Endogeneidad

Gabriel V. Montes-Rojas

Sesgo por variables omitidas

Si el modelo **verdadero** o **estructural** es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

donde $E(v|x_1, x_2) = 0$ satisface los supuestos de Gauss-Markov, pero estimamos (**modelo estimado**)

$$y = \gamma_0 + \gamma_1 x_1 + e,$$

donde donde $E(e|x_1) = 0$, entonces decimos que omitimos x_2 (es decir, debería estar pero no la usamos, x_2 es una variable omitida).

Teorema: $E(\hat{\gamma}_1) = \beta_1 + \beta_2 \times \delta_{12}$ *donde* δ_{12} *es el coeficiente de la regresión* $x_2 = \delta_0 + \delta_{12} x_1 + w$.
Probar:...

Sesgo por variables omitidas

Probemos el teorema.
Partimos de la prueba de insesgadez del modelo de regresión simple.

Regresión simple
Una forma de interpretación

- Una forma de ver los modelos de regresión es la siguiente. Notemos que para el modelo $y = \gamma_0 + \gamma_1 x_1 + e$,

$$\gamma_1 = \frac{Cov(y, x_1)}{Var(x_1)},$$

bajo el supuesto $Cov(x_1, e) = 0$, o sea que la variable explicativa no tiene relación con los errores.

- La prueba es sencilla:

$$\frac{Cov(y, x_1)}{Var(x_1)} = \frac{Cov(\gamma_0 + \gamma_1 x_1 + e, x_1)}{Var(x_1)} = \frac{\gamma_1 Cov(x_1, x_1) + Cov(x_1, e)}{Var(x_1)}$$

$$= \gamma_1 + \frac{Cov(x_1, e)}{Var(x_1)} = \gamma_1$$

(porque $Cov(x_1, x_1) = Var(x_1)$ y $Cov(e, x_1) = 0$) Esto significa que γ_1 mide cuanto y se relaciona con x, estandarizado por la varianza de x.

- Ver la equivalencia con la teoría asintótica: $\hat{\gamma}_1 \xrightarrow{p} \frac{Cov(y,x_1)}{Var(x_1)}$, cuando $N \to \infty$.

Regresión múltiple

- Consideremos una regresión múltiple con K variables explicativas (y una constante)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + v$$

- Para el caso de muchos regresores también tenemos la misma interpretación de los coeficientes de la regresión:

$$\beta_j = \frac{Cov(y, \tilde{x}_j)}{Var(\tilde{x}_j)}, j = 1, 2, 3, ..., K,$$

donde $\tilde{x}_j$ es el residuo de la regresión de x_j en todas las demás variables, $\boldsymbol{x}_{-j}$.
Prueba: $Cov(y, \tilde{x}_j) = Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + v, \tilde{x}_j)$. Por uno de los supuestos de MCO, dado que $\tilde{x}_j$ es una función de los regresores, entonces no está correlacionado con v. Además, tampoco está correlacionado con $\boldsymbol{x}_{-j}$. Entonces, $Cov(y, \tilde{x}_j) = Cov(\beta_j x_j, \tilde{x}_j) = \beta_j Cov(x_j, \tilde{x}_j) = \beta_j Var(\tilde{x}_j)$.

Sesgo por variables omitidas (cont.)

- Entonces el sesgo por variables omitidas lo podemos ver como

$$\gamma_1 = \frac{Cov(y, x_1)}{Var(x_1)} = \frac{Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + v, x_1)}{Var(x_1)} = \beta_1 + \beta_2 \frac{Cov(x_2, x_1)}{Var(x_1)}.$$

- ¿Cómo se interpreta $\frac{Cov(x_2,x_1)}{Var(x_1)}$?
- Ver la equivalencia con la teoría asintótica: $\hat{\gamma}_1 \xrightarrow{p} \frac{Cov(y,x_1)}{Var(x_1)}$, cuando $N \to \infty$.

Sobre-especificación: agregar variables irrelevantes

Si el modelo **verdadero** o **estructural** es

$$y = \gamma_0 + \gamma_1 x_1 + e$$

donde $E(e|x_1, x_2) = 0$ satisface los supuestos de Gauss-Markov, pero se estima (**modelo estimado**)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

donde $E(u|x_1, x_2) = 0$, entonces decimos que x_2 es una variable irrelevante para estimar γ_1 (es decir, no debería estar en el modelo).
Teorema: $E(\hat{\beta}_1) = \gamma_1$. Teorema: $Var(\hat{\gamma}_1) \leq Var(\hat{\beta}_1)$.

