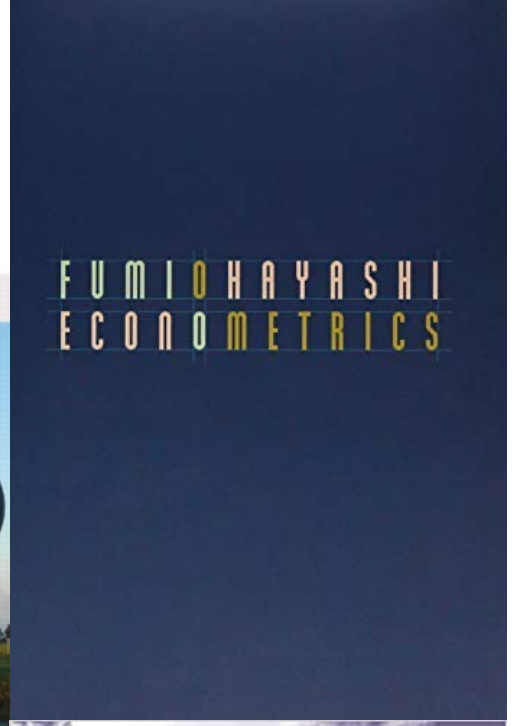
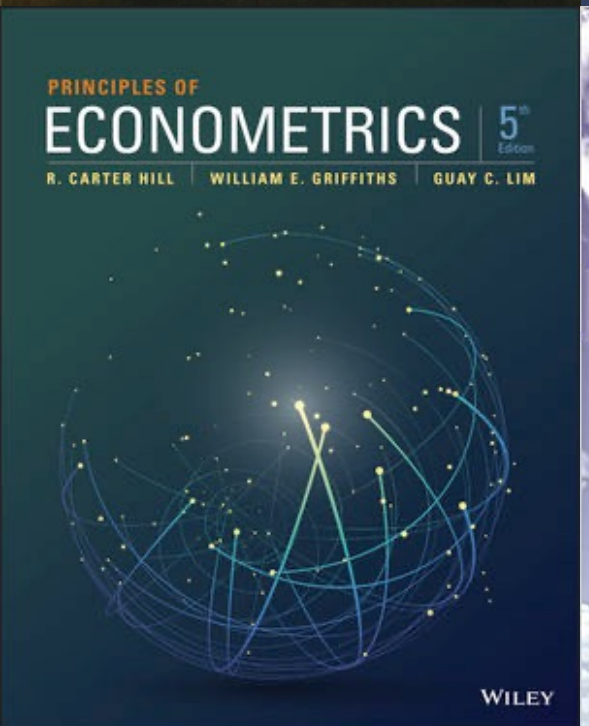
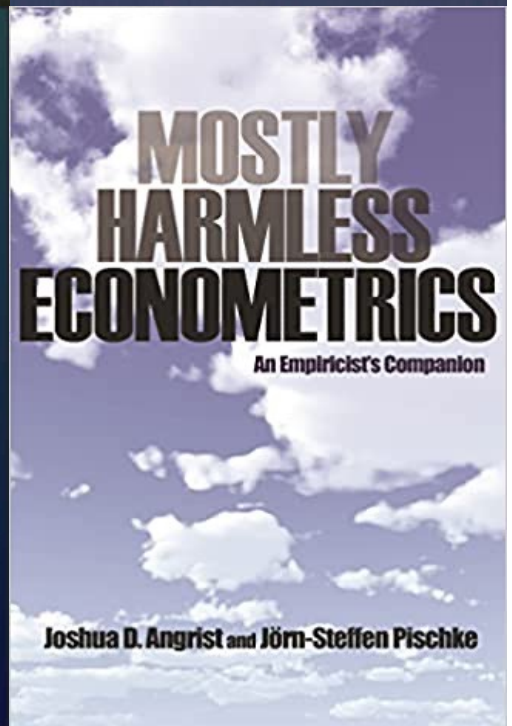

PUCP









MAESTRÍA EN ECONOMÍA
ECONOMETRÍA INTERMEDIA
ECO743 – MÓDULO 2

Sesión 3
Endogeneidad

Docente: Juan Palomino

PUCP

Índice

1 Definición de Endogeneidad

2 ¿Cómo surge la endogeneidad?

3 ¿Qué es un instrumento?

4 Supuestos

5 Estimador de Variables Instrumentales

6 Estimación por Mínimo Cuadrado en Dos Etapas

7 Verificando Condiciones

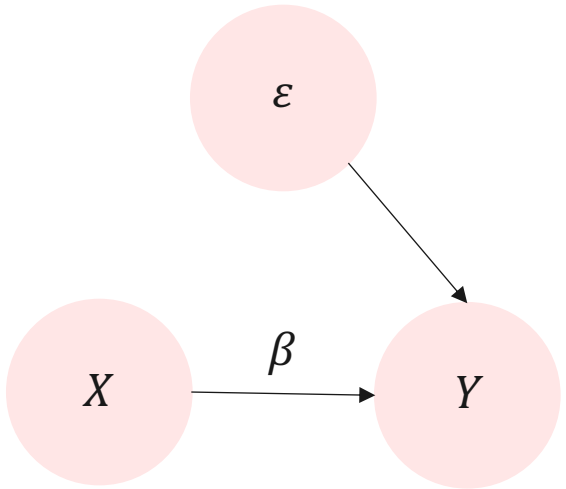
8 Test de Endogeneidad

1. Definición de Endogeneidad

Exogeneidad

- Queremos estimar el efecto causal de un cambio en X sobre Y. Tenemos la siguiente regresión:

