

Econometría II

*Regresores estocásticos y errores de
medición*

Alejandro Mosiño - *Universidad de Guanajuato*

v. 2026.03.22

Regresores estocásticos

Regresores estocásticos (1/6)

Consideremos el modelo:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}$$

Y supongamos que el vector $\mathbf{x}$ es estocástico. En este caso:

$$\begin{aligned}\mathbb{E}(\mathbf{b}) &= \boldsymbol{\beta} + \mathbb{E}[(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'\mathbf{u})] \\ &\neq \boldsymbol{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbb{E}(\mathbf{u})\end{aligned}$$

Regresores estocásticos (2/6)

Se puede demostrar que todas las propiedades de los estimadores de MCO se siguen satisfaciendo si suponemos que el vector $\mathbf{x}$ y el vector $\mathbf{u}$ son independientes. Es decir:

$$\begin{aligned}\mathbb{E}(\mathbf{u}|\mathbf{x}) &= \mathbb{E}(\mathbf{u}) \\ &= \mathbf{0}\end{aligned}$$

Lo anterior nos permite calcular:

$$\begin{aligned}\mathbb{E}(\mathbf{y}|\mathbf{x}) &= \mathbf{x}\beta \\ \mathbb{E}(\mathbf{u}'\mathbf{u}|\mathbf{x}) &= \sigma^2 \mathbf{I}_n\end{aligned}$$

Regresores estocásticos (3/6)

Ejercicio

Muestra que:

$$\begin{aligned}\mathbb{E}(\mathbf{b}|\mathbf{x}) &= \mathbb{E}(\mathbf{b}) \\ &= \boldsymbol{\beta} \\ \text{var}(\mathbf{b}|\mathbf{x}) &= \sigma^2(\mathbf{x}'\mathbf{x})^{-1} \\ \text{var}(\mathbf{b}) &= \sigma^2\mathbb{E}[(\mathbf{x}'\mathbf{x})^{-1}] \\ \mathbb{E}(s^2|\mathbf{x}) &= \mathbb{E}(s^2) \\ &= \sigma^2\end{aligned}$$

Lo anterior implica que el *teorema de Gauss-Markov* sigue siendo válido cuando las $\mathbf{x}$ son estocásticas. ¿Por qué? (**Pista:** Usa la **ley de las expectativas iteradas**.)

Regresores estocásticos (4/6)

Cuando las $\mathbf{x}$ son estocásticas, también resulta importante suponer que:

$$p \lim \left(\frac{1}{n} \mathbf{u}' \mathbf{u} \right) = \mathbf{\Omega}$$

$$p \lim \left(\frac{1}{n} \mathbf{x}' \mathbf{x} \right) = \mathbf{Q}$$

$$p \lim \left(\frac{1}{n} \mathbf{x}' \mathbf{u} \right) = \mathbf{0}$$

Regresores estocásticos (5/6)

Ejercicio

Muestra que el estimador de MCO para β sigue siendo consistente cuando las $\mathbf{x}$ son estocásticas.

Regresores estocásticos (6/6)

Ejercicio

Definamos la matriz de varianzas y covarianzas asintótica de $\mathbf{b}$ como:

$$\text{asy var}(\mathbf{b}) = \frac{1}{n} \lim_{n \rightarrow \infty} \mathbb{E}[\sqrt{n}(\mathbf{b} - \beta)][\sqrt{n}(\mathbf{b} - \beta)]'$$

Muestra que esta es igual a:

$$\text{asy var}(\mathbf{b}) = \frac{1}{n} \sigma^2 \mathbf{Q}^{-1}$$

Nota. Calcula el siguiente límite y comenta.

$$\lim_{n \rightarrow \infty} \mathbb{E}[(\mathbf{b} - \beta)][(\mathbf{b} - \beta)]'$$

Errores de medición en los regresores

Errores de medición, x's (1/5)

