Econometría II

CAPITULO 1 Variables omitidas y error en las variables. Variables instrumentales (VI)

Profesora: Graciela Sanroman

Año: 2016

En este capítulo nos concentraremos en como obtener buenas estimaciones de los efectos parciales con el objetivo de estudiar relaciones causales, no prestaremos atención al tema de la predicción.
El objetivo de la mayoría de los estudios en economía es determinar si el *cambio en una variable* (x) *determina un cambio en otra variable* (y), dejando un conjunto de otras variables constante.

Ejemplos:

- ¿cursar un año adicional de educación tiene algún efecto en el salario?,
- ¿reducir los impuestos municipales determina un aumento en el nivel de actividad de la ciudad?
- ¿reducir el tamaño de los grupos en las escuelas mejora los resultados escolares de los alumnos?
- ¿un cambio en el volumen de producción determina una variación en el costo unitario de producción?
- ¿cuánto influye un aumento de los precios de un bien en su demanda?
- ¿un aumento en el ingreso tiene efectos en el consumo?
- ¿aumentar los años de educación de la población tiene algún efecto en el producto y el crecimiento?

Relaciones causales

- La noción de *ceteris paribus* es el núcleo del establecimiento de una relación causal.
- La pregunta central que intentaremos abordar es ¿cómo podemos trasladar este análisis de ceteris paribus a un *contexto probabilístico*?
- Sean y y $\mathbf{x} = (x_1, x_2, ..., x_k)$ variables aleatorias. Si quiero centrarme en el efecto que tiene sobre y el cambio en una variable digamos x_j, dejando el resto de las variables constantes difícilmente me resultará suficiente el estudiar la correlación entre ambas variables. En cambio utilizamos *métodos econométricos* que permiten identificar ese efecto.

Esperanza condicional

- Covarianza, correlación: se trata de *relaciones simétricas* entre dos variables aleatorias.
- En ciencias sociales quisieramos explicar una variable aleatoria Y (por ejemplo salario por hora) en términos de otra variable aleatoria X (ej: edad, educación, género).
- Podemos resumir la relación entre Y y X considerando la *esperanza condicional de Y dada X* la escribimos como $\mathbf{E}(y \mid x) = \mathbf{E}(y \mid x_1, x_2 .... x_j)$ y la llamamos indistintamente esperanza condicional o media condicional.

Efectos parciales

- Si x_j es una variable continúa podemos obtener el efecto parcial

$$\frac{\partial \mathbb{E}(y \mid \mathbf{x})}{\partial x_j}$$

que mide como espero que cambie en promedio y cuando x_j cambia, manteniendo constantes las restantes X.

- Pero también podría considerar la distribución de probabilidad condicional $\mathbb{P}(y \mid X)$

$$\frac{\partial \mathbb{P}(y \mid \mathbf{x})}{\partial x_j}$$

¿Es eso un análisis ceteris paribus?

Depende de la existencia de variables no observables (omitidas) que estén correlacionadas con x_j.

Consideremos el caso más general donde y es una variable resultado, x son variables observables que influyen en y y u son variables inobservables que influyen en y,

$$y = \phi(\mathbf{x}, u)$$

El efecto ceteris paribus es

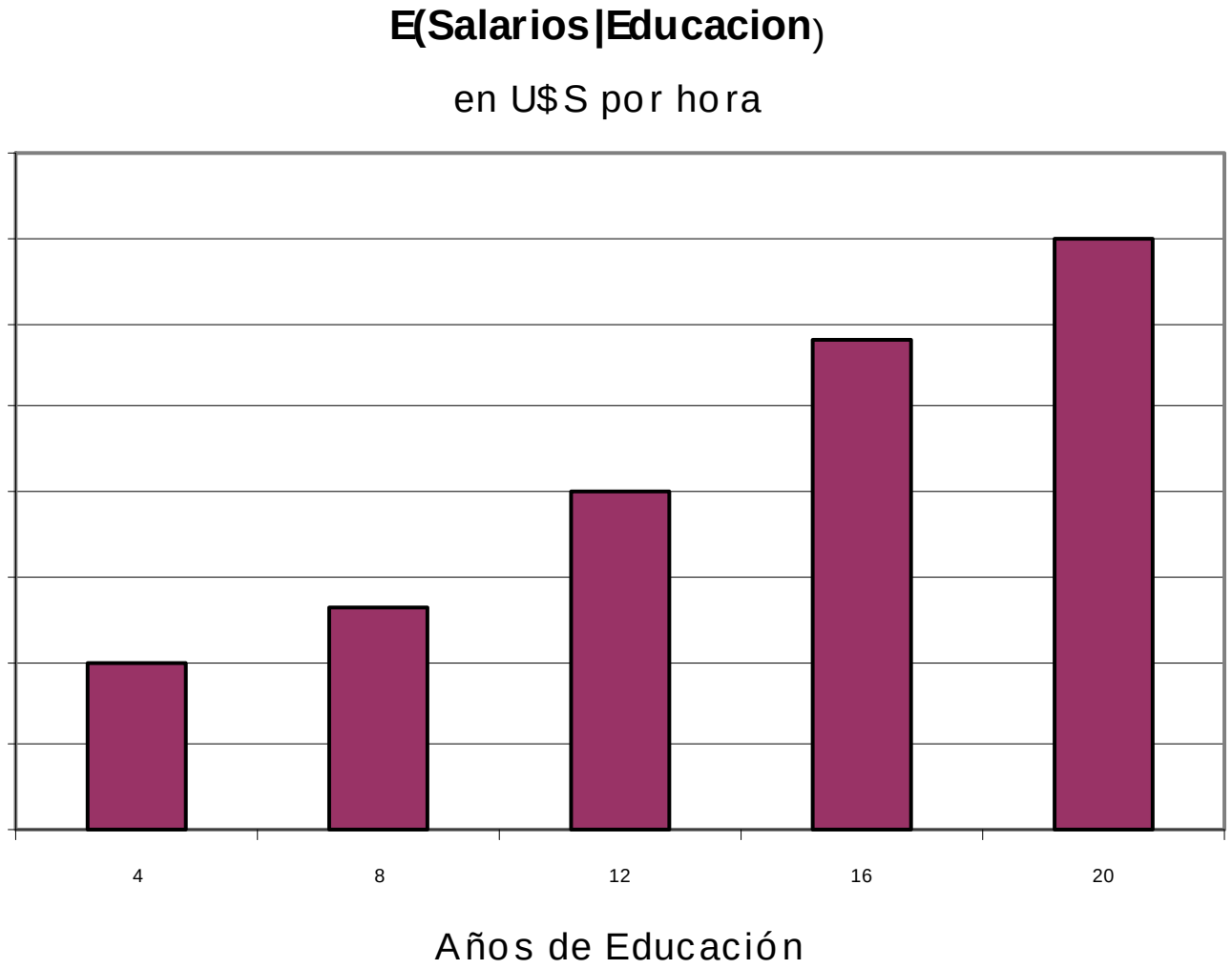
$$\phi_j(\mathbf{x}, u) = \frac{\partial \phi(\mathbf{x}, u)}{\partial x_j}$$

- pero tendremos dificultades en identificarlo, siempre que exista una correlación entre u y x_j ya que

$$\frac{d\phi(\mathbf{x}, u)}{dx_j} = \frac{\partial \phi(\mathbf{x}, u)}{\partial x_j} + \frac{\partial \phi(\mathbf{x}, u)}{\partial u} \frac{\partial u}{\partial x_j}$$

