Econometría I
Variables instrumentales y Mínimos cuadrados en dos etapas (2SLS)

Ramiro de Elejalde

Facultad de Economía y Finanzas
Universidad Alberto Hurtado

Outline

Referencias: Cap. 4 de Angrist and Pischke, cap. 5 de Wooldridge, y cap. 12 de Stock and Watson.

Modelo simple

Modelo simple

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

donde y_i es la variable dependiente, x_i es una variable endógena y z_i es un instrumento.

Supuestos: $\text{Cov}(x_i, u_i) \neq 0$ (x_i es endógena)

IV1 Muestra aleatoria de tamaño $N \implies \{y_i, x_i, z_i\}_{i=1}^{N}$ i.i.d.,

IV2 Exogeneidad del instrumento (Restricción de exclusión): $\text{Cov}(z_i, u_i) = 0$,
Dos condiciones: (1) z_i es *como si fuese* asignada en forma aleatoria, (2) z_i no tiene un efecto directo sobre y_i.

IV3 Relevancia del instrumento: $\text{Cov}(z_i, x_i) \neq 0$

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

- ¿Cuál es el efecto de tener hijos en la oferta laboral de las mujeres?
- Datos: Census Public Use Micro Sample (PUMS) para 1980 y 1990.
- Muestra: Una muestra con mujeres casadas con al menos 2 hijos.
- Variable dependiente: estar empleada (*workedm*), semanas trabajadas (*weeksm*1) y horas trabajadas por semana (*hourswm*).
- Variable explicativa: Tener 3 hijos o más (*morekids*).
- Instrumentos: Hijos del mismo sexo (*samesex*).

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

- ¿El hecho que los dos primeros dos hijos sean del mismo sexo (*samesex*) es un buen instrumento para tener tres hijos o más (*morekids*)?

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

- ¿El hecho que los dos primeros dos hijos sean del mismo sexo (*samesex*) es un buen instrumento para tener tres hijos o más (*morekids*)?
- Exogeneidad
  - ¿Es el sexo de los hijos asignado aleatoriamente?
  - ¿El sexo de los hijos puede estar correlado con la decisión de trabajar de la madre por razones distintas de su efecto sobre el número de niños?
- Relevancia: ¿Tener dos hijos del mismo sexo afecta la decisión de tener un hijo adicional?
  - Se puede inferir de los datos.

Identificación

- ¿Qué momentos poblacionales permiten identificar los parámetros de interés?

Identificación

- ¿Qué momentos poblacionales permiten identificar los parámetros de interés?

$$
\begin{aligned}
&\mathrm{Cov}(z_i, u_i) = 0 \quad \text{usando IV2,} \\
\iff\ &\mathrm{Cov}(z_i(y_i - \beta_0 - \beta_1 x_i)) = 0, \\
\iff\ &\mathrm{Cov}(z_i, y_i) - \beta_1 \mathrm{Cov}(z_i, x_i) = 0, \\
\iff\ &\beta_1 = \frac{\mathrm{Cov}(z_i, y_i)}{\mathrm{Cov}(z_i, x_i)} \quad \text{usando IV3.}
\end{aligned}
$$

Identificación

- ¿Qué momentos poblacionales permiten identificar los parámetros de interés?

$$
\begin{aligned}
&\text{Cov}(z_i, u_i) = 0 \quad \text{usando IV2,} \\
\iff &\text{Cov}(z_i(y_i - \beta_0 - \beta_1 x_i)) = 0, \\
\iff &\text{Cov}(z_i, y_i) - \beta_1 \text{Cov}(z_i, x_i) = 0, \\
\iff &\beta_1 = \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_i)} \quad \text{usando IV3.}
\end{aligned}
$$

- IV2 y IV3 son los supuestos que identifican a β_1.

Mínimos Cuadrados Indirectos

- Primera etapa

$$x_i = \pi_{10} + \pi_{11} z_i + \epsilon_{1i}.$$

- Forma reducida

$$y_i = \pi_{20} + \pi_{21} z_i + \epsilon_{2i}.$$

Entonces,

$$\begin{aligned} \beta_1 &= \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_i)}, \\ &= \frac{\text{Cov}(z_i, y_i)/\text{Var}(z_i)}{\text{Cov}(z_i, x_i)/\text{Var}(z_i)}, \\ \implies \beta_1 &= \frac{\pi_{21}}{\pi_{11}}. \end{aligned}$$

- Intuición: Correlación entre z_i e y_i solamente se puede deber al efecto a través de x_i.

Mínimos Cuadrados en 2 Etapas (2SLS)

- Primera etapa

$$x_i = \pi_{10} + \pi_{11} z_i + \epsilon_{1i} = x_i^* + \epsilon_{1i}.$$

- Luego reemplazar x_i por la predicción x_i^* (que no está correlada con el error).

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i, \\ &= \beta_0 + \beta_1 x_i - \beta_1 x_i^* + \beta_1 x_i^* + u_i, \\ &= \beta_0 + \beta_1 x_i^* + \beta_1 (x_i - x_i^*) + u_i. \end{aligned}$$

  Entonces,

$$\beta_1 = \frac{\text{Cov}(x_i^*, y_i)}{\text{Var}(x_i^*)}.$$

- Intuición: Solamente utilizamos la variación exógena de x_i.

Estimación

- Estimador por variables instrumentales (IV): utilizamos el principio de analogía para reemplazar momentos poblacionales por momentos muestrales.

  Estamos interesados en:

$$\beta_1 = \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_i)}.$$

Estimación

- Estimador por variables instrumentales (IV): utilizamos el principio de analogía para reemplazar momentos poblacionales por momentos muestrales.

  Estamos interesados en:

$$\beta_1 = \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_i)}.$$

  Usando el principio de analogía, lo estimamos con

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)}.$$

Estimación

- Estimador por variables instrumentales (IV): Usando el principio de analogía, lo estimamos con

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)}.$$

Estimación

- Estimador por variables instrumentales (IV): Usando el principio de analogía, lo estimamos con

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)}.$$

- En forma similar podemos motivar el estimador por Mínimos Cuadrados Indirectos

$$\hat{\beta}_{1,\text{ILS}} = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}},$$

Estimación

- Estimador por variables instrumentales (IV): Usando el principio de analogía, lo estimamos con

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)}.$$

- En forma similar podemos motivar el estimador por Mínimos Cuadrados Indirectos

$$\hat{\beta}_{1,\text{ILS}} = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}},$$

y el estimador 2SLS

$$\hat{\beta}_{1,\text{2SLS}} = \frac{\widehat{\text{Cov}}(\hat{x}_i, y_i)}{\widehat{\text{Var}}(\hat{x}_i)}, \quad \text{donde } \hat{x}_i = \hat{\pi}_{10} + \hat{\pi}_{11} z_i.$$

Los estimadores son equivalentes: $\hat{\beta}_{1,IV} = \hat{\beta}_{1,ILS} = \hat{\beta}_{1,2\text{SLS}}$.

Demostración:

- $\hat{\beta}_{1,IV} = \hat{\beta}_{1,ILS}$

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)} = \frac{\widehat{\text{Cov}}(z_i, y_i)/\widehat{\text{Var}}(z_i)}{\widehat{\text{Cov}}(z_i, x_i)/\widehat{\text{Var}}(z_i)} = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}} = \hat{\beta}_{1,ILS}.$$

  donde $x_i = \hat{\pi}_{10} + \hat{\pi}_{11} z_i + \hat{\epsilon}_{1i}$ y $y_i = \hat{\pi}_{20} + \hat{\pi}_{21} z_i + \hat{\epsilon}_{2i}$.

- $\hat{\beta}_{1,2SLS} = \hat{\beta}_{1,IV}$

$$\hat{\beta}_{1,2\text{SLS}} = \frac{\widehat{\text{Cov}}(\hat{x}_i, y_i)}{\widehat{\text{Var}}(\hat{x}_i)} = \frac{\widehat{\text{Cov}}(\hat{x}_i, y_i)}{\widehat{\text{Cov}}(\hat{x}_i, x_i)} = \frac{\hat{\pi}_{11}\widehat{\text{Cov}}(z_i, y_i)}{\hat{\pi}_{11}\widehat{\text{Cov}}(z_i, x_i)} = \hat{\beta}_{1,IV}.$$

  donde $\hat{x}_i = \hat{\pi}_{10} + \hat{\pi}_{11} z_i$.