RESULTADO: Agregar variables irrelevantes no afecta la insesgadez de los estimadores MCO pero incrementa la varianza innecesariamente.
La razón es que $Var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)}$, $j = 2, ..., K$, donde SCT_j es la suma de cuadrados totales de x_j, R_j^2 es el R^2 de regresar x_j en todas las demás variables (cuanto más variables mayor será R_j^2).

El problema de la endogeneidad

En Econometría la endogeneidad tiene una definición particular:
Una variable x_j es **endógena** *si* $Cov(x_j, error) \neq 0$.
Una variable x_j es **exógena** *si* $Cov(x_j, error) = 0$.

El problema de la endogeneidad

- Consideremos el modelo **estructural**

$$\log(wage) = \beta_0 + \beta_1\ educ + \beta_2\ exper + \beta_3\ abil + v$$

- Nuestro interés es estimar β_1 y β_2. Sin embargo, *abil* (ability en inglés) no se puede observar. Por ello obtendríamos estimadores sesgados (por variables omitidas).
- En la práctica tenemos este modelo:

$$\log(wage) = \beta_0 + \beta_1\ educ + \beta_2\ exper + u$$

donde $u \equiv \beta_3\ abil + v$.

- En este caso podemos argumentar que: $Cov(educ, u) \neq 0$, $Cov(exper, u) \neq 0$, siempre y cuando $Cov(educ, abil) \neq 0$, $Cov(exper, abil) \neq 0$.
- Otra forma de verlo es que en el modelo con la variable omitida *abil*,

$$\log(wage) = \gamma_0 + \gamma_1\ educ + \gamma_2\ exper + e,$$

los parámetros γ no van a ser los β.

El problema de la endogeneidad

- Consideremos un modelo estructural general:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \beta_q q + v,$$
$$E(v|x_1, x_2, ..., x_K, q) = 0.$$

- Supongamos que q es no observable. Entonces forma parte del error. Asumamos sin pérdida de generalidad que $E(q) = 0$ (como hay un intercepto no es ningún problema)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u,$$
$$u \equiv \beta_q q + v.$$

- Ahora consideremos la proyección de q en $\boldsymbol{x}$ como

$$q = \delta_0 + \delta_1 x_1 + ... + \delta_K x_K + r,$$

donde por definición $E(r) = 0$, $Cov(x_j, r) = 0, j = 1, 2, ..., K$.

- Entonces,

$$y = (\beta_0 + \beta_q \delta_0) + (\beta_1 + \beta_q \delta_1) x_1 + (\beta_2 + \beta_q \delta_2) x_2 + ... + (\beta_K + \beta_q \delta_K) x_K + (\beta_q r + v),$$

El problema de la endogeneidad

- En el modelo anterior si planteamos

$$abil = \delta_0 + \delta_1 educ + \delta_2 exper + r.$$

- Podemos expresar el sesgo como

$$log(wage) = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)educ + (\beta_2 + \beta_3\delta_2)exper + (\beta_3 r + v).$$

- En este caso, los parámetros γ del modelo con la variable omitida *abil* son $\gamma_0 = \beta_0 + \beta_3\delta_0$, $\gamma_1 = \beta_1 + \beta_3\delta_1$, $\gamma_2 = \beta_2 + \beta_3\delta_2$. También el error $e = \beta_3 r + v$.

Soluciones

Hay 3 posibles soluciones:

1. Medir la variable no observada.
2. Encontrar una **variable proxy**.
3. Encontrar una **variable instrumental**.

Variables proxy

- Consideremos el modelo

$$\log(wage) = \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, abil + v$$

  Tomemos $\boldsymbol{x} = (educ, exper)$.
- Una potencial variable proxy para *abil* es *IQ*.
- La variable proxy debería satisfacer lo siguiente:
  1. $abil = \alpha_0 + \alpha_3 \text{IQ} + v_3$, donde v_3 no esta correlacionado con *educ*, *exper* y *IQ*.
  2. v no esta correlacionado con *educ*, *exper* y *abil*. Otra forma de expresarlo es $E(log(wage)|\boldsymbol{x}, abil, IQ) = E(log(wage)|\boldsymbol{x}, abil)$, y decimos que la proxy es irrelevante para explicar los salarios una vez que las variables observables $\boldsymbol{x}$ y la variable *abil* son usadas.
- Entonces podemos estimar
  $log(wage) = (\beta_0 + \beta_3\alpha_0) + \beta_1 educ + \beta_2 exper + \beta_3\alpha_3 IQ + (v + \beta_3 v_3)$.