$$Y = \beta X + \varepsilon$$

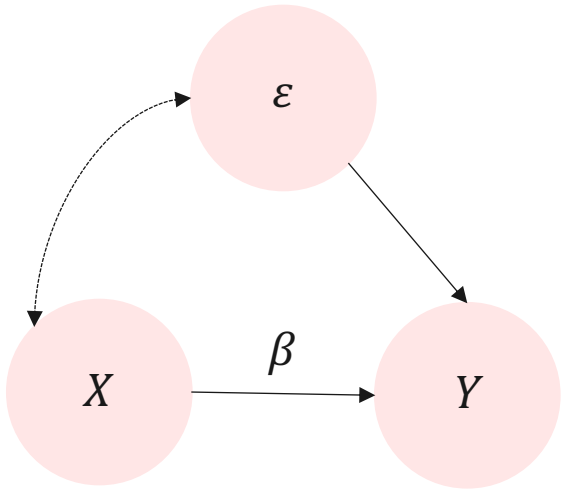


- Si ε no está correlacionado con X, tenemos que el único efecto directo que existe es de X sobre Y vía βX porque no hay asociación entre X y ε

Endogeneidad

- Tenemos $Y = \beta X + \varepsilon$ con la siguiente derivada total:

$$\frac{dY}{dX} = \beta + \frac{d\varepsilon}{dX}$$



- Un efecto directo vía βX y un efecto indirecto vía ε afectando X, el cual afecta también a Y.

Endogeneidad

- En el modelo MCO de regresión múltiple:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

- Si $E(\varepsilon_i | x_i) \neq 0 \; \forall i = 1,2,\ldots,k$ se dice las variables explicativas son endógenas, lo que invalida los estimadores MCO, volviéndolos inconsistentes.

Endogeneidad: $cov(x_i, \varepsilon) \neq 0$

Exogeneidad: $cov(x_i, \varepsilon) = 0$

- Por tanto, se dice que una variable x_j es **endógena** si está correlacionada con ε_i.

2. ¿Cómo surge la endogeneidad?

Endogeneidad

- Endogeneidad se debe a 3 problemas:
    - Simultaneidad
    - Sesgo por variable omitida
    - Error de medición
- Si uno de estos problemas están presentes, los parámetros podrían ser inconsistentes y no podrían medir la magnitud y dirección de la causa, sino solo una simple correlación.

2. ¿Cómo surge la endogeneidad?

2.1 Sesgo por Variable Omitida

Sesgo por Variable Omitida

Un caso común es sospechar que hay una variable omitida q que está correlacionada con x y que explica y:

$$y_i = x_i'\beta_0 + \delta q_i + \varepsilon_i$$

Como no observamos q, no podemos incluirla como control y se encuentra en el término de error, lo que implica $E[\varepsilon_i|x_i] \neq 0$

Sesgo por Variable Omitida

El verdadero modelo $y_i = x_i'\beta_0 + \delta q_i + \varepsilon_i$, pero estimamos $y_i = x_i'\beta_0 + v_i$ donde $v_i = \delta q_i + \varepsilon_i$.

Entonces:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} x_i y_i$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} x_i (x_i'\beta_0 + \delta q_i + \varepsilon_i)$$

$$\hat{\beta} = \beta_0 + \delta \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} x_i q_i + \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} x_i \varepsilon_i$$

Sesgo por Variable Omitida

Por lo tanto:

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n} x_i q_i \xrightarrow{p} Q_{xx}^{-1} E(x_i q_i)$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n} x_i \varepsilon_i \xrightarrow{p} Q_{xx}^{-1} E(x_i \varepsilon_i) = 0$$

Asimismo:

$$\beta_n \xrightarrow{p} \beta_0 + \delta Q_{xx}^{-1} E(x_i q_i)$$

Por lo tanto, el estimador es inconsistente.

2. ¿Cómo surge la endogeneidad?

2.2 Error de medición

Medición de Error

El verdadero modelo: $y_i = x_i'\beta_0 + \varepsilon_i$, pero x_i es medido con errores. Es decir, nosotros observamos $\tilde{x}_i = x_i + v_i$, en vez de x_i.

Asumir que v_i no está correlacionado con x_i, es decir, $E(x_i \cdot v_i) = 0$. Entonces:

$$y_i = x_i'\beta_0 + \varepsilon_i$$

$$y_i = (\tilde{x}_i - v_i)'\,\beta_0 + \varepsilon_i$$

$$y_i = \tilde{x}_i\beta_0 + u_i$$

Donde $u_i = \varepsilon_i - v_i' \cdot \beta_0$

Medición de Error

El problema es que:

$$E[\tilde{x} \cdot u_i] = E[(x_i + v_i)(\varepsilon_i - v_i'\beta_0)]$$

$$= E[x_i\varepsilon_i] - E[x_i v_i'\beta_0] + E[v_i\varepsilon_i] - E[v_i v_i'\beta_0]$$

$$= -E[v_i v_i'\beta_0]$$

$$\neq 0$$

Entonces, para el estimador OLS nosotros tenemos que:

$$plim\hat{\beta}_n = \beta_0 + E(\tilde{x}_i\tilde{x}_i')^{-1}E(\tilde{x}_i u_i)\beta_0 \neq \beta_0$$

Esto es llamado el sesgo por error de medida.