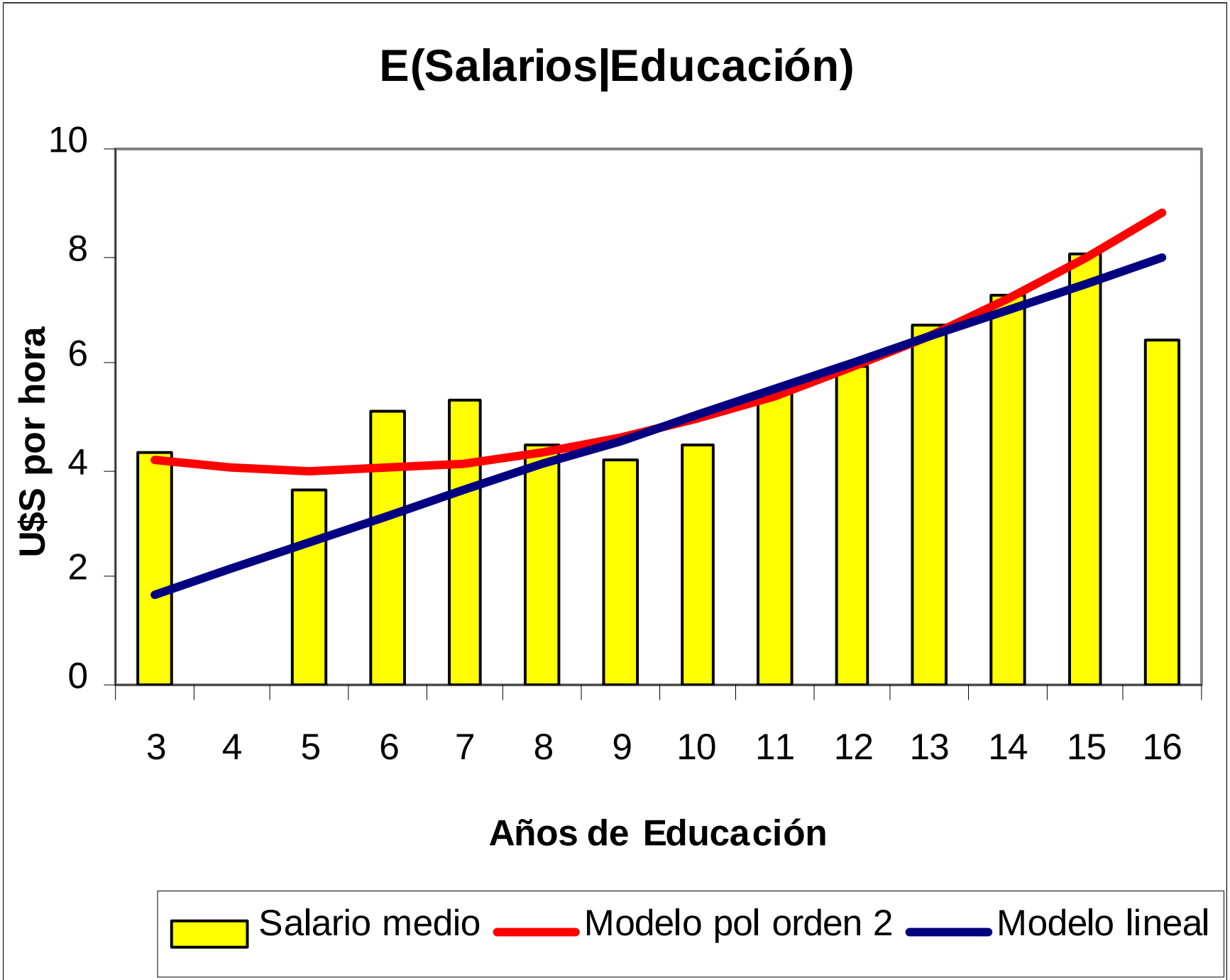
Esperanza condicional

Ejemplo: Salario por hora en función de los años de educación



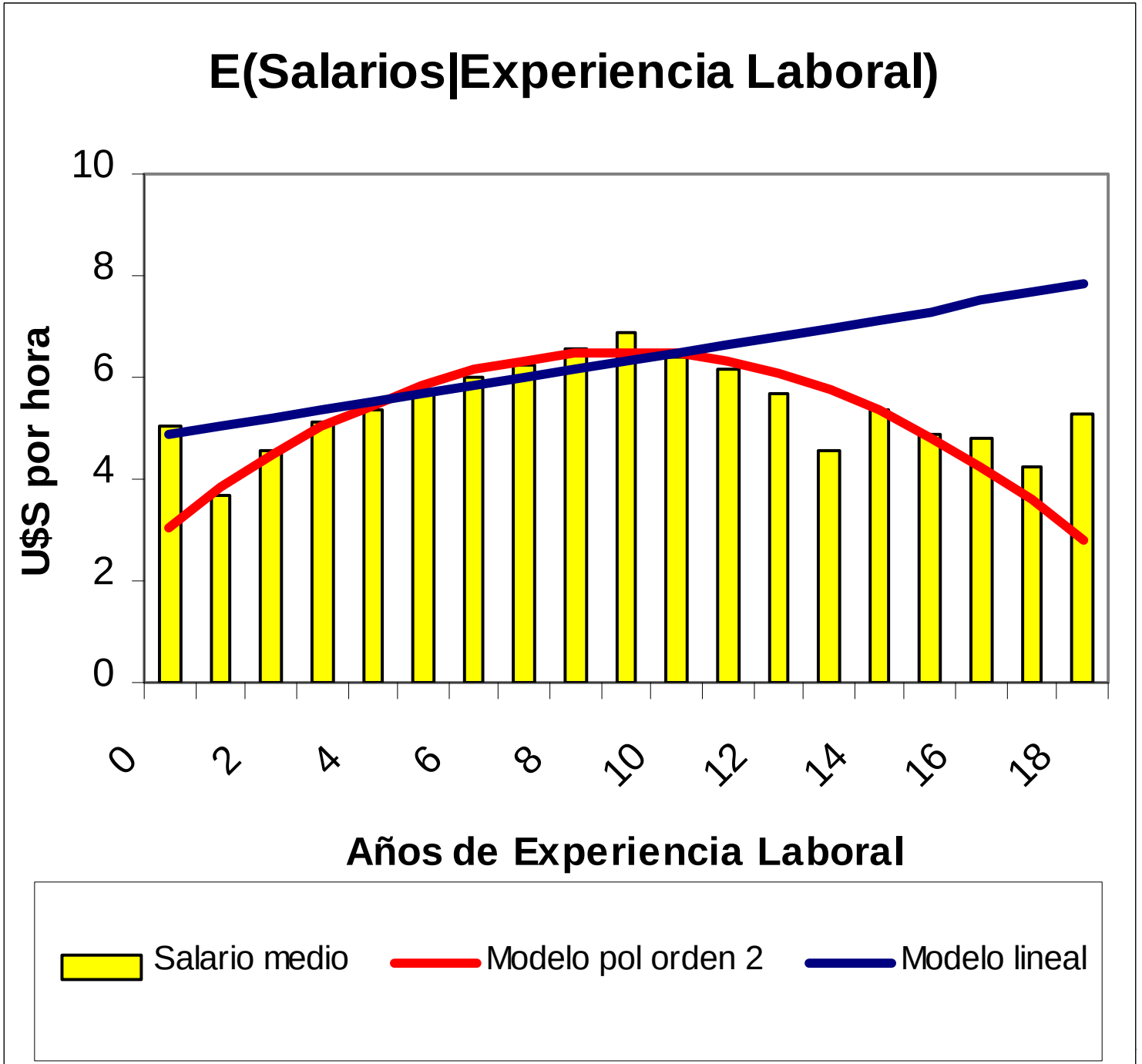
Ejemplo: Salario por hora dado años de educación

Años educ.	No. Observaciones	Salario promedio
3	8	4.32
4	0	sin dato
5	16	3.61
6	40	5.09
7	16	5.32
8	144	4.50
9	136	4.20
10	376	4.49
11	736	5.44
12	1848	5.97
13	432	6.70
14	328	7.28
15	248	8.03
16	32	6.47
Total	**4360**	**5.92**


E(Salarios|Educación)
10
8
6
4
2
0
U$S por hora
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Años de Educación
Salario medio
Modelo pol orden 2
Modelo lineal

Pero pueden haber otros factores que influyen en el salario, por ejemplo la experiencia

Años experiencia	No. Observaciones	Salario promedio
1	57	3.69
2	268	4.58
3	376	5.08
4	464	5.39
5	511	5.79
6	523	6.00
7	531	6.28
8	533	6.60
9	485	6.86
10	275	6.40
11	169	6.14
12	81	5.69
13	34	4.58
14	22	5.35
15	14	4.85
16	10	4.81
17	3	4.25
18	2	5.29
Total	4360	5.92


E(Salarios|Experiencia Laboral)
U$S por hora
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
Años de Experiencia Laboral
Salario medio
Modelo pol orden 2
Modelo lineal

- El problema de mirar la relación de los salarios con la educación y la experiencia separadamente es que (si estas dos variables están correlacionadas) al omitir por ejemplo la experiencia la relación encontrada entre salarios y educación estará también capturando el efecto de la experiencia y por lo tanto lo que medimos no será el efecto causal de la educación.
- En este caso el problema se resuelve fácilmente recurriendo al modelo de regresión múltiple en el que podemos considerar ambos factores simultáneamente.