Supongamos que el modelo verdadero es:

$$y_i = \beta_1 + \beta_2 z_i + v_i$$

Sin embargo no observamos z_i . En su lugar observamos:

$$x_i = z_i + w_i$$

Errores de medición, x's (2/5)

Suponemos que:

$$V_i \sim iid \square(0, \sigma_V^2) ; \quad V = v, w, z$$

Además:

$$v \perp w; \quad v \perp z; \quad w \perp z$$

y:

$$p \lim \frac{1}{n} \sum_{i=1}^n (V_i - \overline{V})^2 = \sigma_V^2$$

Errores de medición, x's (3/5)

La información disponible nos permite reescribir el modelo en una forma más conocida:

$$y_i = \beta_1 + \beta_2 x_i + u_i;$$

$$u_i = v_i - \beta_2 w_i$$

Nota sin embargo que al estimar por MCO:

$$\begin{aligned} \mathbb{E}(b_2) &= \beta_2 + \mathbb{E} \left(\frac{\text{cov}(x, u)}{\text{var}(x)} \right) \\ &\neq \beta_2 \end{aligned}$$

Esto se debe a que x y u *no son independientes*. En otras palabras, nos enfrentamos a un problema de **endogeneidad**.

Errores de medición, x's (4/5)

Dado que no podemos calcular la distribución exacta de b_2 , nos concentraremos en ver qué pasa cuando el número de observaciones tiende a infinito. Se puede demostrar que:

$$p \lim(b_2) = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}$$

Por lo tanto el sesgo de b_2 (para muestras grandes) es:

$$\text{Sesgo} = -\beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}$$

el cual depende del signo de β_2 .

Errores de medición, x's (5/5)

Ejercicio

Muestra que:

$$p \lim(b_2) = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}$$

Nota. ¿Qué pasa si introducimos en el modelo una *variable proxy imperfecta*?

Simulación (1/2)

Una simulación nos muestra los efectos de los errores de medición en los regresores. Sea el modelo verdadero:

$$y_i = 2 + 0.8z_i + v_i$$

$$x_i = z_i + \omega_i$$

$$z_i \sim \square(10, 4)$$

$$v_i \sim \square(0, 4)$$

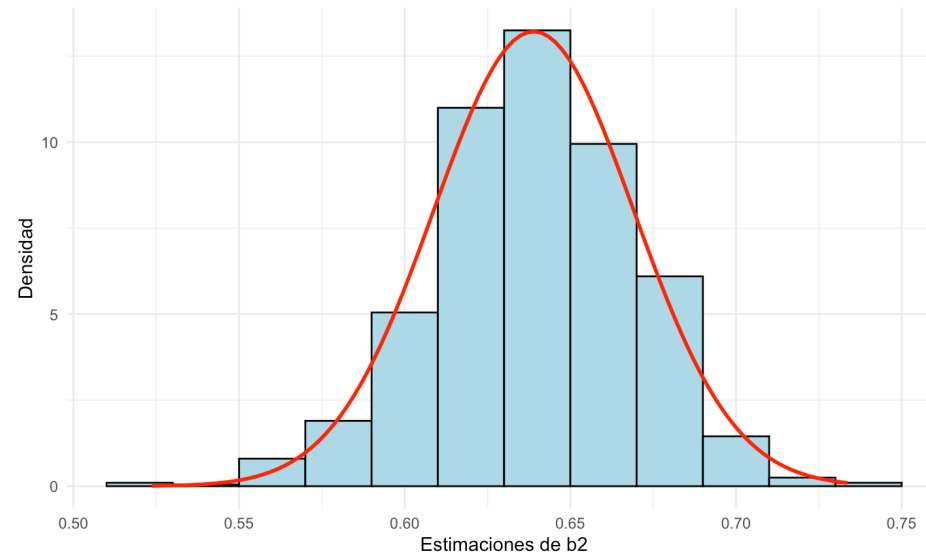
$$\omega_i \sim \square(0, 1)$$

Ejercicio

Muestra que si usamos x_i en lugar de z_i en la regresión, obtendremos que: $p \lim(b_2) = 0.64$.