2. ¿Cómo surge la endogeneidad?

2.3 Sesgo por simultaneidad

Sesgo por simultaneidad

Ésta surge cuando una o más de las variables explicativas se determina conjuntamente con la variable dependiente.

Suponga el modelo de regresión:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Pero $x_i = f(y)$. Entonces x afecta a y, pero también ocurre que y afecta a x. ¿Qué debería suceder con β?

3. ¿Qué es un instrumento?

Definición de un Instrumento

- Consideremos el modelo lineal de k variables

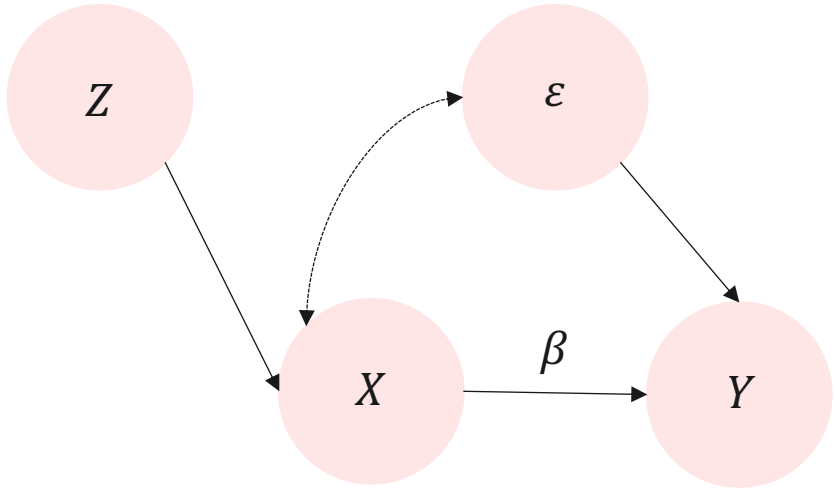
$$y = X\beta + \varepsilon$$

en donde algunos de los regresores están correlacionados con ε (regresores endógenos), mientras que otros no lo están (regresores estrictamente exógenos).

- La idea de las estimaciones con variables instrumentales es detectar los movimientos en x no correlacionados con el error.
- Debemos definir un instrumento. Supongamos que contamos con l variables instrumentales $Z = [Z_1, Z_2, \dots, Z_l]$, donde algunas de las variables en Z podrían ser las mismas que los regresores exógenos. Esta matriz Z es de dimensiones $n \times l$.

Definición de un Instrumento

Un instrumento debe cumplir dos propiedades:



1. **Condición de Exogeneidad o Exclusión**: Z no este correlacionado con el error ε. Tenemos que preguntarnos si Z tiene una asociación con Y, independientemente de su asociación con Y a través de X.

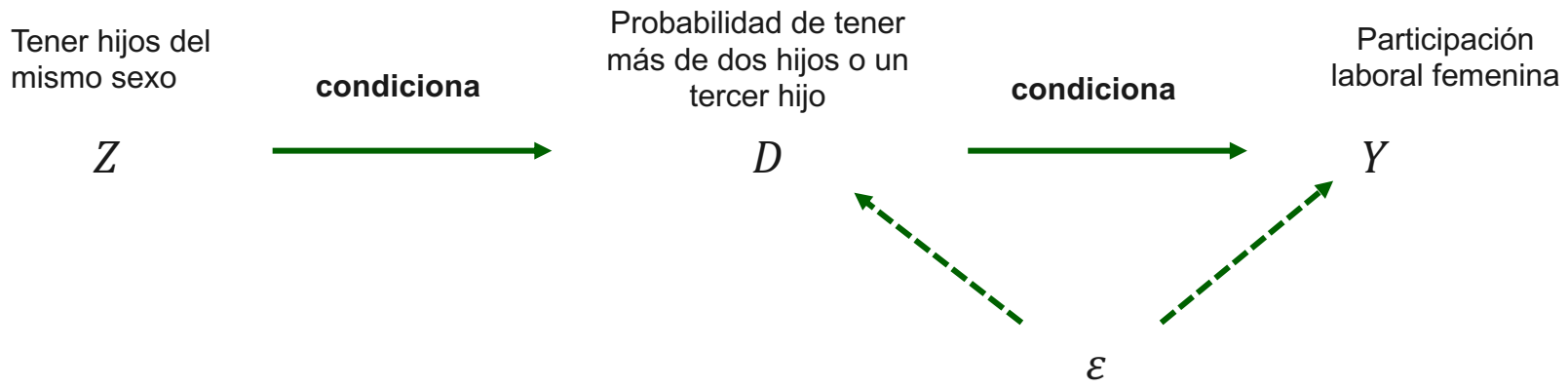
2. **Condición de Relevancia**: Z está correlacionado con el regresor X.