Ejemplo: IQ como proxy para abilidad

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/wage2, clear
reg lwage educ exper tenure married south urban black
reg lwage educ exper tenure married south urban black IQ
gen educIQ=educ*IQ
reg lwage educ exper tenure married south urban black IQ educIQ
```

Variables	(1)	(2)	(3)
educ	.065	.054	.018
	(.006)	(.007)	(.041)
exper	.014	.014	.014
	(.002)	(.002)	(.003)
tenure	.012	.011	.011
	(.002)	(.002)	(.002)
married	.199	.200	.201
	(.039)	(.039)	(.039)
south	-.091	-.080	-.080
	(.026)	(.026)	(.026)
urban	.184	.182	.184
	(.027)	(.027)	(.027)
black	-.188	-.143	-.147
	(.038)	(.039)	(.040)
IQ	-	.0036	-.0009
		(.0010)	(.0052)
educIQ	-	-	-.00034
			(.00038)

Sesgo potencial usando una proxy: proxy imperfecta

Asumamos por el contrario que

$$abil = \alpha_0 + \alpha_1 educ + \alpha_2 exper + \alpha_3 IQ + v_3$$

$$\Rightarrow log(wage) = (\beta_0 + \beta_3\alpha_0) + (\beta_1 + \beta_3\alpha_1)educ$$

$$+(\beta_2 + \beta_3\alpha_2)exper + \beta_3\alpha_3 IQ + (v + \beta_3 v_3)$$

En este caso, IQ se define como una **variable proxy imperfecta**. Como puede verse MCO con proxy imperfecta tiene sesgo.

Ejercicio: Comparar $(\beta_1 + \beta_3\alpha_1)$ con $(\beta_1 + \beta_3\delta_1)$ obtenido en el problema de variables omitidas.

Variables instrumentales

Consideremos la siguiente regresión:

$$y = \beta_0 + \beta_1 x + u$$

donde $Cov(x, u) \neq 0$ (o sea, x is endógena)

Una variable instrumental (VI) z debería satisfacer:

1. **Exogeneidad.** No estar correlacionada con el error: $Cov(z, u) = 0$
2. **Relevancia.** Estar correlacionada con la variable endógena: $Cov(x, z) \neq 0$

Variables instrumentales

¿Cómo podríamos estimar β_1 usando z?
Notar que

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$$

¿Por qué?

$$\begin{aligned} Cov(z,y) &= Cov(z, \beta_0 + \beta_1 x + u) \\ &= Cov(z,\beta_0) + Cov(z,\beta_1 x) + Cov(z,u) \end{aligned}$$

Entonces podemos plantear el siguiente estimador de β_1 usando variables instrumentales:

$$\hat{\beta}_1^{VI} = \frac{\widehat{Cov(z,y)}}{\widehat{Cov(z,x)}} = \frac{1/N \sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{1/N \sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})},$$

y notar que $\hat{\beta}_1^{VI} \xrightarrow{p} \frac{Cov(z,y)}{Cov(z,x)} = \beta_1$ cuando $N \to \infty$.

VI como un estimador en dos etapas

- Consideremos la regresión simple

$$y = \beta_0 + \beta_1 x + u,$$

donde $Cov(x, u) \neq 0$.
- Consideremos la siguiente regresión auxiliar (**etapa 1**): $x = \gamma_0 + \gamma_1 z + r$.
- Construir los valores predichos $\hat{x} \equiv \gamma_0 + \gamma_1 z$. Notemos que $x = \hat{x} + r$ and $\gamma_1 = \frac{Cov(x,z)}{Var(z)}$.
- Notemos que $\hat{x}$, que es una función de z y la podemos escribir con $x(z)$, no esta correlacionado con r (por construcción) y también $Cov(\hat{x}, u) = 0$.
- Consideremos otra regresión (**etapa 2**):

$$y = \beta_0 + \beta_1(\hat{x} + r) + u = \beta_0 + \beta_1 \hat{x} + v,$$

donde $v \equiv \beta_1 r + u$ y $Cov(\hat{x}, v) = Cov(\hat{x}, \beta_1 r + u) = 0$. Entonces,

$$\frac{Cov(y, \hat{x})}{Var(\hat{x})} = \frac{Cov(y, \gamma_0 + \gamma_1 z)}{Var(\gamma_0 + \gamma_1 z)} = \frac{Cov\left(y, \frac{Cov(x,z)}{Var(z)} z\right)}{Var\left(\frac{Cov(x,z)}{Var(z)} z\right)}$$

$$= \frac{Cov(z, y)}{Cov(z, x)} = \beta_1.$$