En el siguiente cuadro se observan los resultados de una regresión de los salarios por hora sobre la educación y la experiencia, se prueban distintas especificaciones

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Educación		0.48	-0.42			0.65	-0.25
Educ. al cuadrado			0.04				0.04
Experiencia				0.16	0.78	0.30	0.67
Exp.al cuadrado					-0.04		-0.03
Constante	5.92	0.22	5.10	4.85	3.07	-3.69	0.50

	Media
Educación	11.77
Educ. al cuadrado	141.51
Experiencia	6.51
Exp.al cuadrado	50.42

Algunas observaciones:

- La constante en la "regresión" sin regresores corresponde a la media no condicional de los salarios, para recuperar la media no condicional en las otras especificaciones es necesario multiplicar los coeficientes por las medias de los regresores y sumar.
- En la regresión (2) el efecto parcial asociado a la educación indicaría que por cada año adicional de educación el salario promedio se incrementa en 0.48 dólares por hora
- En la regresión (6) en cambio el efecto parcial es 0.65, es decir un 0.17 más que en la (2): ello me indica que el rendimiento de la educación es mayor CONTROLANDO POR EXPERIENCIA. Ello se debe a que, en la muestra utilizada:
  - la experiencia tiene un efecto positivo en el salario
  - educaciòn y experiencia están negativamente correlacionados

Esperanza condicional

Entonces analizamos el salario por hora en función de los años de educación: en el ejemplo anterior consideramos unos pocos valores, podriamos considerar el caso de 3,4, 5, 6,7.....20 años de educación, pero tratar con cada valor particular de X se vuelve rápidamente engorroso (en particular cuando tengo más de una X), a veces difícil de interpretar y en algunos casos no cuento con suficiente información en cada "celda".
Recurrimos entonces a un modelo paramétrico.
Consideramos una función general:

$$\mathbf{E}\left(Y \mid X = x\right) = \mu(x)$$

que podemos parametrizar, por ejemplo,

$$\mu(x) = a + bx$$

En el caso de la educación, por ejemplo:

$$\mathbf{E}\left(Salario \mid EDUC = educ\right) = 0.22 + 0.48 educ$$

si agregamos la experiencia

$$\mathbf{E}(Salario|EDUC = educ, EXPER = \text{exper}) = -3.96 + 0.65 educ + 0.30 \text{exper}$$

¿Es lo anterior un análisis ceteris paribus?

- Estamos realizando un análisis CONTROLANDO POR EXPERIENCIA es decir es válido razonar que es el efecto de la educación dejando constante la experiencia, así que en cierta medida puedo pensarlo con un análisis ceteris paribus.
- Sin embargo, podríamos estar omitiendo otros factores que afecten los salarios y tengan correlación con educación o experiencia por lo cual 0.65 no es necesariamente el efecto causal de la educación. Ejemplos de variables omitidas en este caso:
  - Sexo: relativamente fácil de solucionar porque suele observarse en todas las encuestas de salarios
  - Habilidad del individuo: muy difícil de solucionar

Nota: aqui mostramos **estimaciones puntuales** para ilustrar la idea de la relación entre los efectos causales, las esperanzas condicionales y el modelo de regresión, para realizar el análisis en es necesario también tener en cuenta la varianza de los estimadores.

Modelo paramétrico

Lo usual es que no conozcamos la forma de $E(y|\mathbf{x})$, por lo que generalmente se establece que esa esperanza condicional depende de un conjunto finito de parámetros, lo que determina la especificación de un MODELO PARAMETRICO de $E(y|\mathbf{x})$.

Por ejemplo, para el caso de dos variables explicativas, podrìamos considerar las siguientes especificaciones para la media condicional

$$\begin{aligned}
E(y|x_1, x_2) &= E^*(y|x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \\
E(y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 \\
E(y|x_1, x_2) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 \\
E(y|x_1, x_2) &= \exp\left[\theta_0 + \theta_1 \log(x_1) + \theta_2 x_2\right]
\end{aligned}$$

Función de Esperanzas Condicional (FEC)

Definición:

- Sea y una variable aleatoria, a la que nos referiremos como la *variable explicada, o regresando, o variable dependiente, o variable endógena*
- Sea $\mathbf{x} = (x_1, x_2, ..., x_k)$ un vector de variables aleatorias de dimensión $1 \times K$ a las que nos referiremos como *variables explicativas, regresores o covariables.*

Si $\mathbb{E}(|y|) < \infty$, entonces existe una función, digamos $\mu : R^K \rightarrow R$, tal que,

$$\begin{aligned} \mathbb{E}(y \mid x_1, x_2, ..., x_k) &= \mu(x_1, x_2, ..., x_k) \\ &\text{o} \\ \mathbb{E}(y \mid \mathbf{x}) &= \mu(\mathbf{x}) \end{aligned}$$

Esperanza condicional

La idea es entonces, suponiendo que se conociera el valor de X, cual sería el valor esperado o medio de Y. Para obtener la esperanza condicionales hay que *integrar respecto a las función de probabilidad condicional*:
Si *Y es discreta* que puede asumir m valores $\{y_1, y_2; ....y_m\}$

$$\mathbf{E}(y \mid x = \mathbf{x}) = \sum_{i=1}^{m} y_i f_{y|x}(y_i \mid x = \mathbf{x}) = \sum_{i=1}^{m} y_i \Pr(y = y_i \mid x = \mathbf{x})$$

En el caso de *Y continua*

$$\mathbf{E}(y \mid x = \mathbf{x}) = \int_{-\infty}^{+\infty} y_i f_{Y|X}(y_i \mid x = \mathbf{x})\, dy$$

Algunas propiedades de la FEC

A. *Operador lineal:* Si $a_1(x), ..., a_G(x)$ y $b(x)$ son escalares que dependen de x y sean y_1, $y_2 ... y_G$ variables aleatorias, entonces se cumple

$$\mathbb{E}\left(\sum_{j=1}^{G} a_j(x) y_j + b(x) \,|x\right) = \sum_{j=1}^{G} \left[a_j(x)\mathbb{E}\left(y_j \,|x\right) + b(x)\right]$$

siempre que las esperanzas $\mathbb{E}\left(|y_j|\right) < \infty$, $\mathbb{E}\left(|a_j(x)y_j|\right) < \infty$ y $\mathbb{E}\left(|b(x)|\right) < \infty$.

Algunas propiedades de la FEC (cont.)

B. *Ley de las Expectativas Iteradas (LEI):*
$\mathbb{E}(y) = \mathbb{E}(\mathbb{E}(y|x)) = \mathbb{E}[\mu(x)]$

También se cumple por LEI $\mathbb{E}(\mathbb{E}(y|x,z)|x) = \mathbb{E}(y|x)$

.

Algunas propiedades de la FEC (cont.)

C. *Perturbaciones:* Si $u \equiv y - \mathbb{E}(y|x)$ entonces $\mathbb{E}(g(x)u) = 0$ para cualquier función $g(x)$ siempre que $\mathbb{E}(|g(x)u|) < \infty$ y $\mathbb{E}(|u|) < \infty$.