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

Variable	Obs	Mean	Std. Dev.	Min	Max
workedm	262793	.5301892	.4990887	0	1
weeksm1	262793	19.10925	21.88929	0	52
hourswm	262793	16.81808	18.37502	0	99
morekids	262793	.383686	.4862838	0	1
samesex	262793	.5053369	.4999725	0	1

Estimación MCO

```
. reg hourswm morekids, robust

Linear regression                               Number of obs =  262793
                                                F(  1,262791) = 2540.58
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0095
                                                Root MSE      =  18.287

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -3.687237   .0731534    -50.40   0.000    -3.830616   -3.543858
       _cons |   18.23282   .0456501    399.40   0.000     18.14335    18.32229
------------------------------------------------------------------------------
```

Primera etapa

```
.    reg morekids samesex, robust

Linear regression                               Number of obs =  262793
                                                F(  1,262791) = 1253.62
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0047
                                                Root MSE      =  .48513

------------------------------------------------------------------------------
             |               Robust
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0669939   .0018921    35.41   0.000     .0632854    .0707024
       _cons |   .3498315   .0013228   264.47   0.000     .3472389    .3524241
------------------------------------------------------------------------------
```

Se puede utilizar para contrastar el supuesto de relevancia!!!

$$\text{Cov}(x_i, z_i) \neq 0 \iff \pi_{11} \neq 0$$

donde $x_i = \pi_{10} + \pi_{11} z_i + e_{1i}$.

Forma reducida

```
.          reg hourswm samesex, robust

Linear regression                               Number of obs =  262793
                                                F(  1,262791) =   25.55
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0001
                                                Root MSE      =  18.374

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |  -.3623775   .0716935    -5.05   0.000    -.5028947   -.2218602
       _cons |    17.0012   .0510958   332.73   0.000     16.90105    17.10135
------------------------------------------------------------------------------
```

Mínimos cuadrados indirectos: $-0.362/0.067 = -5.409$

Mínimos cuadrados en 2 etapas: A mano

```
.          reg hourswm morekidshat, robust

Linear regression                               Number of obs =  262793
                                                F(  1,262791) =   25.55
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0001
                                                Root MSE      =  18.374

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
morekidshat |  -5.409112   1.070149    -5.05   0.000    -7.506575   -3.311649
       _cons |   18.89348   .4123485    45.82   0.000     18.08529    19.70167
------------------------------------------------------------------------------
```

Variables instrumentales/Mínimos cuadrados en 2 etapas

```
.    ivregress 2sls hourswm (morekids = samesex), robust

Instrumental variables (2SLS) regression              Number of obs =  262793
                                                      Wald chi2(1)  =   25.74
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0074
                                                      Root MSE      =  18.306

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -5.409112   1.066216    -5.07   0.000    -7.498856   -3.319368
       _cons |   18.89348   .4108644    45.98   0.000      18.0882    19.69876
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   samesex
```

Estimador de Wald

```
. tabstat hourswm morekids, statistic(mean) by(samesex)

Summary statistics: mean
by categories of: samesex (first two kids are of same sex)

 samesex |   hourswm  morekids
---------+--------------------
       0 |   17.0012  .3498315
       1 |  16.63882  .4168254
---------+--------------------
   Total |  16.81808   .383686
------------------------------
```

Estimador de Wald=(16.6-17.0)/(.417-.350)=-5.41

Retornos de la educación (Angrist and Krueger, QJE 1991)

- ¿Cuáles son los retornos de la educación?
- Estamos interesados en el modelo

$$y_i = \alpha + \rho s_i + \gamma a_i + \nu_i$$

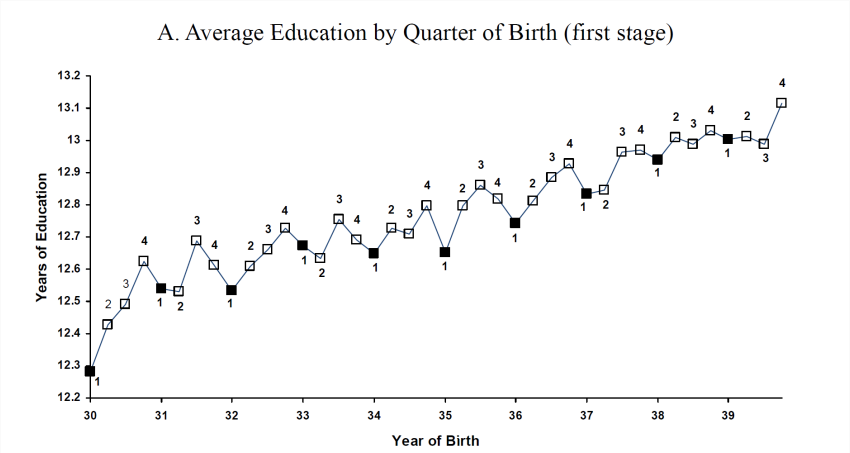
donde y_i es log de ingresos, s_i es escolaridad, a_i es habilidad innata y ν_i son inobservados que cumplen $\text{Cov}(s_i, \nu_i) = \text{Cov}(a_i, \nu_i) = 0$.

- Si observamos habilidad innata, podemos estimar el efecto causal por MCO.
- Desafortunadamente, no lo observamos y el modelo que podemos estimar $y_i = \alpha + \rho s_i + u_i$ donde $u_i = \gamma a_i + \nu_i$ tiene un sesgo por omisión de variable.
- Solución de variables instrumentales: una variable correlada con s_i pero con que no esté correlada con u_i.

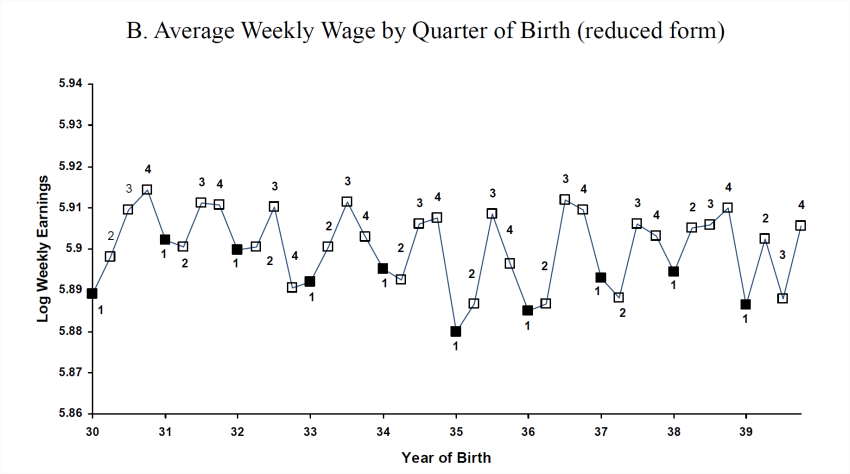
Retornos de la educación (Angrist and Krueger, QJE 1991)

- Instrumento: Angrist and Krueger argumentan que las personas que nacen en los últimos trimestres del año empiezan la escuela más jóvenes, y permanecen más tiempo en la escuela a causa de las leyes de escolaridad obligatoria que dependen de la edad.
- Variables explicativas adicionales: año de nacimiento y estado donde nació.

Retornos de la educación (Angrist and Krueger, QJE 1991)



Retornos de la educación (Angrist and Krueger, QJE 1991)



Retornos de la educación (Angrist and Krueger, QJE 1991)

TABLE 4.1.1
2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years of education	.071	.067	.102	.13	.104	.108	.087	.057
	(.0004)	(.0004)	(.024)	(.020)	(.026)	(.020)	(.016)	(.029)
Exogenous Covariates								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year-of-birth dummies		✓			✓	✓	✓	✓
50 state-of-birth dummies		✓			✓	✓	✓	✓
Instruments								
dummy for QOB = 1			✓	✓	✓	✓	✓	✓
dummy for QOB = 2				✓		✓	✓	✓
dummy for QOB = 3				✓		✓	✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓	✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

Retornos de la educación (Angrist and Krueger, QJE 1991)

- Buckles and Hungerman, "Season of Birth and Later Outcomes: Old Questions, New Answers" (REStat, 2013).
- Idea: Niños nacidos en diferentes meses del año son concebidos por madres con características socieconómicas diferentes.
- Los autores encuentran que niños nacidos en invierno tienen una mayor probabilidad de haber nacido de una madre adolescente, menor probabilidad de que la madre esté casada o la madre tenga secundario completo.
- Mecanismo: el clima en verano afecta en forma diferencial los patrones de fertilidad de los distintos grupos socieconómicos.