Variables instrumentales en regresión múltiple

Consideremos el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u$$

donde $Cov(x_K, u) \neq 0$ (o sea, x_K es endógena) y $Cov(x_j, u) = 0, j = 1, 2, ..., K-1$ (el resto son exógenas).

Una variable instrumental z debe satisfacer dos condiciones:

1. No estar correlacionada con el error: $Cov(z, u) = 0$
2. Estar correlacionada con la variable endógena. Más formalmente, consideremos la proyección lineal de x_K en **todas** las variables exógenas (las exógenas originales más el instrumento):

$$x_K = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta z + r_K,$$

donde por definición $E(r_K) = 0$ y r_K no está correlacionado con $x_1, x_2, ..., x_{K-1}$. El supuesto importante es que $\theta \neq 0$.

Muchos instrumentos

- Cuando hay más de un instrumento $(z_1, z_2, ..., z_M)$ el estimador más eficiente es el de mínimos cuadrados en dos etapas (two-stage least squares, 2SLS):

$$\hat{\beta}_{2SLS} = \left(N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_i' \boldsymbol{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_i' y_i \right) = (\hat{\boldsymbol{X}}' \boldsymbol{X})^{-1} (\hat{\boldsymbol{X}} \boldsymbol{y})$$

donde

$$x_K = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M + r_K$$

$$\hat{x}_K = \hat{\delta}_1 x_1 + ... + \hat{\delta}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + ... + \hat{\theta}_M z_M$$

- Notemos que $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$, es una proyección de $\boldsymbol{x}$ en el espacio de $\boldsymbol{z} \equiv (x_1, ..., x_{K-1}, z_1, ..., z_M)$, donde $\boldsymbol{P}_{\boldsymbol{Z}}$ es la matriz de proyección. Entonces, $\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} = \hat{\boldsymbol{X}}'\boldsymbol{X}$. Así el estimador 2SLS es un estimador de MCO donde $\hat{\boldsymbol{x}}$ se usa en vez de $\boldsymbol{x}$. O sea, $\hat{\beta}_{2SLS} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}(\hat{\boldsymbol{X}}\boldsymbol{y})$

Identificación de VI

- Consideremos el modelo de regresión

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

- Definamos $\boldsymbol{z} \equiv (1, x_1, ..., x_{K-1}, z)$, como el vector de todas las variables exógenas.
- Hay entonces $K+1$ condiciones de ortogonalidad:

$$E(\boldsymbol{z}'u) = 0.$$

- Multiplicamos el modelo de regresión por $\boldsymbol{z}'$, y tomando esperanzas

$$[E(\boldsymbol{z}'\boldsymbol{x})]\boldsymbol{\beta} = E(\boldsymbol{z}'y),$$

donde $E(\boldsymbol{z}'\boldsymbol{x})$ es una matriz $(K+1)\times(K+1)$ y $E(\boldsymbol{z}'y)$ es $(K+1)\times 1$. Este sistema tiene una única solución si y sólo si la primera matriz tiene rango $K+1$, entonces

$$\boldsymbol{\beta} = [E(\boldsymbol{z}'\boldsymbol{x})]^{-1}E(\boldsymbol{z}'y).$$

- El estimador de **variables instrumentales** de $\boldsymbol{\beta}$ es

$$\hat{\boldsymbol{\beta}}_{VI} = \left(N^{-1}\sum_{i=1}^{N}\boldsymbol{z}_i'\boldsymbol{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\boldsymbol{z}_i'y_i\right) = (\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{Z}\boldsymbol{y})$$

Supuestos para identificación y consistencia de 2SLS

Supuesto 2SLS.1: Para un vector $1 \times L$ $\boldsymbol{z}$, $E(\boldsymbol{z}'u) = 0$.

Supuesto 2SLS.2: (a) rango $E(\boldsymbol{z}'\boldsymbol{z}) = L$; *(b) rango* $E(\boldsymbol{z}'\boldsymbol{x}) = K + 1$.