Simulación (2/2)

La siguiente gráfica nos muestra la distribución de b_2 para $n = 1000$.



Nota que la media de la distribución es cercana a 0.64 aun para una muestra pequeña.

Errores de medición en la variable dependiente

Errores de medición, y (1/5)

Supongamos que el modelo verdadero es:

$$Q_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Sin embargo no observamos Q_i . En su lugar observamos:

$$y_i = Q_i + \psi_i$$

Errores de medición, y (2/5)

Suponemos que:

$$W_i \sim iid \square(0, \sigma_W^2); \quad W = \varepsilon, x, \psi$$

Además:

$$\varepsilon \perp x; \quad \varepsilon \perp \psi; \quad x \perp \psi$$

y:

$$p \lim \frac{1}{n} \sum_{i=1}^n \left(W_i - \overline{W} \right)^2 = \sigma_W^2$$

Errores de medición, y (3/5)

Ejercicio

Muestra que, cuando la variable dependiente está medida con error, los estimadores de MCO siguen siendo insesgados. Es decir:

$$\mathbb{E}(b_2) = \beta_2$$

Errores de medición, y (4/5)

Aunque los estimadores de MCO siendo insesgados, se puede demostrar que son ineficientes (para muestras grandes). En particular, se puede demostrar que:

$$\sigma_{b_2}^2 = \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{n\sigma_s^2} \geq \frac{\sigma_\varepsilon^2}{n\sigma_x^2}$$

Errores de medición, y (5/5)

Ejercicio

Muestra que, para muestras grandes:

$$\sigma_{b_2}^2 = \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{n\sigma_x^2}$$

Nota. ¿Qué pasa si tanto x_i como y_i están sujetas a errores de medición?

Variables Instrumentales (IV)

Variables instrumentales (1/2)

Imaginemos que:

$$p \lim \left(\frac{1}{n} \mathbf{x}' \mathbf{u} \right) \neq \mathbf{0}$$

En este caso, el estimador de MCO *deja de ser consistente*:

$$\begin{aligned} p \lim(\mathbf{b}) &= \boldsymbol{\beta} + p \lim \left(\frac{1}{n} \mathbf{x}' \mathbf{x} \right)^{-1} \cdot p \lim \left(\frac{1}{n} \mathbf{x}' \mathbf{u} \right) \\ &\neq \boldsymbol{\beta} \end{aligned}$$

Variables instrumentales (2/2)

Sin embargo, supongamos que existe u vector de variables $\mathbf{z}$ *de la misma dimensión* que $\mathbf{x}$, tal que:

$$p \lim \left(\frac{1}{n} \mathbf{z}' \mathbf{u} \right) = \mathbf{0}$$

$$p \lim \left(\frac{1}{n} \mathbf{z}' \mathbf{x} \right) = \mathbf{\Sigma}$$

$$p \lim \left(\frac{1}{n} \mathbf{z}' \mathbf{z} \right) = \mathbf{\Sigma}^*$$

donde las matrices $\mathbf{\Sigma}$ y $\mathbf{\Sigma}^*$ deben existir y ser no singulares. Las variables en $\mathbf{z}$ pueden utilizarse como **instrumentos** de $\mathbf{x}$.

El estimador de VI (1/3)

¿Cómo utilizamos el vector de instrumentos $\mathbf{z}$? Supongamos que el modelo es el de siempre:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}$$

Pre multiplicamos ambos lados de la ecuación por $\mathbf{z}'$:

$$\mathbf{y}^* = \mathbf{x}^* \boldsymbol{\beta} + \mathbf{u}^*,$$

$$\mathbf{y}^* = \mathbf{z}' \mathbf{y}$$

$$\mathbf{x}^* = \mathbf{z}' \mathbf{x}$$

$$\mathbf{u}^* = \mathbf{z}' \mathbf{u}$$

El estimador de VI (2/3)

Ejercicio

Muestra que si estimamos el modelo resultante por MCO obtenemos:

$$\hat{\beta}_{IV} = (\mathbf{z}'\mathbf{x})^{-1} (\mathbf{z}'\mathbf{y})$$

que es conocido como el **estimador de variables instrumentales** de β .