Propiedades de $\hat{\beta}_{IV}$

- Hasta ahora demostramos, en el modelo de regresión simple con un instrumento, el estimador IV es idéntico a 2SLS.
- Además, el estimador de Wald es un caso particular de IV cuando el instrumento es una variable dummy.

Propiedades de $\hat{\beta}_{IV}$

- Hasta ahora demostramos, en el modelo de regresión simple con un instrumento, el estimador IV es idéntico a 2SLS.
- Además, el estimador de Wald es un caso particular de IV cuando el instrumento es una variable dummy.
- Dado que los tres estimadores son idénticos, analizamos las propiedades del estimador IV:

$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)},$$

bajos los supuestos de

IV1 Muestra aleatoria de tamaño $N \implies \{y_i, x_i, z_i\}_{i=1}^{N}$ i.i.d.,

IV2 Exogeneidad del instrumento: $\text{Cov}(z_i, u_i) = 0$,

IV3 Relevancia del instrumento: $\text{Cov}(z_i, x_i) \neq 0$

Consistencia

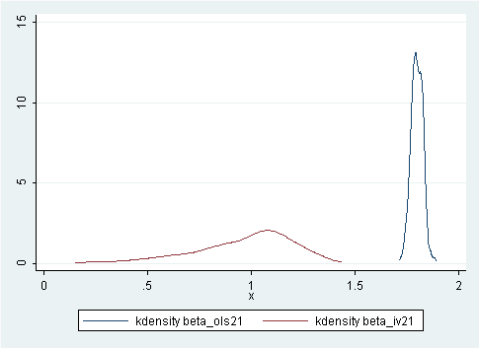
$$\hat{\beta}_{1,IV} = \frac{\widehat{\text{Cov}}(z_i, y_i)}{\widehat{\text{Cov}}(z_i, x_i)} = \frac{\widehat{\text{Cov}}(z_i, \beta_0 + \beta_1 x_i + u_i)}{\widehat{\text{Cov}}(z_i, x_i)},$$
$$= \beta_1 + \frac{\widehat{\text{Cov}}(z_i, u_i)}{\widehat{\text{Cov}}(z_i, x_i)},$$

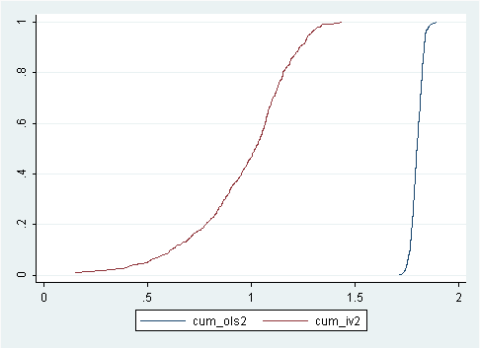
usando los resultados derivados en el capítulo de Teoría asintótica, y IV2 (exogeneidad) y IV3 (relevancia)

$$\text{plim}\, \widehat{\text{Cov}}(z_i, x_i) = \text{Cov}(z_i, x_i) \neq 0,$$
$$\text{plim}\, \widehat{\text{Cov}}(z_i, u_i) = \text{Cov}(z_i, u_i) = 0.$$

Entonces $\text{plim}\, \hat{\beta}_{1,IV} = \beta_1$.

Discusión: $\hat{\beta}_{1,IV}$ no es insesgado.





$\beta_1 = 1$ y $N = 500$.

Distribución asintótica normal

$$\sqrt{N}(\hat{\beta}_{1,IV} - \beta_1) = \frac{1}{\widehat{\text{Cov}}(z_i, x_i)} \sqrt{N}\widehat{\text{Cov}}(z_i, u_i).$$

Usando los resultados derivados en el capítulo de Teoría asintótica, y IV2 y IV3,

$$\text{plim}\, \widehat{\text{Cov}}(z_i, x_i) = \text{Cov}(z_i, x_i) \neq 0,$$
$$\sqrt{N}\widehat{\text{Cov}}(z_i, u_i) \xrightarrow{d} \text{Normal}(0, V).$$

donde

$$\begin{aligned} V &= \mathbb{E}[(z - \mathbb{E}\, z)^2 u^2] - \mathbb{E}[(z - \mathbb{E}\, z)u]^2, \\ &= \mathbb{E}[(z - \mathbb{E}\, z)^2 u^2] - \text{Cov}(z, u)^2, \\ &= \mathbb{E}[(z - \mathbb{E}\, z)^2 u^2]. \end{aligned}$$

Distribución asintótica normal

Usando el teorema de Slutsky

$$\sqrt{N}(\hat{\beta}_{1,IV} - \beta_1) \xrightarrow{d} \text{Normal}\left(0, \frac{\mathbb{E}[(z - \mathbb{E}\, z)^2 u^2]}{\text{Cov}(z,x)^2}\right).$$

- Solamente para obtener intuición y comparar con MCO asumimos homoscedasticidad ($\mathbb{E}(u^2|z) = \mathbb{E}(u^2) = \sigma^2$) y obtenemos:

$$V = \mathbb{E}[(z - \mathbb{E}\, z)^2 u^2] = \text{Var}(z)\, \mathbb{E}(u^2).$$

$$\begin{aligned}\sqrt{N}(\hat{\beta}_{1,IV} - \beta_1) &\xrightarrow{d} \text{Normal}\left(0, \frac{\text{Var}(z)\sigma^2}{\text{Cov}(z,x)^2}\right), \\ &\xrightarrow{d} \text{Normal}\left(0, \frac{\sigma^2}{\rho_{x,z}^2 \text{Var}(x)}\right).\end{aligned}$$

Distribución asintótica normal

- Decimos

$$\hat{\beta}_{1,IV} \overset{a}{\sim} \text{Normal}(\beta, \text{AVar}(\hat{\beta}_{1,IV})), \text{ donde}$$

$$\text{AVar}(\hat{\beta}_{1,IV}) = \frac{1}{N} \frac{\sigma^2}{\rho^2_{x,z} \text{Var}(x)}.$$

Distribución asintótica normal

$$\text{AVar}(\hat{\beta}_{1,OLS}) = \frac{\sigma^2}{N}\frac{1}{\text{Var}(x)},$$
$$\text{AVar}(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{N}\frac{1}{\rho^2_{x,z}\,\text{Var}(x)}.$$

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

Estimación MCO

```
.          reg hourswm morekids, robust

Linear regression                                      Number of obs =  262793
                                                       F(  1,262791) = 2540.58
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0095
                                                       Root MSE      =  18.287

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -3.687237   .0731534   -50.40   0.000    -3.830616   -3.543858
       _cons |   18.23282   .0456501   399.40   0.000     18.14335    18.32229
------------------------------------------------------------------------------
```

Mínimos cuadrados en 2 etapas

```
.    ivregress 2sls hourswm (morekids = samesex), robust

Instrumental variables (2SLS) regression              Number of obs =  262793
                                                      Wald chi2(1)  =   25.74
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0074
                                                      Root MSE      =  18.306

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -5.409112   1.066216    -5.07   0.000    -7.498856   -3.319368
       _cons |   18.89348   .4108644    45.98   0.000      18.0882    19.69876
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   samesex
```

Instrumentos débiles

- Instrumentos débiles: los instrumentos explican poco de la variabilidad de la variable endógena.
- ¿Por qué es un problema?
  - Si el instrumento tiene alguna correlación con el inobservado, el sesgo asintótico de IV puede ser mayor a MCO.
  - Aumentan los errores estándar de la estimación.
  - Aproximación normal es una pobre aproximación y el sesgo del estimador IV (para un tamaño de muestra) aumenta con el número de instrumentos

Instrumentos débiles

- ¿Cuando es un problema?
  - Regla práctica cuando tenemos una variable endógena: Necesitamos un $F_N > 10$ en la primera etapa en el test sobre los instrumentos (Stock and Yogo).

Instrumentos débiles

Recomendaciones

- Reportar la primera etapa y evaluar si los signos son los esperados.
- Reportar estadístico F sobre los instrumentos en la primera etapa. Criterio F $>$10.
- Elegir mejor instrumento y estimar el modelo con un sólo instrumento.
- Estimar el modelo con todos los instrumentos con LIML.
- Comparar los resultados de 2SLS, 2SLS con un instrumento y LIML.

Sesgo asintótico cuando el instrumento no es exógeno.