Varianza condicional

Definición de varianza condicional:

$$Var\left(y\left|x\right.\right)=\sigma^{2}\left(x\right)=\mathbb{E}\left[\left\{y-\mathbb{E}\left(y\left|x\right.\right)\right\}^{2}\left|x\right.\right]=\mathbb{E}\left(y^{2}\left|x\right.\right)-\left[\mathbb{E}\left(y\left|x\right.\right)\right]^{2}$$

Algunas propiedades de la Varianza condicional

A. Si $a(x)$ y $b(x)$ son escalares que dependen de x:

$$Var\left(a(x)y + b(x)\,|x\right) = a(x)^2\, Var\left(y\,|x\right)$$

Algunas propiedades de la Varianza condicional

B. Descomposición de la varianza

$$\begin{aligned} Var(y) &= \mathbb{E}[Var(y|x)] + Var[\mathbb{E}(y|x)] \\ &= \mathbb{E}(\sigma^2(x)) + Var[\mu(x)] \end{aligned}$$

$$Var(y|x) = \mathbb{E}[Var(y|x,z)|x] + Var[\mathbb{E}(y|x,z)|x]$$

C. $\mathbb{E}[Var(y|x)] \geq \mathbb{E}[Var(y|x,z)]$

Independencia

Veremos tres conceptos de independencia:

- Independencia estocástica
- Independencia en media
- Ausencia de correlación

Independencia estocástica

Se dice que dos variables son estocásticamente independiente si se cumple que la función de probabilidad conjunta es igual al producto de las marginales

$$\begin{aligned} x, y \text{ continuas } f_{x,y}(x,y) &= f_x(x) f_y(y) \\ x, y \text{ discretas } \Pr(x &= \mathbf{x}, y = \mathbf{y}) = \Pr(x = \mathbf{x}) \Pr(y = \mathbf{y}) \end{aligned}$$

Si dos variables aleatorias x e y son estocásticamente independientes entonces

- conocer el resultado de x no altera la probabilidad de los posibles resultados de y y viceversa (es una propiedad simétrica)
- $f(y \mid x) = f(y)$ y $f(x \mid y) = f(x)$
- $E(y \mid x) = E(y)$ y $E(x \mid y) = E(x)$
- $E(yx) = E(y)E(x)$
- $cov(y, x) = 0$
- Todos los momentos condicionales de $y \mid x$ y $x \mid y$ de coinciden con los momentos no condicionales

Independencia en media

Se dice que y es independiente en media de x si se cumple que la esperanza condicional de $y \mid x$ es igual a la esperanza incondicional de y

$$E(y \mid x) = E(y)$$

Este concepto no es necesariamente simétrico

$$E(y \mid x) = E(y) \text{ no implica } E(x \mid y) = E(x)$$

Si y es independiente en media de x se cumple:

- $\mathbf{cov(y, x) = 0}$
- $\mathbf{cov(y, g(x)) = 0}$ para toda función continua $g(.)$
  Si y es independiente en media de x y se cumple $E(y \mid x) = 0$
- $\mathbf{E(y) = 0}$
- $\mathbf{E(yg(x)) = 0}$
- $\mathbf{cov(y, g(x)) = 0}$

Ausencia de correlación

Se dice que y y x están incorrelacionados si se cumple que la covarianza entre y y x es igual a 0

$$cov(y, x) = 0$$

Este concepto es simétrico

Si y y x están incorrelacionados se cumple necesariamente:

- $\mathbf{cov}(\mathbf{y}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$ sólo si $g(x) = a + bx$ (lineal)

Modelo de Regresión con Regresores estocásticos

Tanto la insesgadez como la consistencia del estimador MCO en el modelo de regresión con regresores estocásticos

$$y_i = x_i^{'}\beta + u_i$$

depende del supuesto

$$E(u_i \,|X\,) = 0$$

La referencia a este supuesto en la literatura:

- independencia entre el error y los regresores
- ortogonalidad entre el error y los regresores
- exogeneidad de los regresores
- errores no correlacionados con las covariables

Recordemos que en el contexto del modelo de regresión simple con regresores estocásticos,

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

el estimador MCO de β puede ser interpretado en términos de momentos poblacionales de la siguiente forma:

$$\begin{aligned} p\lim \widehat{\beta}_{MCO} &= \frac{cov(x_i, y_i)}{var(x_i)} \\ &= \frac{cov(x_i, \alpha + \beta x_i + u_i)}{var(x_i)} \\ &= \beta + \frac{cov(x_i, u_i)}{var(x_i)} \end{aligned}$$

por lo cual será inconsistente si el regresor y el error están correlacionados

En términos más generales si se cumple la independencia en media del error respecto a los regresores,

$$E(u_i \,|X\,) = 0 \Rightarrow$$

$$E(u_i) = 0$$

$$E(x_i u_i) = 0 \Rightarrow \begin{cases} E(x_{1i} u_i) = 0 \\ E(x_{2i} u_i) = 0 \\ \vdots \\ E(x_{ki} u_i) = 0 \end{cases}$$

$E(u_i \,|X\,) = 0$ también implica $cov(u_i, h(x_i)) = 0$ donde $h(.)$ es una función continua, por ejemplo:

$$\begin{aligned} cov(u_i, x_i) &= 0 \\ cov(u_i, x_i^2) &= 0 \\ cov(u_i, \ln(x_i)) &= 0 \text{ si } x_i > 0 \end{aligned}$$

.

Método de los momentos

Momentos poblacionales

$$E(u_i \,|X\,) = 0 \Rightarrow$$

$$\begin{aligned}
E(u_i) &= E(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki}) = 0 \\
E(x_{1i} u_i) &= E\left[x_{1i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0 \\
E(x_{2i} u_i) &= E\left[x_{2i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0 \\
&\vdots \\
E(x_{ki} u_i) &= E\left[x_{ki}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0
\end{aligned}$$

Si se cuenta con una muestra iid $\{y_i, x_i\}_{i=1,...N}$es posible utilizar el

método de los momentos (el principio de analogía) para estimar los coeficientes del modelo:

Análogo muestral

$$E(u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$E(x_{1i}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} x_{1i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$E(x_{2i}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} x_{2i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$\vdots$$

$$E(x_{ki}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} x_{ki}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

Notar que estas condiciones son idénticas a las que dan lugar a los estimadores MCO y MV en el MRLG.

Es muy importante tener en cuenta que el cumplimiento del. supuesto $E(u_i \mid x) = 0$ no dependerá en general de cuestiones estadísticas sino que dependerá de

- cuáles son las relaciones económicas que esperamos entre las variables del modelo econométrico
- los datos disponibles

Modelos de regresión lineal con regresores endógenos (no independientes)

En econometría no se mantiene el uso tradicional de la palabra "endógeno" que deviene de los modelos económicos, en los cuales este término designa una variable que se determina en el contexto del modelo (no viene "dada"). En econometría, una variable explicativa/regresor/covariable x_j se dice que es

- **endógena** cuando se encuentra correlacionada con el error de la ecuación a estimar,
- **exógena** si está incorrelacionada.con el error de la ecuación a estimar

Endogeneidad

En econometría la endogeneidad surge de tres fuentes:

- la existencia de variables omitidas,
- la presencia de errores de medida
- la simultaneidad en la determinación de alguna variable explicativa x_j y la explicada y (sí relacionado con el concepto económico de endogeneidad).

La estimación consistente en presencia de correlación entre los regresores y el error será posible en algunos casos, un caso es cuando se dispone de INSTRUMENTOS VALIDOS.

Omisión de variables relevantes, inclusión de variables irrelevantes, multicolinealidad (Wooldridge 3.3, 3.4 y Apéndice 3A)

Omisión de variables relevantes

Consideremos el modelo

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad i = 1, ...n$$

$$E\left(u_i \mid x_1, x_2\right) = 0$$

Supongamos que observamos y y x_1 pero no x_2. El modelo que puedo estimar es

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \varepsilon_i \\ \varepsilon_i &= \beta_2 x_{2i} + u_i \end{aligned}$$

En términos de límite de probabilidad:

$$p\lim \widehat{\beta}_{1MCO} = \frac{cov(x_{1i}, y_i)}{var(x_{1i})}$$

como el verdadero modelo para y_i es $x_{1i}\beta_1 + x_{2i}\beta_2 + u_i$ tenemos

$$\begin{aligned} &= \frac{cov(x_{1i}, x_{1i}\beta_1 + x_{2i}\beta_2 + u_i)}{var(x_{1i})} \\ &= \beta_1 + \beta_2 \frac{cov(x_{1i}, x_{2i})}{var(x_{1i})} \end{aligned}$$

Entonces, el estimador será inconsistente a menos que $cov(x_{1i}, x_{2i}) = 0$.y/o $\beta_2 = 0$

En términos de esperanzas:
En el MRLS

$$\widehat{\beta}_{1MCO} = \frac{\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)y_i}{\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)^2}$$

Notar: esto equivale al análogo muestral de $\frac{cov(x_{1i},y_i)}{var(x_{1i})}$.