El estimador de VI (3/3)

Ejercicio

Muestra que el estimador de variables instrumentales de β es consistente y que su varianza asintótica es:

$$\text{asy var}(\hat{\beta}_{IV}) = \frac{1}{n} \sigma^2 \Sigma^{-1} \Sigma^* \Sigma'^{-1}$$

Nota. En la práctica, la varianza asintótica puede estimarse como:

$$\text{est asy var}(\hat{\beta}_{IV}) = \frac{1}{n} s^2 \Sigma^{-1} \Sigma^* \Sigma'^{-1}$$

$$s^2 = \frac{1}{n - k} (\mathbf{y} - \mathbf{X}\hat{\beta}_{IV})' (\mathbf{y} - \mathbf{X}\hat{\beta}_{IV})$$

Simulación (1/3)

Una simulación nos permite comparar los estimadores de MCO y de IV. Sea el modelo verdadero:

$$y_i = 10 + 5x_i + u_i$$

$$x_i = 2z_i + 0.5v_i + u_i$$

$$z_i \sim \square(0, 1)$$

$$v_i \sim \square(0, 1)$$

$$u_i \sim \square(0, 1)$$

Nota que x_i no es independiente de u_i en la regresión, por lo que los estimadores de MCO estarán sesgados. ¿Podemos utilizar z_i o v_i como instrumentos de x_i en la regresión?

Simulación (2/3)

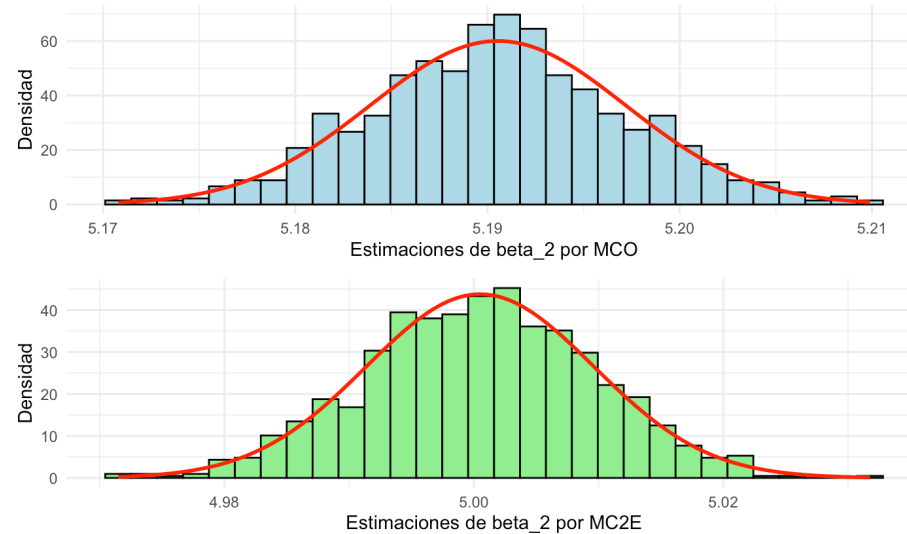
Ejercicio

Utilizando los datos del ejemplo, muestra que si usamos a $\mathbf{z}_i$ como instrumento:

$$\begin{aligned}p \lim b_2^{MCO} &= 5.19 \\p \lim b_2^{VI} &= 5\end{aligned}$$

Ejemplo (3/3)

Las siguientes gráficas comparan la distribución de β_2 estimado por MCO y VI con $n=3200$.



¿Qué estimador te parece mejor? ¿Por qué?