Sesgo asintótico cuando el instrumento no es exógeno.

$$
\begin{aligned}
\text{plim}\,\hat{\beta}_{1,OLS} &= \beta_1 + \frac{\sigma_u}{\sigma_x}\rho_{x,u}, \\
\text{plim}\,\hat{\beta}_{1,IV} &= \beta_1 + \frac{\sigma_u}{\sigma_x}\frac{\rho_{z,u}}{\rho_{z,x}}, \\
sesgo(\hat{\beta}_{1,IV}) &\leq sesgo(\hat{\beta}_{1,OLS}), \\
\iff \frac{\rho_{z,u}}{\rho_{z,x}} &\leq \rho_{x,u}.
\end{aligned}
$$

Sesgo asintótico cuando el instrumento no es exógeno.

- Modelo generador de datos

$$y_i = 1 + x_i + u_i,$$

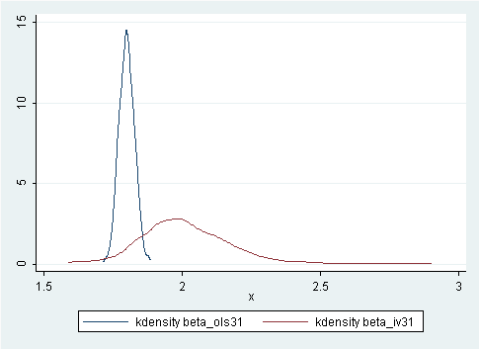
y tenemos un instrumento disponible z_i tal que

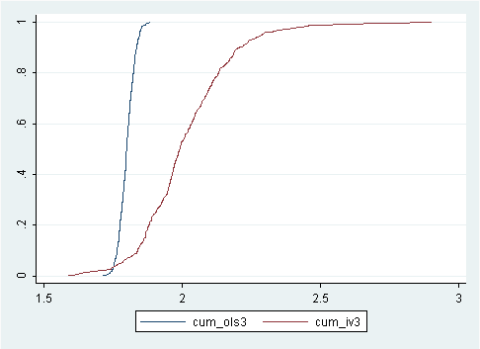
$$\begin{pmatrix} u \\ z \\ x \end{pmatrix} = \text{Normal}\left(0, \begin{pmatrix} 1, \rho_{u,z}, \rho_{u,x} \\ \rho_{u,z}, 1, \rho_{z,x} \\ \rho_{u,x}, \rho_{z,x}, 1 \end{pmatrix}\right)$$

- Tenemos un muestra aleatoria de tamaño $N = 500$, $\{y_i, x_i, z_i\}_{i=1}^{500}$

Sesgo asintótico cuando el instrumento no es exógeno.

x endógena con $\rho_{u,x} = 0.80$, z endógena con $\rho_{z,u} = 0.20$ y $\rho_{z,x} = 0.20$

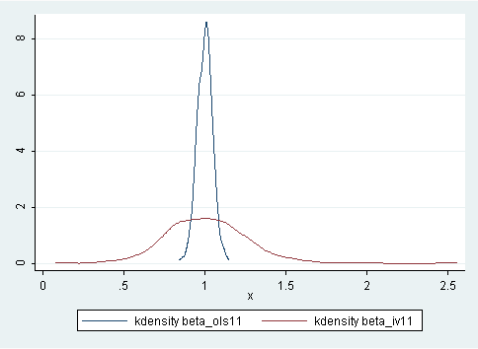


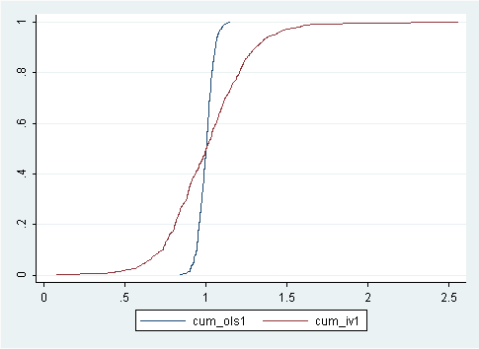


Si el instrumento no es perfecto (está correlado con u), el sesgo de IV puede ser mayor al sesgo MCO.

Aumentan los errores estándar de la estimación en MCO y IV.

x exógena, z exógena y $\rho_{z,x} = 0.20$

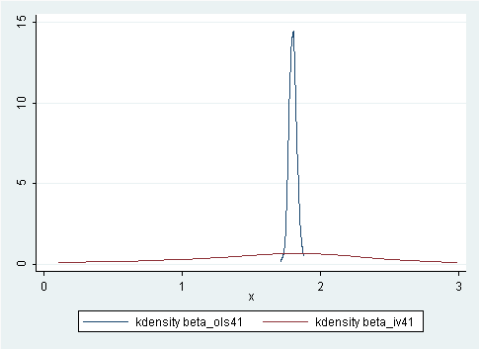


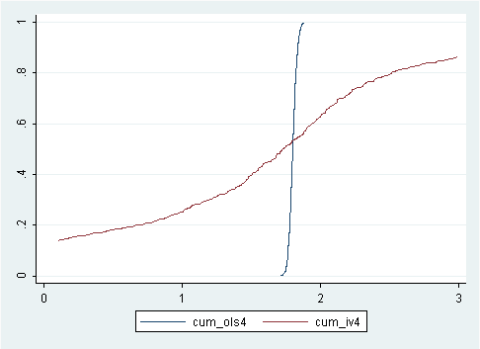


Cuando ambos son consistentes, MCO es más eficiente que IV.

Sesgo del estimador IV el instrumento es irrelevante

x endógena con $\rho_{u,x} = 0.80$, z exógena pero irrelevante $\rho_{z,x} = 0.00$

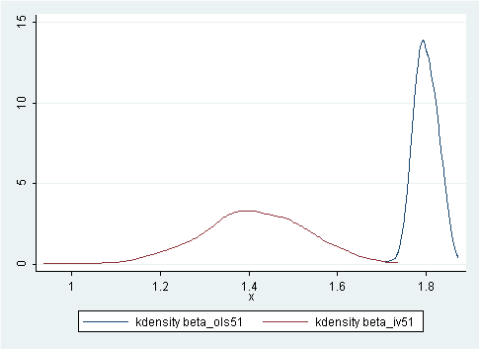


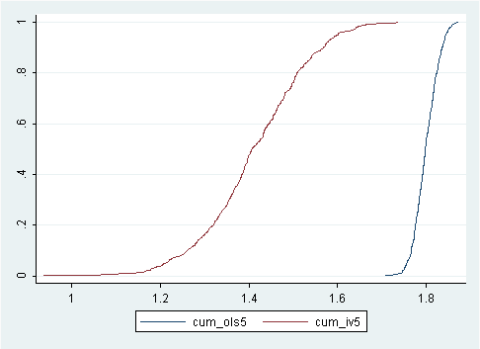


Si el instrumento no cumple la condición de rango, el sesgo de IV es similar al sesgo de MCO. No es tan grave porque la varianza de la estimación IV hace que la estimación no sea informativa.

Sesgo del estimador IV aumenta con el número de instrumentos

x endógena con $\rho_{u,x} = 0.80$, z exógena y $\rho_{z,x} = 0.20$, y 20 instrumentos irrelevantes adicionales





Si tenemos muchos instrumentos débiles, a medida que aumentan los instrumentos el sesgo de IV se acerca al sesgo de MCO.

Modelo general

Modelo general

$$y_i = x_i'\beta + u_i,$$

donde

- x_i son K variables explicativas de donde K_1 son exógenas y K_2 son endógenas ($K = K_1 + K_2$), y
- z_i son L variables exógenas donde K_1 son las variables exógenas incluidas en el modelo y L_1 son variables exógenas excluidas en el modelo (instrumentos) ($L = K_1 + L_1$).

Modelo general

$$y_i = x_i'\beta + u_i,$$

donde

- x_i son K variables explicativas de donde K_1 son exógenas y K_2 son endógenas ($K = K_1 + K_2$), y
- z_i son L variables exógenas donde K_1 son las variables exógenas incluidas en el modelo y L_1 son variables exógenas excluidas en el modelo (instrumentos) ($L = K_1 + L_1$).

Suponemos que $L = dim(z) \geq dim(x) = K \iff L_1 \geq K_2$: Tenemos al menos tantos instrumentos como variables endógenas.

Si $L = K$ (y se cumple una condición de rango) el modelo está exactamente identificado; si $L > K$ el modelo está potencialmente sobreidentificado.

Efecto de la fertilidad en la oferta laboral (Angrist y Evans)

- Modelo

$$hourswm = \beta_0 + \beta_1\, agem1 + \beta_2\, agefstm + \beta_3\, boy1st + \beta_4\, blackm + \\ + \beta_5\, hispm + \beta_6\, othracem + \beta_7\, morekids + u_i.$$

  donde tener 3 hijos o más (*morekids*), edad de la madre (*agem*1), edad de la madre cuando tuvo su primer hijo (*agefstm*), primer hijo varón (*boy*1*st*), raza negra (*blackm*), hispano (*hispm*), otra raza distinto de blanco (*othracem*).