El numerador lo desarrollamos como

$$\sum_{i=1}^{N}(x_{1i}-\overline{x}_1)y_i = \sum_{i=1}^{N}(x_{1i}-\overline{x}_1)\left(x_{1i}\beta_1 + x_{2i}\beta_2 + u_i\right)$$

$$= \beta_1 \sum_{i=1}^{N}(x_{1i}-\overline{x}_1)^2 + \beta_2 \sum_{i=1}^{N}(x_{1i}-\overline{x}_1)x_{2i} + \sum_{i=1}^{N}(x_{1i}-\overline{x}_1)u_i$$

por lo cual:

$$\widehat{\beta}_{1MCO} = \beta_1 + \beta_2 \frac{\sum_{i=1}^{N}(x_{1i}-\overline{x}_1)x_{2i}}{\sum_{i=1}^{N}(x_{1i}-\overline{x}_1)^2} + \frac{\sum_{i=1}^{N}(x_{1i}-\overline{x}_1)u_i}{\sum_{i=1}^{N}(x_{1i}-\overline{x}_1)^2}$$

$$
\begin{aligned}
E\left(\widehat{\beta}_{1MCO} \mid x1, x2\right) &= \beta_1 + \beta_2 \frac{\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)x_{2i}}{\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)^2} \\
&= \beta_1 + \beta_2 * \widehat{\delta}
\end{aligned}
$$

donde $\widehat{\delta}$ es el coeficiente asociado a $x1$ en una regresión de $x1$ sobre $x2$
Por la Ley de las Expectativas iteradas podemos escribir:

$$
\begin{aligned}
E\left(\widehat{\beta}_{1MCO}\right) &= E\left[E\left(\widehat{\beta}_{1MCO} \mid x1, x2\right)\right] = \beta_1 + \beta_2 * E(\widehat{\delta}) \\
&= \beta_1 + \beta_2 * \delta
\end{aligned}
$$

entonces, vemos que $\widehat{\beta}_{1MCO}$ es sesgado excepto que $\beta_2 = 0$ y/o $\delta = 0$.

El sesgo por omisión de variable relevante, entonces depende de β_2 y δ, el signo del sesgo estará dado por

	$corr(x1, x2) > 0$	$corr(x1, x2) < 0$
$\beta_2 > 0$	sesgo positivo	sesgo negativo
$\beta_2 < 0$	sesgo negativo	sesgo positivo

Omisión de variables relevantes

Otra forma de verlo es:

$$\begin{aligned} cov(x_i, \varepsilon_i) &= cov(x_{1ii}, x_{2i}\beta_2 + u_i) \\ &= \beta_2 \frac{cov(x_{1i}, x_{2i})}{var(x_{1i})} \end{aligned}$$

se viola el supuesto de ausencia de correlación entre los regresores y el término de error

La conclusión entonces es que si la variable omitida está correlacionada con algún regresor el estimador MCO será sesgado e inconsistente.

Omisión de variables relevantes

Otro aspecto que es necesario analizar es la varianza del estimador. Notar que

$$Var(\widehat{\beta}_{1MCO}) = \sigma^2 \left(X_1' X_1 \right)^{-1}$$

Mientras que si hubieramos estimado el modelo correcto (aquel que incluye también x_2) la varianza del estimador estaría dada por el bloque superior izquierdo de la matriz $\left(X'X \right)^{-1}$ el cual utilizando las propiedades de la matriz particionada

$$\begin{pmatrix} X_1' X_1 & X_1' X_2 \\ X_2' X_1 & X_2' X_2 \end{pmatrix}^{-1}$$

Omisión de variables relevantes

$$(X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1} = \left(X_1'M_2X_1\right)^{-1}$$

$$Var(\widehat{\beta}_{12MCO}) = \sigma^2 \left(X_1'M_2X_1\right)^{-1}$$

$$Var(\widehat{\beta}_{1MCO})^{-1} - Var(\widehat{\beta}_{12MCO})^{-1} = \frac{1}{\sigma^2}X_1'X_2(X_2'X_2)^{-1}X_2'X_1$$

que es una matriz definida positiva, por lo cual la varianza del estimador en el modelo que omite X_2 es menor que la que incluye todas las variables. Notar que este resultado es independiente del hecho que las variables estén correlacionadas.

Omisión de variables relevantes en el MRLM

Consideremos ahora el caso del modelo de regresión lineal múltiple (se incluyen 3 variables, 2 observables y una inobservable):

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad i = 1, ...n$$

$$E\left(u_i \mid x_1, x_{2,}, x_3\right) = 0$$

$$plim\ \widehat{\beta}_{MCO} = \left[Var(x)\right]^{-1} cov(x, y) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$= \begin{bmatrix} Var(x1) & cov(x1, x2) & cov(x1, x3) \\ & Var(x2) & cov(x2, x3) \\ & & Var(x3) \end{bmatrix}^{-1} \begin{bmatrix} cov(x1, y) \\ cov(x2, y) \\ cov(x3, y) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{12} \\ & & \sigma_3^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{1Y} \\ \sigma_{2Y} \\ \sigma_{3Y} \end{bmatrix}$$

Si no disponemos de mediciones de $x3$ (o la omitimos por error) estimaremos un modelo

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad i = 1, ...n \\ \varepsilon_i &= \beta_3 x_{3i} + u_i \end{aligned}$$

en el cual el supuesto $E(u_i \mid x_1, x_2, x_3) = 0$ no asegura $E(\varepsilon_i \mid x_1, x_2) = 0$
En este modelo

$$plim \begin{pmatrix} \widetilde{\beta}_{1MCO} \\ \widetilde{\beta}_{2MCO} \end{pmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{1Y} \\ \sigma_{2Y} \end{bmatrix}$$

podemos re-escribir

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \sigma_{1Y} \\ \sigma_{2Y} \\ \sigma_{3Y} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} Var(x) * [Var(x)]^{-1} \begin{bmatrix} \sigma_{1Y} \\ \sigma_{2Y} \\ \sigma_{3Y} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} Var(x) * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

operando obtenemos

$$plim \begin{pmatrix} \widetilde{\beta}_{1MCO} \\ \widetilde{\beta}_{2MCO} \end{pmatrix} = \begin{bmatrix} 1 & 0 & \frac{\sigma_2^2\sigma_{13}-\sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} \\ 0 & 1 & \frac{\sigma_1^2\sigma_{23}-\sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$= \begin{matrix} \beta_1 + \beta_3 \left(\frac{\sigma_2^2\sigma_{13}-\sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} \right) \\ \beta_2 + \beta_3 \left(\frac{\sigma_1^2\sigma_{23}-\sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} \right) \end{matrix}$$

Multicolinealidad

Definición de multicolinealidad: dos o más regresores incluidos en el modelo están (altamente) correlacionados.

- Multicolinealidad perfecta: al menos uno de los regresores puede escribirse como una combinación lineal de otro(s) regresor(es). Aquí se viola el supuesto de rango(x)=K con probabilidad 1, el modelo no se puede calcular
- Multicolinealidad: La correlación es alta pero no perfecta (a veces se establece un valor de 0.9 pero este número es sólo una guía).

Multicolinealidad

- Cómo detectar multicolinealidad
  - una señal de multicolinealidad surge cuando al estimar el modelo de interés se observa un R2 alto con coeficientes individualmente no significativos
  - también es aconsejable analizar la matriz de correlaciones de los regresores o hacer regresiones de unos regresores sobre otros antes de estimar el modelo de interés
- Consecuencias de la multicolinealidad:
  - Si todos los regresores son exógenos el estimador MCO de los coeficientes sigue siendo consistente e insesgado
  - La varianza del estimador $\widehat{\beta}_j$ se "infla", no es aconsejable realizar inferencia sobre cada coeficiente aisladamente.
    $Var(\widehat{\beta}_j) = \frac{\sigma^2}{Var(x_j)\left(1-R_j^2\right)}$ donde R_j^2 es el R2 de una regresión de x_j sobre $x_1, x_2, \ldots x_{j-1}, x_{j+1}, \ldots x_k$

Multicolinealidad y omisión de variables relevantes

- Intentar corregir el "problema" de la multicolinealidad omitiendo una variable muy correlacionada con otro regresor puede llevar a un mal mayor: omisión de variable relevante que afecta la consistencia e insesgadez del estimador
- Lo que se aconseja es realizar contrastes de significación conjunta de los regresores correlacionados
- El tamaño de la muestra también juega un rol importante en este caso

Inclusión de variables irrelevantes

La inclusión de variables irrelevantes:

- no afectará la insesgadez ni la consistencia del estimador de los coeficientes del modelo
- aumentará la varianza (precisión) del estimador

Errores de medida (Wooldridge: 9.2 y 9.3)

Una dificultad que subyace a casi todo trabajo empírico en Economía es la imposibilidad de disponer de observaciones muestrales de las variables que se pretende relacionar.Por ejemplo,

- las variables de Contabilidad Nacional como el PBI, el consumo, no son sino estimaciones de conceptos teóricos que no se observan en la realidad
- la Renta Permanente, la inteligencia o la habilidad de un trabajador, no disponemos ni siquiera de estimaciones, por lo que, en el mejor de los casos suelen utilizarse variables aproximadas, (variables proxy)
- hay variables que se recolectan en encuestas pero estas mediciones presentan errores (ingresos del hogar, educación de los individuos, etc.)