Una condición necesaria para estas condiciones es que $L \geq K + 1$, *o sea, más instrumentos que variables endógenas.*

Identificación

Identificación: *Si asumimos que* $E(\boldsymbol{z}'\boldsymbol{z})$ *es no singular definamos la proyección* $\boldsymbol{x}^* = \boldsymbol{z}\Pi$, *donde* $\Pi = [E(\boldsymbol{z}'\boldsymbol{z})]^{-1}E(\boldsymbol{z}'\boldsymbol{x})$ *es una matriz* $L \times (K+1)$. *Multiplicando por* $x^{*'}$, *y tomando esperanzas tenemos*

$$E(\boldsymbol{x}^{*'}y) = E(\boldsymbol{x}^{*'}\boldsymbol{x})\boldsymbol{\beta} + E(\boldsymbol{x}^{*'}u) = E(\boldsymbol{x}^{*'}\boldsymbol{x})\boldsymbol{\beta}$$

Así $\boldsymbol{\beta}$ *esta identificado por* $\boldsymbol{\beta} = [E(\boldsymbol{x}^{*'}\boldsymbol{x})]^{-1}E(\boldsymbol{x}^{*'}y)$. *Para esto necesitamos que* $E(\boldsymbol{x}^{*'}\boldsymbol{x}^*)$ *sea no singular. Pero*

$$E(\boldsymbol{x}^{*'}\boldsymbol{x}) = E(\boldsymbol{\Pi}'\boldsymbol{z}'\boldsymbol{x}) = E(\boldsymbol{x}'\boldsymbol{z})[E(\boldsymbol{z}'\boldsymbol{z})]^{-1}E(\boldsymbol{z}'\boldsymbol{x})$$

Entonces esta matriz es no singular si $E(\boldsymbol{z}'\boldsymbol{x})$ *tiene rango* $K+1$ *(Supuesto 2SLS.2b). Para esto también necesitamos* $E(\boldsymbol{z}'\boldsymbol{z})$ *no singular y entonces con rango* L *(Supuesto 2SLS.2a).*

Consistencia de 2SLS

$$\hat{\beta}_{2SLS} = \left[\left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{z}_i \right) \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i' \boldsymbol{z}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i' \boldsymbol{x}_i \right) \right]^{-1}$$
$$\left(N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i' \boldsymbol{z}_i \right) \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i' \boldsymbol{z}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i' y_i \right)$$

Consistencia: Bajo los Supuestos 2SLS.1 y 2SLS.2, plim $\hat{\beta}_{2SLS} = \beta$.

Prueba: Ley de los grandes números y teorema de Slutsky.

Normalidad asintótica de 2SLS

Supuesto 2SLS.3: $E(u^2 \boldsymbol{z}'\boldsymbol{z}) = \sigma^2 E(\boldsymbol{z}'\boldsymbol{z})$, *donde* $\sigma^2 = E(u^2)$.

Normalidad asintótica: Bajo los supuestos 2SLS.1, 2SLS.2 y 2SLS.3,
$\sqrt{N}(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2([E(\boldsymbol{x}'\boldsymbol{z})][E(\boldsymbol{z}'\boldsymbol{z})]^{-1}[E(\boldsymbol{z}'\boldsymbol{x})])\right)$.

Contrastes para endogeneidad

- El estimador de 2SLS es menos eficiente (mayor varianza) que MCO con variable enxógenas.
- La estimación de modelos 2SLS es más demandante en términos computacionales.
- Entonces es importante chequear primero si hay endogeneidad para evitar usar un estimator ineficiente innecesariamente.

Tomemos el modelo

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u$$

donde y_2 es (potencialmente) endógena; z_1 and z_2 son variables explicativas exógenas; z_3 and z_4 son IV. Para contrastar por endogeneidad:

1. $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2$ y construir los residuos $\hat{v}_2$
2. $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + +\beta_3 z_2 + \delta_1 \hat{v}_2 + error$
3. Contrastar por la significancia estadística de $\hat{v}_2$, $H_0 : \delta_1 = 0$.