- Instrumentos: Dos hijas mujeres (*girls*2) y dos hijos varones (*boys*2).

Efecto de la fertilidad en la oferta laboral (Angrist y Evans)

- Modelo

$$hourswm = \beta_0 + \beta_1\, agem1 + \beta_2\, agefstm + \beta_3\, boy1st + \beta_4\, blackm + \\ + \beta_5\, hispm + \beta_6\, othracem + \beta_7\, morekids + u_i.$$

  donde tener 3 hijos o más (*morekids*), edad de la madre (*agem*1), edad de la madre cuando tuvo su primer hijo (*agefstm*), primer hijo varón (*boy*1*st*), raza negra (*blackm*), hispano (*hispm*), otra raza distinto de blanco (*othracem*).

- Instrumentos: Dos hijas mujeres (*girls*2) y dos hijos varones (*boys*2).
- Entonces $K = 8$, $L = 9$.
  $x = (1,\ agem1,\ agefstm,\ boy1st,\ blackm,\ hispm,\ othracem,\ morekids)$
  $z = (1,\ agem1,\ agefstm,\ boy1st,\ blackm,\ hispm,\ othracem,\ girls2,\ boys2)$

Testear competencia imperfecta en el mercado de pescado de Fulton, NY, Graddy (Rand, 1995)

- ¿Es el mercado de pescado de Fulton perfectamente competitivo?
- Evidencia de discriminación de precios (asiáticos pagan precios más bajos).
- Nos enfocamos en la estimación de la demanda de pescado.
- Datos: ventas del día, precio promedio del día, día de la semana y otras variables de un puesto de venta de pescado (merluza) en el mercado de Fulton en Nueva York desde Diciembre de 1991 a Marzo 1992

Modelo econométrico

$$ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t + \alpha\, lavgprc_t + u_t,$$

donde *lavgprc*: Logaritmo del precio promedio del día ($ por libra), *ltotqty*: Logaritmo de ventas totales del día en libras, $mon = 1$ para Lunes, $tues = 1$ para Martes, $wed = 1$ para Miércoles, $thurs = 1$ para Jueves.

- Instrumentos: *speed*2: Mínimo en los últimos 2 días de la velocidad media del viento, *speed*3: Mínimo hace 3 días de la velocidad media del viento, *wave*2: Máximo en los últimos 2 días de la altura promedio de las olas, *wave*3: Máximo hace 3 y 4 días de la altura promedio de las olas.

Modelo econométrico

$$ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t + \alpha\, lavgprc_t + u_t,$$

- Instrumentos: $speed2$, $speed3$, $wave2$, $wave3$.

Modelo econométrico

$$ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t + \alpha\, lavgprc_t + u_t,$$

- Instrumentos: *speed*2, *speed*3, *wave*2, *wave*3.
- Entonces $K = 6$, $L = 9$.
  $x = (1,\ mon,\ tues,\ wed,\ thurs,\ lavgprc)$
  $z = (1,\ mon,\ tues,\ wed,\ thurs,\ speed2,\ speed3,\ wave2,\ wave3)$

Modelo general

$$y_i = x_i'\beta + u_i.$$

Supuestos: $\text{Cov}(x_j, u) \neq 0$ para $j = K_1 + 1, ..., K$.

IV1 Muestra aleatoria de tamaño $N \implies \{y_i, x_i, z_i\}_{i=1}^N$ i.i.d.,

IV2 Exogeneidad: $\mathbb{E}(zu) = 0$,

IV3 No multicolinealidad perfecta entre exógenas: $rango(\mathbb{E}(zz')) = L$

IV4 Condición de rango: $rango(\mathbb{E}(zx')) = K$.
Una condición necesaria es la condición de orden: $L \geq K \iff L_1 \geq K_2$, tenemos tantos instrumentos como variables endógenas.

IV5 $\mathbb{E}(u^2 zz')$ existe.

Condición de Rango

- La condición $rango(\mathbb{E}(zx')) = K$ es la generalización del supuesto de relevancia del modelo simple.
- Para modelo simple tenemos $x = (1 \quad x_1)$ y $z = (1 \quad z_1)$.

Condición de Rango

- La condición $rango(\mathbb{E}(zx')) = K$ es la generalización del supuesto de relevancia del modelo simple.
- Para modelo simple tenemos $x = (1 \quad x_1)$ y $z = (1 \quad z_1)$.

  Entonces:

$$\mathbb{E}(zx') = \begin{pmatrix} 1 & \mathbb{E}(x_1) \\ \mathbb{E}(z_1) & \mathbb{E}(z_1 x_1) \end{pmatrix}.$$

  $rango(\mathbb{E}(zx')) = 2$ si y sólo sí $|\mathbb{E}(zx')| \neq 0$.

Condición de Rango

- La condición $rango(\mathbb{E}(zx')) = K$ es la generalización del supuesto de relevancia del modelo simple.
- Para modelo simple tenemos $x = (1 \quad x_1)$ y $z = (1 \quad z_1)$.

  Entonces:

$$\mathbb{E}(zx') = \begin{pmatrix} 1 & \mathbb{E}(x_1) \\ \mathbb{E}(z_1) & \mathbb{E}(z_1 x_1) \end{pmatrix}.$$

  $rango(\mathbb{E}(zx')) = 2$ si y sólo sí $|\mathbb{E}(zx')| \neq 0$.

  $|\mathbb{E}(zx')| = \mathbb{E}(z_1 x_1) - \mathbb{E}(z_1)\mathbb{E}(x_1) = \text{Cov}(z_1, x_1) \neq 0.$

Condición de Rango

- Para una variable endógena y un instrumento, la condición de rango es equivalente a contrastar si el instrumento es significativo en la primera etapa.
- Para una variable endógena y múltiples instrumentos, la condición de rango es equivalente a contrastar si al menos un instrumento es significativo en la primera etapa.
- Para más de una variable endógena, una condición suficiente es que instrumentos distintos sean significativos para distintas v. endógenas en la primera etapa. Existen contrastes que evalúan si la matriz $\mathbb{E}(zx')$ tiene rango K.
- Contraste de Wald (o F) en la primera etapa H_0 : los coeficientes de los instrumentos son cero, versus H_1 : al menos uno de los coeficientes de los instrumentos es distinto de cero.

Identificación para $L = K$

- ¿Qué momentos poblacionales permiten identificar los parámetros de interés?

Identificación para $L = K$

- ¿Qué momentos poblacionales permiten identificar los parámetros de interés?

$$
\begin{aligned}
&\mathbb{E}(zu) = 0 \quad \text{usando IV2,} \\
\iff\ &\mathbb{E}(z(y - x'\beta)) = 0, \\
\iff\ &\mathbb{E}(zy) - \mathbb{E}(zx')\beta = 0, \\
\iff\ &\beta = \mathbb{E}(zx')^{-1}\,\mathbb{E}(zy) \quad \text{usando IV3.b.}
\end{aligned}
$$

- IV2 y IV3.b son los supuestos que identifican β.

Estimador por variables instrumentales (IV)

- Principio de analogía: reemplazar momentos poblacionales por momentos muestrales.

Estimador por variables instrumentales (IV)

- Principio de analogía: reemplazar momentos poblacionales por momentos muestrales.

  Estamos interesados en:

$$\beta = \mathbb{E}(zx')^{-1}\,\mathbb{E}(zy).$$

  Usando el principio de analogía, lo estimamos con

$$\hat{\beta}_{IV} = \left(\frac{1}{N}\sum z_i x_i'\right)^{-1} \frac{1}{N}\sum z_i y_i.$$

- En notación matricial: $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$ donde $Z = (z_1, ..., z_N)'$.

Estimación para $L \geq K$

- Supongamos que tenemos una variable endógena y dos instrumentos: ¿Podemos combinar ambos instrumentos en la estimación? ¿Cómo?

Estimación para $L \geq K$

- Supongamos que tenemos una variable endógena y dos instrumentos: ¿Podemos combinar ambos instrumentos en la estimación? ¿Cómo?
- Bajo homoscedasticidad, el mejor instrumento es la combinación lineal de las variables exógenas que se obtiene del modelo de regresión de la primera etapa.