Ejemplo 1

Consideremos el siguiente modelo:

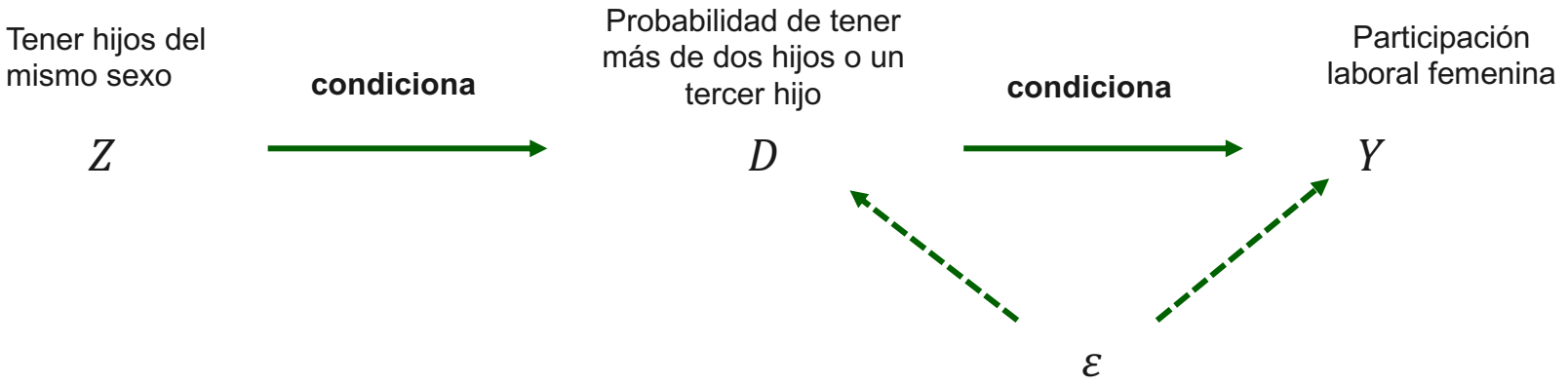
$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

Entonces, podemos establecer una variable instrumental Z que afecta a D y que a su vez afecta a Y solo a través de D



No habría porqué pensar que tener dos hijos del mismo sexo vayan incidir en mis outcomes laborales.

Ejemplo 1



Si podemos aislar la parte que no está correlacionada con ε, podemos estimar β_1 (el efecto del tratamiento) de modo consistente. En otras palabras las variables instrumentales solo identifican un efecto causal para cualquier grupo de unidades cuyos comportamientos se modifican como resultado del instrumento (cumplidores o compliers). Es un efecto Local. Solo de los que pueden tener más de dos hijos.

Incumplimiento

La idea de cumplimiento (**compliance**) refiere a si el tratamiento aplicado coincide con la asignación de este. Full compliance: 100% de los asignados al grupo de tratamiento reciben el tratamiento.

Incumplimiento (**noncompliance**) ocurre cuando el tratamiento aplicado y la asignación de este no coinciden. En otras palabras, algunos sujetos asignados al grupo de tratamiento en realidad no reciben el tratamiento. Ningún sujeto en el grupo de control es tratado.

- **Compliers**: Los individuos que toman el tratamiento son efectivamente los grupos asignados a este. La subpoblación cuyo estado de tratamiento se ve afectado por el instrumento en la dirección correcta.
- **Never Takers**: Independiente de su condición de asignación, los individuos nunca tomaran el tratamiento. La subpoblación de unidades que nunca toman el tratamiento independientemente del valor del instrumento.
- **Always Takers**: Independiente de su condición de asignación, los individuos siempre van a tomar el tratamiento. La subpoblación de unidades que siempre toman el tratamiento independientemente del valor del instrumento.
- **Defiers**: la subpoblación cuyo estado de tratamiento se ve afectado por el instrumento en la dirección incorrecta.

Incumplimiento

Nota:

Las variables instrumentales lo que hacen es identificar un efecto causal de un X sobre un Y para los compliers

Los compliers son las personas que deciden tener un tercer hijo dado que sus dos hijos fueron del mismo sexo.

Ejemplo 2

Tenemos:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + \varepsilon$$

donde:

- $health$ es una medida del estatus de salud de los individuos,
- $work$ son las horas trabajadas semanalmente, y
- $exercise$ son las horas de ejercicio por semana.

¿Por qué exercise puede ser una variable endógena?

Instrumentos: distancia desde casa ($dhome$) y distancia desde el trabajo ($dwork$) al gimnasio o al club de salud más cercano.

4. Supuestos

Supuestos

Supuesto 1: Linealidad

La ecuación a ser estimada es lineal:

$$y_i = x_i'\beta_0 + \varepsilon_i, \qquad i = 1, \dots, n$$

Donde x_i es un vector de regresores de dimensión K, β_0 es un vector de coeficientes de dimensión K, y ε_i son términos de errores no observables.

Supuestos

Supuesto 2: Muestra Aleatoria

Sea z_i un vector de instrumentos de dimensión L, y sea w_i el elemento no constante y único de $(y_i, \mathrm{x_i}, \mathrm{z_i})$. $\{w_i\}$ es i.i.d.