Cabe esperar que cuando hay error en la medición de las variables, las estimaciones no mantendrán las propiedades ideales que se satisfacen bajo los supuestos del MRLG. Recordemos que estas propiedades son la ausencia de sesgo, la consistencia, la eficiencia y la normalidad asintótica.

Errores de medida en la variable dependiente

Veamos que sucede si sólo puedo observar la variable dependiente sujeta a un error de medida:
Es decir en el modelo:

$$\begin{aligned} y_i^* &= \beta x_i + u_i \quad i = 1, 2, ...N \\ E(u_i &\mid x_i) = 0 \\ V(u_i &\mid x_i) = \sigma_u^2 \end{aligned}$$

observamos

$$\begin{aligned} & \{y_i, x_i\}_{i=1,2....N} \\ y_i &= y_i^* + v_i \end{aligned}$$

suponemos que v_i se distribuye con media 0 y varianza σ_v^2, .independiente de u_i y x_i

Errores de medida en la variable dependiente

Para hacer nuestro modelo estimable tenemos que expresarlo en términos de variables observables

$$\begin{aligned} y_i &= \beta x_i + (u_i + v_i) \\ &= \beta x_i + \varepsilon_i \\ E(\varepsilon_i \mid x_i) &= 0 \\ V(\varepsilon_i \mid x_i) &= \sigma^2_\varepsilon = \sigma^2_u + \sigma^2_v \end{aligned}$$

Por consiguiente, el suponer que existe errores de medida en la variable endógena no afectará la insesgadez y consistencia del estimador MCO, aunque si afectará la varianza del error y por tanto de los estimadores (aumenta en relación a un modelo donde no hay errores de medida).

Errores de medida en variables independientes (regresores, covariables)

La situación es diferente cuando la variable que se mide con error es la variable independiente (o regresor).
Supongamos

$$\begin{aligned} y_i &= \beta x_i^* + u_i \quad i = 1, 2, ...N \\ E(u_i \mid x_i^*) &= 0 \\ V(u_i \mid x_i^*) &= \sigma_u^2 \end{aligned}$$

observamos

$$\begin{aligned} & \{y_i, x_i\}_{i=1,2....N} \\ x_i &= x_i^* + v_i \end{aligned}$$

Suponemos que v_i está incorrelacionado tanto con x_i^* como con u_i.

Errores de medida en variables independientes (regresores, covariables)

Entonces el modelo, expresado en términos de las variables observables será

$$\begin{aligned} y_i &= \beta x_i + (u_i - \beta v_i) \\ &= \beta x_i + \varepsilon_i \end{aligned}$$

El principal problema que enfrentamos es que ya no podemos suponer $E(\varepsilon_i \mid x_i) = 0$, debido a que

$$\begin{aligned} cov(x_i, \varepsilon_i) &= cov(x_i^* + v_i, u_i - \beta v_i) \\ &= cov(x_i^*, u_i) - \beta cov(x_i^*, v_i) + cov(v_i, u_i) - \beta \sigma_v^2 \\ &= -\beta \sigma_v^2 \end{aligned}$$

Errores de medida en variables independientes (regresores, covariables)

Asintóticamente

$$
\begin{aligned}
p\lim \widehat{\beta}_{MCO} &= \frac{cov(x_i, y_i)}{var(x_i)} = \frac{cov(x_i^* + v_i, \beta x_i^* + u_i)}{var(x_i^* + v_i)} \\
&= \beta \frac{var(x_i^*)}{var(x_i^*) + var(v_i)} \\
&= \beta \frac{1}{1 + \frac{\sigma_v^2}{\sigma_x^2}} = \beta \frac{1}{1 + \lambda}
\end{aligned}
$$

Entonces $\widehat{\beta}_{MCO}$ es inconsistente, excepto que $\sigma_v^2 = 0$, o sea en el caso que no existe error de medida (o el mismo sea constante).

Notas:

1) El sesgo es hacia 0 $\left| p\lim \widehat{\beta}_{MCO} \right| < |\beta|$ por ello el sesgo por error de medida se conoce también con el nombre de "sesgo de atenuación"

2) El sesgo es mayor cuando mayor sea el ratio $\frac{\sigma_v^2}{\sigma_x^2}$

Simultaneidad (Wooldridge, 16.2 Hayashi, 3.1)

La simultaneidad surge cuando al menos una de las variables explicativas es determinada simultáneamente con y. Si, digamos x_j es determinada parcialmente como una función de y, entonces x_j y u están en general correlacionadas.

Por ejemplo, si y es la tasa de criminalidad en una ciudad y x_j el tamaño de la fuerza de policía, este último puede ser función de la tasa de criminalidad. Ésta es una situación conceptualmente difícil de analizar, ya que debemos pensar en una situación en que x_j puede variar exógenamente, aún cuando en los datos que observamos y y x_j son generadas simultáneamente.

Un modelo simple de oferta y demanda ilustra el caso:

$$q_i^d = \alpha_0 + \alpha_1 p_i + u_i \qquad (demanda)$$

$$q_i^s = \beta_0 + \beta_1 p_i + v_i \qquad (oferta)$$

Los términos u_i y v_i desplazan cada una de las curvas as representan el conjunto de otras influencias además del precio en cada curva. Suponemos que $E(u_i) = E(v_i) = 0$ así como el equilibrio del mercado:

$$q_i^s = q_i^d$$

(una tercera ecuación que reduce el sistema a dos ecuaciones).

$$q_i = \alpha_0 + \alpha_1 p_i + u_i \qquad (demanda)$$

$$q_i = \beta_0 + \beta_1 p_i + v_i \qquad (oferta)$$

La idea de endogeneidad es que un regresor no satisface la condición de ortogonalidad con el término de error (si existe una constante en el modelo, cuando está correlacionado con el término de error). En este caso, en ambas ecuaciones p_i está correlacionado con el término de error.

Resolviendo para q_i, p_i se obtiene:

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}$$

$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1}$$

El regresor p_i es una función de ambos términos de error. $Cov\left(p_i, v_i\right) \neq 0$ y $Cov\left(p_i, u_i\right) \neq 0$. En este ejemplo la endogeneidad resulta del equilibrio de mercado.

Variables instrumentales (Wooldridge, 15.1 a 15.6)

Variables instrumentales y variables aproximadas (en el caso de omisión una variable inobservable)

Para el caso de variables relevantes omitidas dos soluciones pueden llegar a ser válidas: utilizar variables proxy o utilizar variables instrumentales. Es importante notar que la naturaleza de ambas variables es muy diferente

- En el caso de variables aproximadas se sustituye la variable inobservable por la variable PROXY que deseablemente debe estar MUY CORRELACIONADA CON LA VARIABLE INOBSERVABLE. Cuando se cuenta con una variable PROXY se sustituye la variable inobservable por ésta y se estima el modelo por MCO.
- En el caso del método de las variables instrumentales la variable inobservable pasa a formar parte del error y necesitamos instrumento que cumpla con varias condiciones (que veremos más adelante) una condición es que debe estar INCORRELACIONADA CON LA VARIABLE INOBSERVABLE.

Variables instrumentales

En el caso de modelos estimados por el método VI se observan tres tipos de variables (además de la dependiente):

- regresor(es) no correlacionado(s) con el error
- regresor(es) correlacionado(s) con el error
- el/los instrumentos.