Estimación para $L \geq K$

- Supongamos que tenemos una variable endógena y dos instrumentos: ¿Podemos combinar ambos instrumentos en la estimación? ¿Cómo?
- Bajo homoscedasticidad, el mejor instrumento es la combinación lineal de las variables exógenas que se obtiene del modelo de regresión de la primera etapa.
- Podemos escribir la primera etapa como

$$x = \Pi' z + e,$$

donde $\Pi : L \times K$ es igual a $\Pi = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')$ y $\mathbb{E}(ze') = 0$.

- El vector de instrumentos de x es:

$$x^* = \Pi' z.$$

- Note que para x_j exógena usamos la misma variable como instrumento: $x_j^* = x_j$.

- Podemos escribir la primera etapa como

$$x = \Pi' z + e,$$

donde $\Pi = \mathbb{E}(zz')^{-1}\,\mathbb{E}(zx')$ y $\mathbb{E}(ze) = 0$.

- Demostración:

El sistema de ecuaciones de la primera etapa es

$$\begin{aligned} x_1 &= \pi_1' z + e_1, \\ &\vdots \\ x_K &= \pi_K' z + e_K. \end{aligned}$$

Se puede escribir

$$x = \begin{bmatrix} \pi_1' \\ \vdots \\ \pi_K' \end{bmatrix} z + e = [\pi_1 \dots \pi_K]' z + e = \Pi' z + e.$$

Estimación para $L \geq K$

- Podemos escribir la primera etapa como

$$x = \Pi' z + e,$$

donde $\Pi = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')$ y $\mathbb{E}(ze') = 0$.

- Demostración:

El sistema de ecuaciones de la primera etapa es

$$x = \Pi' z + e,$$

donde $\Pi = [\pi_1 \dots \pi_K]$.

Dado que $\pi_j = \mathbb{E}(zz')^{-1}\mathbb{E}(zx_j)$ entonces

$$\Pi = [\pi_1 \dots \pi_K] = \mathbb{E}(zz')^{-1}[\mathbb{E}(zx_1) \dots \mathbb{E}(zx_K)] = \mathbb{E}(zz')^{-1}\mathbb{E}(zx').$$

Estimación para $L \geq K$

- Supongamos que tenemos una variable endógena y dos instrumentos: ¿Qué instrumento utilizamos?
- Bajo homoscedasticidad, el mejor instrumento es la combinación lineal de las variables exógenas que se obtiene del modelo de regresión de la primera etapa.
- Podemos escribir la primera etapa como

$$x = \Pi' z + e,$$

donde $\Pi : L \times K$ es igual a $\Pi = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')$ y $\mathbb{E}(ze') = 0$.

- El vector de instrumentos de x es:

$$x^* = \Pi' z.$$

- Note que para x_j exógena usamos la misma variable como instrumento: $x_j^* = x_j$.

Instrumentos

Queremos hacer una estimación de IV con $x^* = \Pi' z$. Verifiquemos que los instrumentos cumplen la condición de exogeneidad y rango.

Instrumentos

Queremos hacer una estimación de IV con $x^* = \Pi' z$. Verifiquemos que los instrumentos cumplen la condición de exogeneidad y rango.

- Exogeneidad

  $\mathbb{E}(x^* u) = \mathbb{E}(\Pi' z u) = \Pi' \mathbb{E}(zu) = 0$ por exogeneidad de z.

- Condición de rango: $rango(\mathbb{E}(x^* x')) = K$.

  $\mathbb{E}(x^* x') = \mathbb{E}(\Pi' z x') = \Pi' \mathbb{E}(z x') = \mathbb{E}(x z') \mathbb{E}(z z')^{-1} \mathbb{E}(z x')$
  Como $rango(\mathbb{E}(z z')) = L$, $rango(\mathbb{E}(z x') = K$ y $L \geq K$ entonces
  $rango(\mathbb{E}(x^* x')) = K$.

Estimación

Estimación en dos etapas:

- Regresar x_i en z_i y obtener valores predichos

$$\hat{x}_i = \hat{\Pi}' z_i.$$

  Nota: Es lo mismo que regresar solamente las variables endógenas y utilizar las variables exógenas incluidas como su propio instrumento.

- Usar $\hat{x}_i$ como instrumentos en un IV

$$\hat{\beta}_{IV} = \left(\frac{1}{N} \sum \hat{x}_i x_i' \right)^{-1} \frac{1}{N} \sum \hat{x}_i y_i.$$

Estimación

- Dado que $x_i = \hat{x}_i + \hat{e}_i$ y de las CPO de MCO

$$\frac{1}{N}\sum_i \hat{x}_i \hat{e}_i' = 0,$$

entonces

$$\frac{1}{N}\sum_i \hat{x}_i x_i' = \frac{1}{N}\sum_i \hat{x}_i \hat{x}_i',$$

y el estimador IV se puede escribir como un estimador 2SLS

$$\hat{\beta}_{2SLS} = \left(\frac{1}{N}\sum \hat{x}_i \hat{x}_i'\right)^{-1} \frac{1}{N}\sum \hat{x}_i y_i,$$

donde la segunda etapa es un regresión de y en $\hat{x}_i$.

- Recuerde: No es recomendable computar el estimador usando el procedimiento en dos etapas porque los s.e. están mal calculados.

Propiedades

- Para probar consistencia es conveniente escribir el estimador en función de las variables originales.
- El coeficiente estimado de regresar x_j en los instrumentos es $\hat{\pi}_j = (\sum z_i z_i')^{-1} \sum z_i x_{ji}$.
- La matriz de dichos coeficientes es $\hat{\Pi} = [\hat{\pi}_1 \hat{\pi}_2 ... \hat{\pi}_K] = (\sum z_i z_i')^{-1} \sum z_i x_i'$.
- Por lo tanto podemos escribir:

$$\begin{aligned}\hat{\beta}_{2SLS} &= \left(\sum \hat{x}_i \hat{x}_i'\right)^{-1} \sum \hat{x}_i y_i, \\ &= \left[\left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1}\left(\sum z_i x_i'\right)\right]^{-1} \\ &\quad \left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1}\left(\sum z_i y_i\right).\end{aligned}$$

- Si $K = L$, tenemos el estimador IV.

Consistencia

$$\hat{\beta}_{2SLS} = \beta + \left[\left(\sum x_i z_i' \right) \left(\sum z_i z_i' \right)^{-1} \left(\sum z_i x_i' \right) \right]^{-1}$$
$$\left(\sum x_i z_i' \right) \left(\sum z_i z_i' \right)^{-1} \left(\sum z_i u_i \right)$$

Entonces usando los resultados derivados en el capítulo de Teoría asintótica, y IV2 y IV3

$$\text{plim}\, \hat{\beta}_{2SLS} = \beta + [\mathbb{E}(xz')\, \mathbb{E}(zz')^{-1}\, \mathbb{E}(zx')]^{-1}\, \mathbb{E}(xz')\, \mathbb{E}(zz')^{-1}\, \mathbb{E}(zu).$$

Entonces $\text{plim}\, \hat{\beta}_{2SLS} = \beta$.

Distribución asintótica normal

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta) = \left[\left(\sum x_i z_i' \right) \left(\sum z_i z_i' \right)^{-1} \left(\sum z_i x_i' \right) \right]^{-1} \left(\sum x_i z_i' \right) \left(\sum z_i z_i' \right)^{-1} \sqrt{N} \left(\sum z_i u_i \right)$$

Usando los resultados derivados en el capítulo de Teoría asintótica, y IV2, IV3, y IV4, y el teorema de Slutsky,

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} \text{Normal}(0, B^{-1} C B^{-1}),$$

donde

$$B = \mathbb{E}(xz') \, \mathbb{E}(zz')^{-1} \, \mathbb{E}(zx'),$$
$$C = \mathbb{E}(xz') \, \mathbb{E}(zz')^{-1} \, \mathbb{E}(u^2 zz') \, \mathbb{E}(zz')^{-1} \, \mathbb{E}(zx').$$

Distribución asintótica normal

- Decimos

$$\hat{\beta}_{2SLS} \overset{a}{\sim} \text{Normal}(\beta, \text{AVar}(\hat{\beta}_{2SLS})), \text{ donde}$$
$$\text{AVar}(\hat{\beta}_{2SLS}) = \frac{1}{N} B^{-1} C B^{-1}.$$