Supuestos

Supuesto 3: Condiciones de Ortogonalidad

Instrumentos no están correlacionados con el término de error. Todas las L variables en z_i son predeterminadas en el sentido que ellos son ortogonales al término de error: $E(z_{il}\varepsilon_i) = 0$ para todo i y l $(l = 1,2,\dots,\mathrm{L})$.

$$\mathrm{E}[z_i \cdot (y - x_i'\beta_0)] = 0$$

Se denota también como:

$$\mathrm{E}(g_i) = 0$$

Donde $g_i = z_i \cdot \varepsilon_i$ y $\varepsilon_i = y_i - x_i'\beta_0$

Supuestos

Ejemplo: Salarios

Considerar

$$\text{wage}_i = \beta_1 + \beta_2 sch_i + \beta_3 exper_i + \varepsilon_i$$

Nuestro instrumento es $prox_i$. Por lo tanto:

$$y_i = \text{wage}_i, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \qquad z_i = \begin{pmatrix} 1 \\ exper_i \\ prox_i \end{pmatrix} \quad K = 3, L = 3$$

De tal manera que:

$$E\begin{pmatrix} (\text{wage}_i - \beta_1 - \beta_2 sch_i - \beta_3 exper_i) \\ \text{exper}_i(\text{wage}_i - \beta_1 - \beta_2 sch_i - \beta_3 exper_i) \\ \text{prox}_i(\text{wage}_i - \beta_1 - \beta_2 sch_i - \beta_3 exper_i) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Supuestos

Supuesto 4: Condición Rango para identificación

La matriz $E(z_i x_i') = Q_{zx}$ de dimensión $L \times K$ es de rango completo columna (es decir, su rango igual K).

Identificación

Como un ejemplo, considerar un modelo con solo una covariable y un instrumento: $z_i = (1, z_i)'$ y $x_i = (1, x_i)'$. Entonces:

$$Q_{zx} = \begin{pmatrix} 1 & E(z_i) \\ E(z_i) & E(z_i x_i) \end{pmatrix}$$

El determinante de Q_{zx} no es cero (es rango columna completo) si y solo si $cov(z_i, x_i) = E(z_i x_i) - E(x_i) \cdot E(z_i) \neq 0$

Identificación

Escribimos la condición de momentos de ortogonalidad como:

$$E[g(w_i, \beta)] = 0 \text{ donde } g_i = g(w_i; \beta) \equiv z_i \cdot (y_i - x_i'\beta)$$

Considerar un estimador $\hat{\beta}_{K\times 1}$ de β_0. Entonces, tenemos un sistema de L ecuaciones simultáneas en K incógnitas:

$$E\left[g\left(w_i, \hat{\beta}\right)\right] = 0$$

Identificación

Ya que el modelo es lineal se puede escribir como:

$$E[g(w_i, \beta)] = E[z_i \cdot (y_i - x_i'\beta)] = E(z_i \cdot y_i) - E(z_i x_i')\hat{\beta} = 0$$

o

$$\underset{(L\times K)}{Q_{zx}} \quad \underset{(K\times 1)}{\hat{\beta}} = \underset{(L\times 1)}{q_{zy}}$$

La única solución es que $\hat{\beta} = \beta_0$ si y solo si Q_{zx} es de rango completo.

Condición de Orden para Identificación

- Ya que el rango $(Q_{zx}) < K$ si $L < K$, una condición necesaria para identificación es que $L \geq K$.
- En otras palabras, el número de variables predeterminados debe ser mayor o igual al número de variables exógenas.
- El número de instrumentos debe ser mayor o igual al número de variables endógenas.

1. La ecuación es **sobreidentificada** si la condición de rango se cumple y $L > K$
2. La ecuación es **identificada** exactamente si la condición de rango se cumple y $L = K$
3. La ecuación es **subidentificada** (o no identificada) si la condición de orden no se cumple, es decir, $L < K$.

5. Estimador de Variable Instrumental

Estimador Variable Instrumental

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).

Estimador Variable Instrumental

Si reemplazamos las condiciones de momento por los momentos muestrales, tenemos:

$$g_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g(w_i;\hat{\beta})$$

$$= \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\hat{\beta})$$

$$= \frac{1}{n}\sum_{i=1}^{n} z_i y_i - (\frac{1}{n}\sum_{i=1}^{n} z_i x_i')\,\hat{\beta}$$

$$= \frac{1}{n}(\mathrm{Z}'y - Z'X\hat{\beta})$$

$$g_n(\hat{\beta}) = s_{zy} - S_{zx}\,\hat{\beta}$$

Estimador Variable Instrumental

Entonces, la muestra analógica $g_n(\hat{\beta}) = 0$ es un sistema de ecuación lineal L en K incógnitas:

$$S_{zx}\hat{\beta} = s_{zy}$$

Si $K = L$ (la ecuación es exactamente identificada), entonces Q_{zx} es cuadrada e invertible y el sistema de ecuaciones simultaneas tiene una solución única dada por:

$$\hat{\beta}_{IV} = S_{zx}^{-1} s_{zy}$$

$$\hat{\beta}_{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} z_i y_i$$

$$\hat{\beta}_{IV} = (Z'X)^{-1}\mathrm{Z}'y$$

Se llama el estimador de variables instrumentales.

Si $z_i = x_i$, es decir, los regresores son ortogonales al término de error, entonces $\hat{\beta}_{IV}$ se reduce al estimador MCO.

Estimador VI en modelo simple

La idea del método de VI es que dado el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Donde x es una variable endógena $(corr(x_i, \varepsilon_i) \neq 0)$ que hace inconsistente el estimador MCO.

En consecuencia, necesitamos un instrumento (z) para aislar la parte de x no correlacionada con ε. Por lo que, este instrumento debe cumplir las condiciones de exogeneidad y relevancia.