Variables instrumentales: Ejemplo (1/2)

Ejercicio

Utiliza los datos: [E206-mroz](#). Estima el modelo utilizando el método de los MCO. ¿Crees que exista un problema de endogeneidad? ¿Por qué? *Pista:* La variable endógena es la variable educación.

$$l_wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

Variables instrumentales: Ejemplo (2/2)

En el ejemplo anterior, estimamos los coeficientes utilizando la variable *mothereduc*, que representa la educación de la mamá de la i -ésima observación. (¿Por qué?)

```
##  
## Call:  
## ivreg(formula = log(wage) ~ educ + exper + I(exper^2) | mothereduc +      exper + I(exper^2), data = data, subset = (v  
##  
## Coefficients:  
## (Intercept)      educ      exper  I(exper^2)  
##    0.1981861    0.0492630    0.0448558   -0.0009221
```

Mínimos cuadrados en dos etapas (2SLS)

Mínimos cuadrados en dos etapas (1/4)

¿Qué pasa si contamos con más instrumentos que variables explicativas?

Sea:

$$\mathbf{z} = \{\mathbf{z}_1 \quad \mathbf{z}_2 \dots \mathbf{z}_l\}$$

donde $l \geq k$. En este caso, podemos usar el método de los **mínimos cuadrados en dos etapas**.

Mínimos cuadrados en dos etapas (2/4)

1. Estimamos una regresión por cada variable explicativa usando el método de los mínimos cuadrados ordinarios:

$$\mathbf{x}_j = \mathbf{z}\boldsymbol{\delta}_j + \mathbf{e}_j$$

donde:

$$\boldsymbol{\delta}_j = \begin{pmatrix} \delta_{j1} \\ \delta_{j2} \\ \vdots \\ \delta_{jl} \end{pmatrix}, \quad j = 1, 2, \dots, k$$

Mínimos cuadrados en dos etapas (3/4)

2. Recuperamos los valores ajustados:

$$\hat{\mathbf{x}}_j = \mathbf{z}\hat{\delta}_j$$

donde:

$$\hat{\delta}_j = (\mathbf{z}'\mathbf{z})^{-1} (\mathbf{z}'\mathbf{x}_j)$$

Y, naturalmente:

$$\begin{aligned}\mathbf{x}_j &= \mathbf{z}\hat{\delta}_j + \hat{\mathbf{e}}_j \\ &= \hat{\mathbf{x}}_j + \hat{\mathbf{e}}_j\end{aligned}$$

Mínimos cuadrados en dos etapas (4/4)

3. Formamos el vector:

$$\hat{\mathbf{x}} = \{\hat{\mathbf{x}}_1 \quad \hat{\mathbf{x}}_2 \quad \dots \quad \hat{\mathbf{x}}_k\}$$

y usamos MCO para estimar el modelo:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u}$$

De aquí obtenemos el **estimador de mínimos cuadrados en dos etapas** para $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\mathbf{x}'\hat{\mathbf{x}})^{-1} \mathbf{x}'\mathbf{y}$$

Mínimos cuadrados en dos etapas: Ejemplo

En el ejemplo de antes, estimamos los coeficientes utilizando las variables *mothereduc* y *fathereduc*, donde esta última representa la educación del papá de la *i*-ésima observación.

```
##  
## Call:  
## ivreg(formula = log(wage) ~ educ + exper + I(exper^2) | mothereduc + fathereduc + exper + I(exper^2), data = data)  
##  
## Coefficients:  
## (Intercept)      educ      exper  I(exper^2)  
##    0.048100    0.061397    0.044170   -0.000899
```

2SLS vs. IV (1/2)

Ejercicio

Muestra que:

1. $\mathbf{z}'\hat{\mathbf{e}}_j = 0 \quad j = 1, 2, \dots, k$
2. $\mathbf{x}_j'\hat{\mathbf{e}}_j = 0 \quad j = 1, 2, \dots, k$
3. $\mathbf{x}_m'\hat{\mathbf{e}}_j = 0 \quad \forall \quad m \neq j, \quad j, m = 1, 2, \dots, k$