- Estimación consistente de $\text{AVar}(\hat{\beta}_{2SLS})$: Reemplazamos momentos poblacionales por muestrales y errores poblacionales (u_i) por errores estimados ($\hat{u}_i$). Es decir,

$$\widehat{\text{AVar}}(\hat{\beta}_{2SLS}) = \frac{1}{N} \hat{B}^{-1} \hat{C} \hat{B}^{-1},$$

donde

$$\hat{B} = \left(\frac{1}{N}\sum x_i z_i'\right)\left(\frac{1}{N}\sum z_i z_i'\right)^{-1}\left(\frac{1}{N}\sum z_i x_i'\right),$$
$$\hat{C} = \left(\frac{1}{N}\sum x_i z_i'\right)\left(\frac{1}{N}\sum z_i z_i'\right)^{-1}\left(\frac{1}{N}\sum \hat{u}_i^2 z_i z_i'\right)\left(\frac{1}{N}\sum z_i z_i'\right)^{-1}\left(\frac{1}{N}\sum z_i x_i'\right).$$

Eficiencia

- Si asumimos homoscedasticidad, es decir IV5 $\mathbb{E}(u^2|z) = \mathbb{E}(u^2) = \sigma^2$:

$$\text{AVar}(\hat{\beta}_{2SLS}) = \frac{1}{N}\sigma^2 B^{-1}.$$

Efecto de la fertilidad en la oferta laboral (Angrist y Evans, AER 1998)

- Modelo

$$hourswm = \beta_0 + \beta_1\, agem1 + \beta_2\, agefstm + \beta_3\, boy1st + \beta_4\, blackm + \\ + \beta_5\, hispm + \beta_6\, othracem + \beta_7\, morekids + u_i.$$

  donde tener 3 hijos o más (*morekids*), edad de la madre (*agem*1), edad de la madre cuando tuvo su primer hijo (*agefstm*), primer hijo varón (*boy*1*st*), raza negra (*blackm*), hispano (*hispm*), otra raza distinto de blanco (*othracem*).
- Instrumentos: Dos hijas mujeres (*girls*2) y dos hijos varones (*boys*2).

MCO

```
.     reg hourswm morekids agem1 agefstm boy1st blackm hispm othracem, robust

Linear regression                                       Number of obs =  262793
                                                        F(  7,262785) = 2784.79
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.0653
                                                        Root MSE      =  17.766

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |   -5.95623   .0728213   -81.79   0.000    -6.098958   -5.813503
       agem1 |   .8097462    .011472    70.58   0.000     .7872614     .832231
     agefstm |  -1.300151   .0132688   -97.99   0.000    -1.326158   -1.274145
      boy1st |   .0378813   .0693342     0.55   0.585    -.0980118    .1737745
      blackm |   9.631453   .1541123    62.50   0.000     9.329397    9.933509
       hispm |   3.014593   .2337595    12.90   0.000      2.55643    3.472755
    othracem |   5.103861   .2220974    22.98   0.000     4.668556    5.539166
       _cons |   20.70811   .3460431    59.84   0.000     20.02988    21.38635
------------------------------------------------------------------------------
```

Primera etapa

```
.    * First stage
.     reg morekids boys2 girls2 agem1 agefstm boy1st blackm hispm othracem, robust

Linear regression                                      Number of obs =  262793
                                                       F(  8,262784) = 3192.69
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0794
                                                       Root MSE      =  .46659

------------------------------------------------------------------------------
             |               Robust
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       boys2 |    .058245   .0025344    22.98   0.000     .0532776    .0632124
      girls2 |   .0796582   .0026165    30.44   0.000       .07453    .0847865
       agem1 |   .0302071   .0002879   104.91   0.000     .0296427    .0307714
     agefstm |  -.0440882   .0003225  -136.72   0.000    -.0447203   -.0434562
       _cons |   .3374692   .0088555    38.11   0.000     .3201126    .3548258
------------------------------------------------------------------------------

.     test  boys2 girls2
 ( 1)  boys2 = 0
 ( 2)  girls2 = 0

       F(  2,262784) =  727.61
            Prob > F =    0.0000
```

IV con prediccion de la primera etapa

```
ivregress 2sls hourswm (morekids = morekidshat3) agem1 agefstm boy1st blackm hispm othrace
> m, robust

Instrumental variables (2SLS) regression              Number of obs =  262793
                                                      Wald chi2(7)  =11845.37
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0637
                                                      Root MSE      =   17.78

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -4.426483   .9981875    -4.43   0.000    -6.382894   -2.470071
       agem1 |   .7636138   .0321289    23.77   0.000     .7006423    .8265852
     agefstm |  -1.232783   .0457746   -26.93   0.000      -1.3225   -1.143067
      boy1st |   .0519022   .0699756     0.74   0.458    -.0852475    .189052
      blackm |    9.52702   .1684359    56.56   0.000     9.196891    9.857148
       hispm |   2.761302   .2861099     9.65   0.000     2.200536    3.322067
    othracem |   5.010299   .2303647    21.75   0.000     4.558793    5.461806
       _cons |   20.13108   .5105939    39.43   0.000     19.13033    21.13182
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   agem1 agefstm boy1st blackm hispm othracem morekidshat3
```

Dos veces MCO

```
.     reg hourswm morekidshat3 agem1 agefstm boy1st blackm hispm othracem, robust

Linear regression                                     Number of obs =  262793
                                                      F(  7,262785) = 1663.58
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0423
                                                      Root MSE      =  17.982

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
morekidshat3 |  -4.426483   1.009505    -4.38   0.000    -6.405085    -2.44788
       agem1 |   .7636138   .0324912    23.50   0.000       .699932    .8272956
     agefstm |  -1.232783   .0462865   -26.63   0.000    -1.323504   -1.142063
      boy1st |   .0519022   .0707778     0.73   0.463    -.0868204    .1906249
      blackm |    9.52702   .1704203    55.90   0.000        9.193    9.861039
       hispm |   2.761302   .2896682     9.53   0.000      2.19356    3.329043
    othracem |   5.010299   .2325027    21.55   0.000       4.5546    5.465998
       _cons |   20.13108   .5159306    39.02   0.000     19.11987    21.14229
------------------------------------------------------------------------------
```

Mínimos cuadrados en 2 etapas

```
.     ivregress 2sls hourswm (morekids = boys2 girls2) agem1 agefstm boy1st blackm hispm othracem, robust

Instrumental variables (2SLS) regression              Number of obs =  262793
                                                      Wald chi2(7)  =11845.37
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0637
                                                      Root MSE      =   17.78

------------------------------------------------------------------------------
             |               Robust
     hourswm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -4.426483   .9981875    -4.43   0.000    -6.382894   -2.470071
       agem1 |   .7636138   .0321289    23.77   0.000     .7006423    .8265852
     agefstm |  -1.232783   .0457746   -26.93   0.000      -1.3225   -1.143067
      boy1st |   .0519022   .0699756     0.74   0.458    -.0852475     .189052
      blackm |    9.52702   .1684359    56.56   0.000     9.196891    9.857148
       hispm |   2.761302   .2861099     9.65   0.000     2.200536    3.322067
    othracem |   5.010299   .2303647    21.75   0.000     4.558793    5.461806
       _cons |   20.13108   .5105939    39.43   0.000     19.13033    21.13182
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   agem1 agefstm boy1st blackm hispm othracem boys2 girls2
```

Testear competencia imperfecta en el mercado de pescado de Fulton, NY

- Modelo

$$\begin{aligned} ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t \\ + \alpha\, lavgprc_t + u_t, \end{aligned}$$

  donde *lavgprc*: Logaritmo del precio promedio del día ($ por libra), *ltotqty*: Logaritmo de ventas totales del día en libras, $mon = 1$ para Lunes, $tues = 1$ para Martes, $wed = 1$ para Miércoles, $thurs = 1$ para Jueves.

- Instrumentos: *speed*2: Mínimo en los últimos 2 días de la velocidad media del viento, *speed*3: Mínimo hace 3 días de la velocidad media del viento, *wave*2: Máximo en los últimos 2 días de la altura promedio de las olas, *wave*3: Máximo hace 3 y 4 días de la altura promedio de las olas.