Estimador VI en modelo simple

Dada la exogeneidad del instrumento $Cov(Z_i, \varepsilon_i) = 0$, utilizando el método de los momentos:

$$E(\varepsilon) = E(y - \beta_0 - \beta_1 x) = 0$$
$$E(\varepsilon z) = Cov(Z_i, y_i - \beta_0 - \beta_1 x_{i1}) = 0$$

De la primera ecuación se obtiene la constante:

$$\frac{1}{N}\sum(y - \beta_0 - \beta_1 x) = 0$$
$$\tilde{\beta}_0^{IV} = \bar{y} - \tilde{\beta}_1^{IV}\bar{x}$$

El estimador de VI se puede obtener (si $z = x, \tilde{\beta}_i^{IV} = \tilde{\beta}_i^{MCO}$):

$$\tilde{\beta}_i^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Estimador VI en modelo simple

Por tanto, cuando el $corr(Z_i, \varepsilon_i) \neq 0$ el estimador de VI es inconsistente (sesgo de consistencia):

$$\tilde{\beta}_i^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})Y_i}{\sum_{i=1}^{n}(z_i - \bar{z})X_i} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^{n}(z_i - \bar{z})X_i}$$

$$\tilde{\beta}_i^{IV} = \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\varepsilon_i)}{\sum_{i=1}^{n}(z_i - \bar{z})X_i} \quad \xrightarrow{p} \quad \beta_1 + \frac{cov(z_i, \varepsilon_i)}{cov(z_i, x_i)} = \beta_1$$

Varianza del estimador VI

En general, el estimador de VI tendrá una varianza mayor que el de MCO. Wooldridge (2009) muestra que la varianza asintótica del estimador es:

$$var(\tilde{\beta}_i^{IV}) = \frac{\hat{\sigma}^2}{n\sigma_x^2\rho_{xz}^2}$$

Siendo $\hat{\sigma}^2 = \frac{\sum \tilde{\varepsilon}^2}{n-k}$, estimado con el residuo del modelo de VI; ρ_{xz}^2 es el cuadrado de la correlación poblacional entre x y z (solo en el caso de regresión simple); σ_x^2 es la varianza poblacional de x.

Varianza del estimador VI

$$var(\tilde{\beta}_i^{IV}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 R_{xz}^2}$$

Por tanto, si x es exógena, realizar VI en vez de MCO tiene un coste en término de eficiencia, en tal sentido, a menor correlación, mayor varianza de VI respecto a MCO (recordar que la varianza muestral de x, $\sigma_x^2 = \frac{STC_x}{n}$).

Dado que esta estimación difiere de la de MCO $\left(var(\tilde{\beta}_i^{MCO}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$ por R_{xz}^2, que al ser siempre menor que 1, $var(\tilde{\beta}_i^{IV}) > var(\tilde{\beta}_i^{MCO})$

Varianza del estimador VI

La desviación estándar del coeficiente se puede utilizar para obtener los estadísticos t y realizar inferencia de la forma habitual.

$$t_{\hat{\beta}_j}^{IV} = \frac{\hat{\beta}_j - \beta_{ho}}{de(\tilde{\beta}_i^{VI})}$$

6. Estimación por Mínimos Cuadrados en Dos Etapas

Estimación por MC2E

El método permite emplear más de una variable explicativa exógena como instrumento (Wooldridge, 2009).

Este estimador se presenta como un procedimiento en dos etapas.

En una primera etapa se elimina la correlación entre la endógena y el error, mediante instrumentos (variables exógenas) que están altamente correlacionadas con la variable explicativa de interés.

Dado dos instrumentos válidos (z_1 y z_2), se podría utilizar cualquiera de estos para obtener VI, utilizar una combinación de ambos será siempre más eficiente.

Estimación por MC2E

Un caso sencillo con $k = 3$ variables explicativas (incluyendo a la constante de unos), en donde la última variable presenta correlación con el error.

$$Y_i = \underbrace{\beta_1 + \beta_2 X_{2i}}_{\text{(No correlacionados con } \varepsilon_i)} + \underbrace{\beta_3 X_{3i}}_{\text{(Correlacionados con } \varepsilon_i)} + \varepsilon_i$$

Matricialmente:

$$y = X_2\beta_2 + X_3\beta_3 + \varepsilon$$

En donde X_2 es una matriz $n\times 2$ y X_3 es una matriz $n\times 1$ que contiene al regresor endógeno, donde $Cov(X_2, \varepsilon) = 0$ y $Cov(X_3, \varepsilon) \neq 0$.

Estimación por MC2E

Supongamos que contamos con m variables $W_{1i}, W_{2i}, \dots, W_{mi}$, que cumple las condiciones de relevancia y exogeneidad de las variables instrumentales. Agrupamos a estas variables en una matriz W de dimensión $n \times m$.

Procedimientos (Primera Etapa)

1. Regresionar por MCO al regresor endógeno X_{3i} contra la constante, la variable X_{2i} y todas las variables en la matriz W. Explícitamente se estima la regresión:

$$X_{3i} = \gamma_1 + \gamma_2 X_{2i} + \gamma_k W_{1i} + \gamma_{k+1} W_{2i} + \cdots + \gamma_{k-1+m} W_{mi} + \xi_{1i}$$

2. Luego se calcula la predicción $\hat{X}_{3i}$. Matricialmente, la regresión se escribe como:

$$X_3 = X_2\gamma_1 + W\gamma_2 + \xi = \mathrm{Z}\gamma + \xi$$

3. El estimador MCO es $\hat{\gamma} = (Z'Z)^{-1}Z'X_3$ y las predicciones son:

$$\hat{X}_3 = Z(Z'Z)^{-1}Z'X_3 = P_Z X_3$$

donde $P_Z = Z(Z'Z)^{-1}Z'$ es la matriz de proyección.