Variables instrumentales

Consideremos un modelo

$$\begin{aligned} y_i &= x_i^{'}\beta + u_i \\ x_i^{'} &= (1, x_{1i}, x_{2i}, ...x_{ki}) \\ cov(x_{ji}, u_i) &= 0 \text{ para algun(os) j} \\ cov(x_{ji}, u_i) &\neq 0 \text{ para otro(s) j (supongamos } j = k) \end{aligned}$$

Tengo una muestra $\{y_i, x_i, w_i\}_{i=1,2....N}$

Un instrumento (z)será una variable que no está incluida en el modelo original (por ejemplo pertenece a w) y que cumple con las siguientes condiciones:

- no está correlacionada con el término de error del modelo de regresión que quiero estimar (esto, en general, no podrá ser contrastado empíricamente)
- está correlacionada con la variable independiente que presenta el problema de endogeneidad (si se puede contrastar empíricamente)

Supongamos ahora que tenemos

$$\begin{aligned} y_i &= \beta x_i^* + u_i \quad i = 1, 2, ...N \\ E(u_i &\mid x_i^*) = 0 \\ V(u_i &\mid x_i^*) = \sigma_u^2 \end{aligned}$$

Pero sólo puedo observar x_i sujeto a error de medida,

$$x_i = x_i^* + v_i$$

Suponemos que v_i (el error de medida) está incorrelacionado tanto con x_i^* como con u_i. Entonces el modelo, expresado en términos de las variables observables será

$$\begin{aligned} y_i &= \beta x_i + (u_i - \beta v_i) \\ &= \beta x_i + \varepsilon_i \end{aligned}$$

Variables instrumentales

Ejemplo de un instrumento: tengo una segunda medición de la variable en cuestión, también con error, pero que el error de medida de la primera y la segunda medición son independientes. Supongamos entonces que

$$z_i = x_i^* + h_i$$

con $cov(h_i, x_i^*) = cov(h_i, v_i) = cov(h_i, u_i) = 0$.
El estimador por variables instrumentales será:

$$\begin{aligned} p\lim \widehat{\beta}_{IV} &= \frac{cov(z_i, y_i)}{cov(z_i, x_i)} \\ &= \frac{cov\left[x_i^* + h_i, \beta x_i^* + u_i\right]}{cov(x_i^* + h_i, x_i^* + v_i)} \\ &= \beta \end{aligned}$$

El método de variables instrumentales (VI) proporciona una solución general al problema de la presencia de una variable explicativa endógena.

Para usar el método de VI necesitamos de una variable adicional z, que no esté presente en la ecuación original y que satisfaga dos condiciones básicas. En primer lugar deberá estar incorrelacionada con el error de la ecuaciòn a estimar:

$$Cov\left(z,u\right)=0$$

En otras palabras, como x_1, x_2,..., x_{K-1}, z deberá ser exógena en la ecuación a estimar

El segundo requerimiento se refiere a la relación entre z y la variable endógena x_K. La manera precisa de establecerlo se refiere a la regresión de x_K en **todas** las variables exógenas (esta regresión suele denominarse "modelo de forma reducida":

$$x_k = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + ... + \pi_{k-1} x_{k-1} + z'\theta + v_k$$

en que por definición $E\left(u_K \mid x_{1,....} x_{k-1}, z\right) = 0$. Lo que se requiere es

$$\begin{aligned} \text{Rechazar Ho:} \quad & \theta = 0 \\ \text{No rechazar H1:} \quad & \theta \neq 0 \end{aligned}$$

Este requerimiento se suele enunciar en forma laxa diciendo "z debe estar correlacionada con x_K". En realidad el enunciado $\theta \neq 0$ está indicando algo más, es decir que z está correlacionada con x_K una vez que la influencia de x_1, x_2,..., x_{K-1}, ha sido tomada en cuenta.

No se pone restricciones sobre la distribución de z y x_K. Pueden ser ambas continuas, o una o ambas ser variables discretas.

Cuando z cumple con ambas condiciones, se dice que es una candidata a variable instrumental (o instrumento) para x_K.

Las variables x_1, x_2,..., x_{K-1}, se suponen incorrelacionadas con u, y en ese sentido son sus propios instrumentos en la ecuación.

La lista completa de instrumentos en realidad es la lista de todas las variables exógenas, aunque usualmente se hace referencia al instrumento para la variable endógena.

Variables instrumentales (caso exactamente identificado)

En modelo de regresión múltiple, tenemos

- un vector de regresores X de dimensión **k**,
- un vector de instrumentos Z de dimensión **r=k**, entonces estamos en un caso exactamente identificado

$$\begin{aligned} \widehat{\beta}_{IV} &= \left[Z'X\right]^{-1} Z'Y \\ V(\widehat{\beta}_{IV}) &= \sigma^2 \left[Z'X\right]^{-1} Z'Z \left[X'Z\right]^{-1} \end{aligned}$$

Es importante notar que si tengo menos instrumentos **(r)** que regresores **(k)** no podré realizar la estimación.

Si el único regresor endógeno es x_k y hay un único instrumento z_k entonces

$$Z = \begin{bmatrix} 1 & z_{11} & \cdots & z_{k-1,1} & z_{k,1} \\ 1 & z_{12} & \cdots & z_{k-1,2} & z_{k,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & z_{1N} & \cdots & z_{k-1,N} & z_{k,N} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k-1,1} & z_{k,1} \\ 1 & x_{12} & \cdots & x_{k-1,2} & z_{k,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1N} & \cdots & x_{k-1,N} & z_{k,N} \end{bmatrix}$$

es decir algunas columnas de la matriz Z pueden coincidir con columnas de la matrix X, esto se da para las x exógenas (a veces se dice que son sus propios instrumentos).

Método de los momentos

$$E(u_i \,|Z\,) = 0 \Rightarrow$$

$$\begin{aligned}
E(u_i) &= E(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki}) = 0 \\
E(x_{1i} u_i) &= E\left[x_{1i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0 \\
E(x_{2i} u_i) &= E\left[x_{2i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0 \\
&\vdots \\
E(z_{ki} u_i) &= E\left[z_{ki}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0
\end{aligned}$$

Si se cuenta con una muestra iid $\{y_i, x_i, z_i\}_{i=1,...N}$es posible utilizar el principio de analogía:

Análogo muestral

$$E(u_i) = 0 \overset{\text{ppio analogía}}{\rightarrow} \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$E(x_{1i}u_i) = 0 \overset{\text{ppio analogía}}{\rightarrow} \frac{1}{N}\sum_{i=1}^{N} x_{1i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$E(x_{2i}u_i) = 0 \overset{\text{ppio analogía}}{\rightarrow} \frac{1}{N}\sum_{i=1}^{N} x_{2i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

$$\vdots$$

$$E(z_{ki}u_i) = 0 \overset{\text{ppio analogía}}{\rightarrow} \frac{1}{N}\sum_{i=1}^{N} z_{ki}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki}) = 0$$

Variables instrumentales (caso sobreidentificado)

Si tengo más intrumentos que regresores

- un vector de regresores X de dimensión **k**,
- un vector de instrumentos Z de dimensión **r**>**k**, entonces estamos en un caso sobreidentificado

Método de los momentos

Momentos poblacionales

$$E(u_i|Z) = 0 \Rightarrow$$
$$E(u_i) = E(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki}) = 0$$
$$E(z_{1i} u_i) = E\left[z_{1i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0$$
$$E(z_{2i} u_i) = E\left[z_{2i}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0$$
$$\vdots$$
$$E(z_{ri} u_i) = E\left[z_{ri}(y_i - \beta_0 - \beta_1 x_{1i} ..... - \beta_k x_{ki})\right] = 0$$

Si se cuenta con una muestra iid $\{y_i, x_i, z_i\}_{i=1,...N}$es posible utilizar el principio de analogía:

Análogo muestral

$$E(u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki})$$

$$E(z_{1i}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} z_{1i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki})$$

$$E(z_{2i}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} z_{2i}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki})$$

$$\vdots$$

$$E(z_{ri}u_i) = 0 \overset{\text{ppio analogía}}{\longrightarrow} \frac{1}{N}\sum_{i=1}^{N} z_{ri}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} ..... - \widehat{\beta}_k x_{ki})$$

Si $r > k$ tengo más ecuaciones que parámetros: no podré igualar a 0 todas las ecuaciones: usaré MC2E o Método Generalizado de los Momentos

En el caso sobreidentificado la fórmula matricial del estimador VI es,

$$\begin{aligned} \widehat{\beta}_{IV} &= \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'Y \\ V(\widehat{\beta}_{IV}) &= \sigma^2 \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} \end{aligned}$$

Este estimador también puede interpretarse como el estimador de mínimos cuadrados en dos etapas MC2E

Variables instrumentales: Mínimos Cuadrados en 2 Etapas (MC2E)
ETAPA 1: Regresar X sobre Z utilizando MCO y obtener una predicción de X : $\widehat{X} = Z\widehat{\pi}$
ETAPA 2: Regresar Y sobre $\widehat{X}$ utilizando MCO y obtener

$$\widehat{\beta}_{MC2E} = \left[\widehat{X}'\widehat{X}\right]^{-1}\widehat{X}'Y$$

Notar que $\widehat{X} = Z\widehat{\pi} = Z\left[Z'Z\right]^{-1}Z'X$
Se puede probar que esto es equivalente estimar por VI la ecuación original utilizando como instrumento para $X, \widehat{X}$, aquí

$$\widehat{\beta}_{MC2E} = \left[\widehat{X}'X\right]^{-1}\widehat{X}'Y,$$

esto es debido que $\widehat{X}'\widehat{X} = X'Z\left[Z'Z\right]^{-1}Z'X = \widehat{X}'X$.