Usando lo anterior, muestra además que:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2SLS} &= (\mathbf{x}'\hat{\mathbf{x}})^{-1} \mathbf{x}'\mathbf{y} \\ &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}\end{aligned}$$

2SLS vs. IV (2/2)

Ejercicio

Demuestra que las estimaciones por mínimos cuadrados en dos etapas y por variables instrumentales son equivalentes cuando tenemos igual número de instrumentos que de variables endógenas. Es decir:

$$\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$$

Nota. ¿Puedes decir cuál es la ventaja de utilizar el método de los mínimos cuadrados en dos etapas?

Selección de instrumentos

Instrumentos fuertes (1/3)

Para seleccionar los instrumentos, primero tenemos que asegurarnos que la correlación entre estos y las variables que van a instrumentar sea alta.

Para ver lo anterior, considera el estimador de variables instrumentales:

$$\begin{aligned}\hat{\beta}_{IV} &= (\mathbf{z}'\mathbf{x})^{-1} (\mathbf{z}'\mathbf{y}) \\ &= \beta + (\mathbf{z}'\mathbf{x})^{-1} (\mathbf{z}'\mathbf{u})\end{aligned}$$

O, en un modelo de regresión simple:

$$\hat{\beta}_{2,IV} = \beta_2 + \frac{\text{cov}(z, u)}{\text{cov}(z, x)}$$

Cuando la $\text{cov}(\mathbf{z}, \mathbf{x})$ se acerca a cero, el sesgo del estimador de β se acerca a infinito. En este caso decimos que $\mathbf{z}$ es un instrumento débil de $\mathbf{x}$. Es fuerte en caso contrario.

Instrumentos fuertes (2/3)

Para hacer una prueba seria para **instrumentos fuertes**, supongamos que x_k es endógena y que contamos con un instrumento z_1 . Entonces:

- Corremos la regresión (que corresponde a la primera etapa del método de 2SLS):

$$x_k = \gamma_1 + \gamma_2 x_2 + \dots + \gamma_{k-1} x_{k-1} + \theta_1 z_1 + v_k$$

- Para que el instrumento z_1 sea fuerte, es necesario que el coeficiente θ_1 sea (muy) significativo. Como regla de uso común, esto será el caso *si el estadístico F de la prueba de significancia $H_0 : \theta_1 = 0$ es mayor a 10.*

Instrumentos débiles (3/3)

- Naturalmente, en caso de tener L instrumentos, corremos la siguiente regresión:

$$x_k = \gamma_1 + \gamma_2 x_2 + \dots + \gamma_{k-1} x_{k-1} + \theta_1 z_1 + \dots + \theta_l z_l + \nu_k$$

- En este caso, el estadístico F de la prueba $\theta_1 = \theta_2 = \dots = \theta_L = 0$ tiene que ser mayor a 10.

En el ejemplo anterior tenemos:

##	df1	df2	statistic	p-value
## Weak instruments	2	423	55.4003	4.268909e-22

Los instrumentos deben ser exógenos

La relación anterior también deja en claro que el sesgo del estimador de β tiende a cero conforme la $cov(z, u)$ se acerca a cero. Cuando esto ocurre decimos que tenemos **instrumentos exógenos**.

Una forma de saber si tenemos instrumentos exógenos es realizando la **prueba de Sargan**.