Procedimientos (Segunda Etapa)

4. Utilizar a la predicción $\hat{X}_3$ en lugar de X_3 en la ecuación (1) y estimar por MCO la ecuación:

$$y = X_2\beta_2 + \hat{X}_3\beta_3 + \eta$$

5. En términos matriciales:

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

$$= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

El estimador MCO de esta ecuación es el estimador de MC2E, el cual es consistente de los parámetros poblacionales.

Procedimientos (Segunda Etapa)

6. La estimación por MCO de la segunda etapa no entrega las desviaciones estándar correctas del estimador MC2E.

   La matriz de varianzas y covarianzas correcta es:

$$Var(\hat{\beta}_{MC2E}|X) = \hat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X)^{-1}$$
$$= \hat{\sigma}^2(X'PX)^{-1}$$

   Donde σ^2 puede ser estimado mediante:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

$$\hat{\varepsilon} = y - X\hat{\beta}_{MC2E}$$

Consistencia

1. Reemplazando $y = X\beta + \varepsilon$ en el estimador $\hat{\beta}_{MC2E}$ y multiplicando y dividiendo por n, se obtiene:

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

$$= \beta + (X'P_Z X)^{-1} X'P_Z \varepsilon$$

$$= \beta + \underbrace{\left(\frac{1}{n} X'P_Z X\right)^{-1}}_{(a)} \underbrace{\left(\frac{1}{n} X'P_Z \varepsilon\right)}_{(b)}$$

Consistencia

2. Tomando plim al argumento entre paréntesis del término (a):

$$plim\left(\frac{1}{n}X'P_ZX\right) = plim(\frac{1}{n}X'Z(Z'Z)^{-1}Z'X)$$

$$= plim\left(\frac{1}{n}X'Z\right)plim\left(\frac{1}{n}Z'Z\right)^{-1}plim(\frac{1}{n}Z'X)$$

$$= Q_{XZ}Q_{ZZ}^{-1}Q'_{XZ} \neq 0$$

Consistencia

3. Tomando plim al argumento entre paréntesis del término (b):

$$plim\left(\frac{1}{n}X'P_Z\varepsilon\right) = plim(\frac{1}{n}X'Z(Z'Z)^{-1}Z'\varepsilon)$$

$$= plim\left(\frac{1}{n}X'Z\right)plim\left(\frac{1}{n}Z'Z\right)^{-1}plim(\frac{1}{n}Z'\varepsilon)$$

$$= Q_{XZ}Q_{ZZ}^{-1}0 = 0$$

4. Reemplazando estos dos argumentos, se obtiene:

$$plim(\hat{\beta}_{MC2E}) = \beta$$

7. Verificando Condiciones

7.1 El problema de los instrumentos débiles

Instrumentos Débiles

Propiedades del estimador IV pueden ser pobres y el estimador puede ser severamente sesgado, si el instrumento exhibe solamente correlación con los regresores endógenos.

Considerar el siguiente modelo $y = \beta_0 + \beta_1 x_1 + \varepsilon$ donde z_1 como instrumento para x_1. El estimador IV es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})y_i}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Entonces, si $cov(z, x) \neq 0$, el plim del estimador IV es:

$$plim\hat{\beta}_1 = \beta_1 + \frac{cov(z, \varepsilon)}{cov(z, x)}$$

Cuando $cov(z, \varepsilon) = 0$ obtenemos resultados consistentes. Sin embargo, si z tiene alguna correlación con ε, el estimador es inconsistente.

Instrumentos Débiles

Reescribimos:

$$plim\hat{\beta}_1 = \beta_1 + \frac{\sigma_\varepsilon corr(z,\varepsilon)}{\sigma_x corr(z,x)}$$

De aquí vemos:

- Si z y ε están correlacionados, la inconsistencia en el estimador IV se vuelve grande a medida que $corr(z,x)$ se acerca a cero.
- Una correlación pequeña entre z y ε puede causar inconsistencia severa y un sesgo de muestra finito severo, si z solo está debilmente correlacionado con x.

Instrumentos Débiles

En tales casos, puede ser mejor usar MCO, incluso si solo nos enfocamos en la inconsistencia en los estimadores: tenga en cuenta que el límite del estimador MCO es:

$$plim\hat{\beta}_{MCO,1} = \beta_1 + \frac{\sigma_\varepsilon}{\sigma_x} corr(x, \varepsilon)$$

La comparación de estas fórmulas muestra que se prefiere IV a MCO en el terreno de sesgo asintótico cuando:

$$\frac{corr(z, \varepsilon)}{corr(z, x)} < corr(x, \varepsilon)$$

Además:

$$\frac{plim\hat{\beta}_{IV} - \beta}{plim\hat{\beta}_{MCO} - \beta} = \frac{corr(z, \varepsilon)}{corr(x, \varepsilon)} < \frac{1}{corr(z, x)}$$

Por lo tanto, con un instrumento invalido y una baja correlación entre el instrumento y el regresor, el estimador IV puede ser aún más inconsistente que MCO.