Mínimos Cuadrados en 2 Etapas (MC2E)

NOTAS sobre el estimador MC2E:

1) No es correcto obtener la varianza del estimador $\widehat{\beta}_{MC2E}$ a través de los errores estándar de la segunda etapa.

2) Si no hay intrumentos válidos para los regresores endógenos (variables a instrumentar) los parámetros no estarán identificados

3) Si hay instrumentos válidos pero debilmente correlacionados con las variables a instrumentar el estimador tendrá poca precisión.

4) El método Generalizado de los Momentos (MGM) se utiliza como alternativa más eficiente que MC2E (no lo veremos en el curso)

Propiedades de los estimadores VI/MC2E

a) **CONSISTENCIA** Si se cumple $E(u_i|Z) = 0$ el estimador VI/MC2E será sesgado en pequeñas muestras pero consistente

b) **EFICIENCIA** El estimador VI/MC2E será ineficiente con relación al MCO, pero la comparación tiene sentido sólo bajo los supuestos $E(u_i|X) = 0$ y $E(u_i|Z) = 0$. En dicho caso (considerando el MRLS y bajo el supuesto $E(\mathrm{u}\,\mathrm{u}'|X, Z) = \sigma^2 I$) las varianzas asintóticas estarán dadas por

$$Var\left(\widehat{\beta}_{jMCO}\right) = \frac{\sigma_u^2}{N\sigma_x^2} \leq Var\left(\widehat{\beta}_{jVI}\right) = \frac{\sigma_u^2}{N\sigma_x^2\rho_{zx}^2}$$

donde $\rho_{zx}^2 \epsilon\,[0, 1]$ es el cuadrado de la correlación entre x y z.

c) **NORMALIDAD ASINTÓTICA:** El estimador $\widehat{\beta}_{VI}/\widehat{\beta}_{MC2E}$ es asintóticamente normal bajo condiciones generales (no se requiere normalidad de los errores)

En resumen: necesitaremos muestras grandes para poder hacer inferencia a través del estimador VI o MC2E, y sólo se utilizará este estimador en el caso que no se satisfaga el supuesto de exogeneidad de los regresores. En caso contrario se utiliza MCO que es eficiente.

Contraste de Hausman (Wooldridge 15.5)

Hay varios contrastes que permiten someter a prueba la exogeneidad de los regresores. En este curso sólo veremos el contraste de Hausman.
Hausman (1978) propuso un test que permite contrastar la presencia de problemas de endogeneidad en los regresores. En términos bien generales el contraste es el siguiente.
Tengo dos estimadores:

$$\widehat{\theta}_E$$
$$\widehat{\theta}_R$$

$\widehat{\theta}_E$ es un estimador consistente y eficiente bajo la hipótesis nula (por ejemplo MCO en una regresión lineal de un modelo homocedástico, bajo la hipótesis de no correlación entre las x y el error) pero inconsistente bajo la hipótesis alternativa
$\widehat{\theta}_R$ es un estimador consistente tanto bajo la hipótesis nula como bajo la hipótesis alternativa, pero (probablemente) menos eficiente (por ejemplo VI) que $\widehat{\theta}_E$ bajo la nula

Contraste de Hausman

Ho: $\widehat{\theta}_E$ consistente y eficiente, $\widehat{\theta}_R$ consistente
H1: $\widehat{\theta}_E$ inconsistente, $\widehat{\theta}_R$ consistente

Hausman demostró que este contraste se puede realizar a través del estadístico

$$\begin{aligned} h &= \left(\widehat{\theta}_R - \widehat{\theta}_E\right)' \left[\widehat{V}\left(\widehat{\theta}_R - \widehat{\theta}_E\right)\right]^{-1} \left(\widehat{\theta}_R - \widehat{\theta}_E\right) \qquad (1) \\ &= \left(\widehat{\theta}_R - \widehat{\theta}_E\right)' \left[\widehat{V}\left(\widehat{\theta}_R\right) - \widehat{V}\left(\widehat{\theta}_E\right)\right]^{-1} \left(\widehat{\theta}_R - \widehat{\theta}_E\right) \end{aligned}$$

y que este estadístico se distribuye asintóticamente (bajo la hipótesis nula) como una χ^2_k.

Hausman aplicado a los estimadores MCO y VI (MC2E):

$$\widehat{\theta}_E = \widehat{\beta}_{MCO}$$
$$\widehat{\theta}_R = \widehat{\beta}_{MC2E}$$

Los grados de libertad de la distribución chi-cuadrado corresponden a la dimensión del vector $\left(\widehat{\theta}_R - \widehat{\theta}_E\right)$.

Contraste de Hausman (procedimiento alternativo)

Consideremos un modelo

$$\begin{aligned} y_i &= x_i^{'}\beta + u_i \\ x_i^{'} &= (1, x_{1i}, x_{2i}, ...x_{ki}) \\ cov(x_{ji}, u_i) &= 0 \text{ para } j < k \\ cov(x_{ji}, u_i) &\neq 0 \text{ para } j = k \end{aligned}$$

Tengo una muestra $\{y_i, x_i, z_i\}_{i=1,2....N}$

Para realizar Contraste de Hausman procedo en dos etapas:

Etapa 1:

Regreso por MCO x_k sobre **todas** las variables exógenas (modelo reducido):

$$x_k = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + ... + \pi_{k-1} x_{k-1} + z'\theta + v_k$$

obtengo

$$\begin{aligned} \widehat{x}_{ki} &= \widehat{\pi}_0 + \widehat{\pi}_1 x_{1i} + \widehat{\pi}_2 x_{2i} + ... + \widehat{\pi}_{k-1} x_{k-1i} + z_i'\widehat{\theta} \\ \widehat{v}_{ki} &= x_{ki} - \widehat{x}_{ki} \end{aligned}$$

Etapa 2:
Regreso por MCO:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k + \delta \widehat{v}_k + u_k$$

El contraste de Hausman consiste en probar la significación de $\widehat{v}_k$

$$\begin{aligned} Ho &: \delta = 0 \\ H_1 &: \delta \neq 0 \end{aligned}$$

Si rechazo Ho, tengo evidencia que me permite afirmar que x_k es endógena y por lo tanto es recomendable utilizar VI para la estimación del modelo.
Nota: La validez del contraste de Hausman requiere del cumplimiento del supuesto de homocedasticidad y ausencia de correlación de los errores del modelo a estimar.

Bibliografía:

*** Wooldridge, J. M. (2001) Introducción a la Econometría: un enfoque moderno, Thomson Learning, México. (2a. Edición en español, 2006). Capítulos 1, 9 (9.2 y 9.3) y 15. Apéndice B.

** Hayashi, F. (2000) Econometrics, Princeton University Press. Capítulo 3 (3.1).

** Card, D. (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp, ed. L. N. Christophides, E.K. Grant, and R. Swidinsky. Toronto: University of Toronto Press, 201–222.

* Cameron A. C. y P.K. Trivedi (2009) Microeconometrics Using Stata, Stata Press. Capítulo 6.

* Wooldridge, J. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press. Capítulo 4 (4.3 y 4.4) y Capítulo 5 (5.1 y 5.3).

* Greene, W. H. (1999) Análisis Econométrico. 3a. Edición. Prentice Hall Iberia, Madrid. Capítulo 9.