Prueba de Sargan (1/3)

Para realizar la prueba de Sargan:

- Calculamos los estimadores de variables instrumentales:

$$\hat{\beta}_{IV} = (\mathbf{z}'\mathbf{x}')^{-1} (\mathbf{z}'\mathbf{y})$$

- Recuperamos los residuales:

$$\mathbf{e}_{IV} = \mathbf{y} - \mathbf{x}\hat{\beta}_{IV}$$

Prueba de Sargan (2/3)

- Corremos la siguiente regresión:

$$\mathbf{e}_{IV} = \mathbf{z}\gamma + \mathbf{v}$$

- Calculamos el estadístico nR^2 , donde n es el tamaño de la muestra y R^2 es el coeficiente de determinación de la regresión precedente. Tenemos que:

$$nR^2 \sim \chi^2_{(l-k \text{ g.l.})}$$

donde l es el número de instrumentos y k es el número de variables endógenas.

- El punto anterior implica que, para que la prueba de Sargan pueda realizarse, es necesario tener más instrumentos que variables endógenas $l > k$. Por esta razón, a la prueba de Sargan también se le conoce como **prueba de sobre identificación**.

Prueba de Sargan (3/3)

- Realizamos la siguiente prueba:

$$H_0 : \text{cov}(\mathbf{z}, \mathbf{e}_{IV}) = 0$$

$$H_1 : \text{cov}(\mathbf{z}, \mathbf{e}_{IV}) \neq 0$$

Naturalmente, *si rechazamos* H_0 los instrumentos no son exógenos. En nuestro ejemplo de antes tenemos:

##		df1	df2	statistic	p-value
##	Sargan	1	NA	0.3780715	0.5386372

Prueba de Hausman (1/5)

Finalmente, también necesitamos una prueba para saber si nuestra regresión debe estimarse por el método de las variables instrumentales, o si el método de los MCO es suficiente.

La prueba que debemos realizar en este caso se conoce como la **prueba de Hausman**.

Prueba de Hausman (2/5)

Para mostrar como funciona la prueba de Hausman, supongamos que el modelo es:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

y sospechamos que podría haber una relación entre x_i y u_i . Suponemos que tenemos dos buenos instrumentos para x_i : z_1 y z_2 .

Prueba de Hausman (3/5)

- Estimamos por MCO la siguiente ecuación:

$$x_i = \gamma_1 + \theta_1 z_{1i} + \theta_2 z_{2i} + u_i$$

y recuperamos los residuales:

$$\hat{u}_i = x_i - \hat{\gamma}_1 - \hat{\theta}_1 z_{1i} - \hat{\theta}_2 z_{2i}$$

Prueba de Hausman (4/5)

- Ahora estimamos por MCO:

$$y_i = \beta_1 + \beta_2 x_i + \delta v_i + v_i$$

Una *prueba t* nos permitirá probar la siguiente hipótesis:

$$H_0 : \delta = 0$$

$$H_1 : \delta \neq 0$$

Si rechazamos H_0 el método de las variables instrumentales es apropiado. De otra forma usamos MCO.

¿Por qué?

Prueba de Hausman (5/5)

En el ejemplo de antes tenemos:

##	df1	df2	statistic	p-value
## Wu-Hausman	1	423	2.792592	0.09544055

Nota.

En caso de tener más de una variable explicativa relacionada con el error, u , hacemos tantas regresiones como sea necesario en el paso 1 e incluimos los residuales de estas en la regresión del paso 2. La prueba de hipótesis en este caso implicaría el uso de una F en lugar de una t .

Práctica

Práctica

Utiliza los datos: [E207-fertility](#) para explicar ¿cómo afecta la fertilidad a la oferta de trabajo? Es decir, ¿cuánto disminuye la oferta de trabajo de una mujer cuando tiene un hijo más?

`morekids` = 1 if mom had more than 2 children

`boy1st` = 1 if 1st child was a boy

`boy2nd` = 1 if 2nd child was a boy

`samesex` = 1 if 1st two children same sex

`agem1` = age of mom at census

`black` = 1 if mom is black

`hispan` = 1 if mom is Hispanic

`othrace` = 1 if mom is not black, Hispanic or white

`weeksm1` = mom's weeks worked in 1979

Realiza un análisis completo del modelo:

$$morekids_i = \beta_1 + \beta_2 weeksworked_i + u_i$$