Aplicación Monte Carlo

El proceso generador de datos:

$$y_i = \theta + \beta x_i + u_i$$

$$x_i = \alpha + \gamma z_i + p u_i + \varepsilon_i$$

Donde $u_i \sim N(0, \sigma_u^2), \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ y el vector de innovación (u_i, ε_i) es independientemente distribuido.

El intercepto es $\theta = 0.5$. El γ controla la fuerza de los instrumentos z_i, ρ controla el monto de la correlación entre x_i y u_i, σ_ε^2 puede ser usado para controlar la variabilidad relativa de x_i y u_i es un problema de error en variables.

Aplicación Monte Carlo

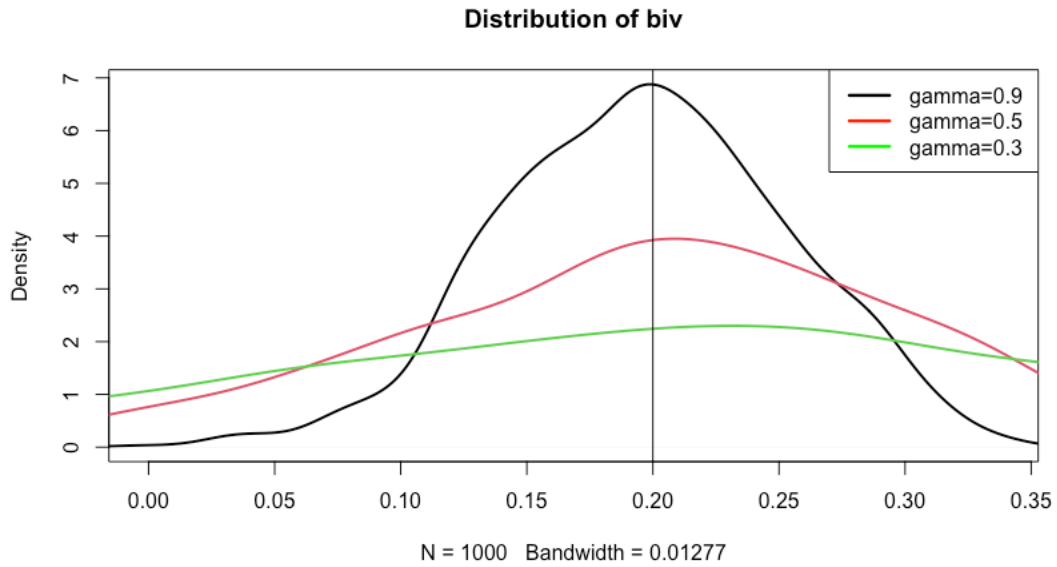


Figura 1. Distribución de β_{IV}

Instrumentos Débiles

Una forma de ver si los instrumentos están correlacionados con el regresor endógeno es a través de la prueba F en la primera etapa del procedimiento de dos etapas (Staiger y Stock, 1997).

La regla de oro aplicable para el caso de un solo regresor endógeno dice que:

> Si el estadístico F de significancia conjunta que prueba la hipótesis H_0: $\gamma = 0$ es mayor a 10, entonces los instrumentos son relevantes.

7. Verificando Condiciones

7.2 Validez de la exogeneidad de los instrumentos

Validez de la exogeneidad de los instrumentos

Test de la validez de la exclusión de W de la ecuación principal asignándoles un valor de cero a sus hipotéticos parámetros (restricción de exclusión).

Test de Sargan y su generalización para errores robustos en el test J de Hansen (Hansen, 1982), puede aplicarse al caso en que el número de instrumentos excluidos es mayor al número de regresores endógenos, o caso sobreidentificado.

La única diferencia entre ambos tests es que el de Sargan asumen homocedasticidad condicional.

Validez de la exogeneidad de los instrumentos

Los pasos del test de Sargan son:

1. Estimar los parámetros de la ecuación (1) por MC2E utilizando los instrumentos propuestos. Calcular $\hat{Y}_1 = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
2. Calcular $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$
3. Regresionar $\hat{\varepsilon}_i$ sobre todos los instrumentos $X_{2i}, W_{1i}, \dots, W_{mi}$
4. Hallar el estadístico F que contrasta la hipótesis que los coeficientes de $W_{1i}, \dots, W_{mi}$ son iguales a cero.
5. Bajo la hipótesis nula de instrumentos exógenos, el valor $J = mF$ se distribuye asintóticamente como un χ^2_{m-1}
6. Si J supera al valor crítico respectivo, se rechaza la hipótesis nula de instrumentos exógenos; si es inferior, se acepta la nula.

8. Test de Endogeneidad

Test de Endogeneidad

Se puede hacer una prueba estadística que confirme o rechace la hipótesis que un regresor sea endógeno.

Test de Hausman: comparar estimadores MCO y MC2E.

- Si todos los regresores son exógenos (H_0), entonces tanto MCO como MC2E son consistentes, pero MCO es más eficiente.
- Si hay regresores endógenos (hipótesis alternativa), solo MC2E es consistente.

Test de Endogeneidad

El test de Hausman:

$$H = n(\hat{\beta}_{IV} - \hat{\beta}_{MCO})'[Var(\hat{\beta}_{IV}) - Var(\hat{\beta}_{MCO})]^{-1}(\hat{\beta}_{IV} - \hat{\beta}_{MCO})$$

Bajo la H_0, H se distribuye asintóticamente como una chi-cuadrado con un grado de libertad (el número de regresores endógenos).


PUCP