4. Si rechazamos la hipótesis nula entonces hay evidencia que u y v_2 están correlacionados y y_2 es endógena.

Contrastes para endogeneidad

Consideremos ahora el contraste de **Durbin-Wu-Hausman** (DWH) que esta basado en la comparación de $\hat{\beta}_{2SLS}$ y $\hat{\beta}_{OLS}$. (La misma idea se ve en datos en panel para comparar RE y FE.)
Bajo la hipótesis nula de exogeneidad, $H_0 : E(\boldsymbol{x}'u) = 0$. Entonces,

1. Ambos estimadores son consistentes para β.
2. Entonces la hipótesis nula se puede redefinir con $H_0 : \hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$.
3. Bajo H_0 (y asumiendo homoscedasticidad)
   $Avar[\sqrt{N}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})] = \sigma^2 \left([E(\boldsymbol{x}^{*\prime}\boldsymbol{x}^*)]^{-1} - [E(\boldsymbol{x}'\boldsymbol{x}]^{-1} \right)$.
4. Dado que MCO es más eficiente, entonces la varianza es definida semipositiva.
5. En particular,

$$DWH = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'[(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1} - (\boldsymbol{X}'\boldsymbol{X})](\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})/\hat{\sigma}^2 \overset{a}{\sim} \chi^2_{L-K}$$

Contraste para la validez de los instrumentos

Requerimiento importante: Necesitamos **más** variables instrumentales que variables endógenas.

1. Supongamos que en el modelo anterior usamos 2SLS con z_3 como la **única** variable instrumental.
2. Computar $\hat{u}_3 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 y_2 - \hat{\beta}_2 z_1 - \hat{\beta}_3 z_2$.
3. Correr la regresión auxiliar $\hat{u}_3 = \delta_0 + \hat{\delta}_1 z_1 + \hat{\delta}_2 z_2 + \delta_4 z_4$.
4. Chequer la significancia de z_4.
5. Esto nos da un contraste válido para la validez de z_4 como VI. Pero tenemos que asumir que z_3 es una VI válida.

Contraste para la validez de los instrumentos

Contraste de Sargan-Hausman

1. Si tenemos más VIs que variables endógenas, entonces el modelo esta sobre-identificado (over-identified).
2. Consideremos H_0 : todas las VIs son exgógenas. Si rechazamos entonces alguna de las VIs es endgógena.
3. Estimar el modelo con todos las VIs usando 2SLS. Obtener los residuos $\hat{u}$.
4. Correr la regresión de $\hat{u}$ en TODAS las variables exógenas (VIs, X exógenas, constante).
5. Computar $NR_u^2 \overset{a}{\sim} \chi^2_{L-K}$, donde R_u^2 es el de la última regresión.

VI en STATA

- Asumamos que `x1` is (potentialmente endógena y `x2` is exógena. Asumamos la existencia de 2 VI: `z1`, `z2`
- `ivregress 2sls y (x1=z1 z2) x2`
- `ivregress 2sls y (x1=z1 z2) x2, first` (para que muestre la primera etapa)
- `estat firststage` (significancia de los instrumentos - necesitamos $F > 10$)
- También podemos usar `reg x1 z1 z2` y `test z1 z2`
- `estat overid` (validez de los instrumentos)
- `estat endogenous` (exogeneidad de todas las variables)

VI en STATA

- Para entender VI se puede correr un estimador en dos etapas a mano para reproducir

```
ivregress 2sls y (x1=z1 z2) x2
```

- Los mismos coeficientes se pueden obtener con

```
reg x1 z1 z2 x2
predict x1hat
reg y x1hat x2
```

- Notar que los errores estándar son diferentes. ¿Por qué?

Ejemplos en la web

- http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge9.html
- http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge15.html
- https://www.stata.com/manuals13/rivregress.pdf