MCO

```
.     reg ltotqty lavgprc mon tues wed thurs, robust

Linear regression                                      Number of obs =      97
                                                       F(  5,    91) =    8.63
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2168
                                                       Root MSE      =  .69504

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.5246552    .161579    -3.25   0.002    -.8456121   -.2036984
         mon |  -.3109273   .2445861    -1.27   0.207    -.7967675    .1749129
        tues |  -.6827902   .2044422    -3.34   0.001     -1.08889   -.2766908
         wed |  -.5338937   .2133237    -2.50   0.014    -.9576351   -.1101524
       thurs |   .0672272   .1656234     0.41   0.686    -.2617634    .3962178
       _cons |   8.244317   .1345196    61.29   0.000     7.977111    8.511524
------------------------------------------------------------------------------
```

Primera etapa

```
.     reg lavgprc mon tues wed thurs speed2 speed3 wave2 wave3, robust

Linear regression                                      Number of obs =      97
                                                       F(  8,    88) =    6.98
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3048
                                                       Root MSE      =  .35232

------------------------------------------------------------------------------
             |               Robust
     lavgprc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mon |  -.0090364   .1184382    -0.08   0.939    -.2444074    .2263345
        tues |  -.0141502   .1234927    -0.11   0.909    -.2595659    .2312656
         wed |   .0494936   .1111364     0.45   0.657    -.1713665    .2703538
       thurs |   .1253236    .104251     1.20   0.233    -.0818534    .3325006
      speed2 |   -.002625   .0087642    -0.30   0.765    -.0200421    .0147921
      speed3 |   .0014381    .007503     0.19   0.848    -.0134725    .0163487
       wave2 |    .096806   .0221069     4.38   0.000     .0528733    .1407388
       wave3 |   .0494724   .0220505     2.24   0.027     .0056516    .0932932
       _cons |  -1.017331   .1584286    -6.42   0.000    -1.332175   -.7024874
------------------------------------------------------------------------------
```

Test relevancia de los instrumentos

```
.
.     test speed2 speed3 wave2 wave3

 ( 1)  speed2 = 0
 ( 2)  speed3 = 0
 ( 3)  wave2 = 0
 ( 4)  wave3 = 0

       F(  4,    88) =   10.33
            Prob > F =    0.0000
.     test speed2 speed3

 ( 1)  speed2 = 0
 ( 2)  speed3 = 0

       F(  2,    88) =    0.06
            Prob > F =    0.9416
```

Primera etapa

```
.     reg lavgprc mon tues wed thurs wave2 wave3, robust

Linear regression                                      Number of obs =      97
                                                       F(  6,    90) =    9.31
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3041
                                                       Root MSE      =  .34856

------------------------------------------------------------------------------
             |               Robust
     lavgprc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mon |  -.0120799   .1148977    -0.11   0.917    -.2403442    .2161844
       thurs |   .1241913   .1030767     1.20   0.231    -.0805885     .328971
       wave2 |   .0944805   .0180429     5.24   0.000     .0586352    .1303258
       wave3 |    .052566   .0168191     3.13   0.002     .0191519    .0859801
       _cons |  -1.022801    .136276    -7.51   0.000    -1.293537   -.7520651
------------------------------------------------------------------------------
.     test wave2 wave3

 ( 1)  wave2 = 0
 ( 2)  wave3 = 0

       F(  2,    90) =   20.77
            Prob > F =    0.0000
```

Mínimos cuadrados en 2 etapas

```
Instrumental variables (2SLS) regression              Number of obs =      97
                                                      Wald chi2(5)  =   29.85
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.1933
                                                      Root MSE      =  .68324

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.8158181   .3234293    -2.52   0.012    -1.449728   -.1819083
         mon |  -.3074355   .2374609    -1.29   0.195    -.7728503    .1579793
        tues |  -.6847291   .2005469    -3.41   0.001    -1.077794   -.2916644
         wed |  -.5206142   .2126399    -2.45   0.014    -.9373808   -.1038476
       thurs |   .0947567   .1647731     0.58   0.565    -.2281926     .417706
       _cons |   8.164099   .1569425    52.02   0.000     7.856497    8.471701
------------------------------------------------------------------------------
Instrumented:  lavgprc
Instruments:   mon tues wed thurs wave2 wave3
```

Validez de los instrumentos

Validez de los instrumentos

La consistencia y la distribución asintótica del estimador dependen de la validez de los instrumentos.

- Relevancia: Los instrumentos están (parcialmente) correlados con la variables endógena.
- Exogeneidad: Los instrumentos no están correlados con los inobservados de la ecuación de interés.

Relevancia de los instrumentos

- Condición de rango: $rango(\mathbb{E}(zx')) = K$.
- Para una variable endógena y un instrumento, la condición de rango es equivalente a contrastar si el instrumento es significativo en la primera etapa.
- Para una variable endógena y múltiples instrumentos, la condición de rango es equivalente a contrastar si al menos un instrumento es significativo en la primera etapa.
- Para más de una variable endógena, una condición suficiente es que instrumentos distintos sean significativos para distintas v. endógenas en la primera etapa. Existen contrastes que evalúan si la matriz $\mathbb{E}(zx')$ tiene rango K.
- Contraste de Wald (o F) en la primera etapa H_0 : los coeficientes de los instrumentos son cero, versus H_1 : al menos uno de los coeficientes de los instrumentos es distinto de cero.

Condición de rango

- Podemos escribir los instrumentos de la siguiente forma

$$x^* = \Pi' z,$$

  donde $\Pi = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')$
- Recuerde que Π es una matriz de $L \times K$ donde cada columna corresponde a la primera etapa de cada variable explicativa.
- Podemos contrastar la condición de rango con la matriz Π.

Condición de rango

- Podemos escribir los instrumentos de la siguiente forma

$$x^* = \Pi' z,$$

donde $\Pi = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')$

- Recuerde que Π es una matriz de $L \times K$ donde cada columna corresponde a la primera etapa de cada variable explicativa.
- Podemos contrastar la condición de rango con la matriz Π. Demostración:

$$\mathbb{E}(zx') = \mathbb{E}(z(x^* + e)') = \mathbb{E}(zx^{*'}) = \mathbb{E}(zz')\Pi.$$

Dado que $rango(\mathbb{E}(zz')) = L$, necesitamos $rango(\Pi) = K$.

Ejemplo 1: Modelo simple

- Modelo: $y = \beta_0 + \beta_1 x_1 + u$ con un instrumento z_1. Escribimos la primera etapa como $x_1 = \pi_{10} + \pi_{11} z_1 + e_1$. Entonces

$$\Pi = \begin{bmatrix} 1 & \pi_{10} \\ 0 & \pi_{11} \end{bmatrix}.$$

  La condición de rango es $\pi_{11} \neq 0$.

Ejemplo 2: Regresión múltiple con una variable endógena x_K

- Escribimos primera etapa como $x_K = x'_{-K}\pi_{10} + \pi'_{11}z + e_i$ donde $\pi_{10} = (K-1) \times 1$ y $\pi_{11} = L_1 \times 1$. Entonces

$$\Pi = \begin{bmatrix} I_{K-1} & \pi_{10} \\ 0_{L_1,K-1} & \pi_{11} \end{bmatrix}.$$

  Se necesita por lo menos un elemento de π_{11} distinto de cero.

Contraste de sobreidentificación (Sargan-Hansen)

- Queremos contrastar la condición de exogeneidad:

$$H_0 : \mathbb{E}(zu) = 0,$$
$$H_1 : \mathbb{E}(zu) \neq 0.$$

Contraste de sobreidentificación (Sargan-Hansen)

- Queremos contrastar la condición de exogeneidad:

$$H_0 : \mathbb{E}(zu) = 0,$$
$$H_1 : \mathbb{E}(zu) \neq 0.$$

- Idea: Podemos utilizar

$$\sqrt{N}\frac{1}{N}\sum z_i u_i \xrightarrow{d}_{H_0} \text{Normal}(0, \mathbb{E}(u^2 zz')),$$

para construir el test estadístico

$$J_N = N\left(\frac{1}{N}\sum z_i u_i\right)' \left(\frac{1}{N}\sum u_i^2 z_i z_i'\right)^{-1} \left(\frac{1}{N}\sum z_i u_i\right) \xrightarrow{d}_{H_0} \chi^2_L.$$

Contraste de sobreidentificación (Sargan-Hansen)

- Queremos contrastar la condición de exogeneidad:

$$H_0 : \mathbb{E}(zu) = 0,$$
$$H_1 : \mathbb{E}(zu) \neq 0.$$

- Idea: Podemos utilizar

$$\sqrt{N}\frac{1}{N}\sum z_i u_i \xrightarrow{d}_{H_0} \text{Normal}(0, \mathbb{E}(u^2 zz')),$$

para construir el test estadístico

$$J_N = N\left(\frac{1}{N}\sum z_i u_i\right)' \left(\frac{1}{N}\sum u_i^2 z_i z_i'\right)^{-1} \left(\frac{1}{N}\sum z_i u_i\right) \xrightarrow{d}_{H_0} \chi^2_L.$$

- Problema: No observamos u_i. Podemos reemplazar u_i por $\hat{u}_i$. Si el modelo está exactamente identificado no podemos realizar el contraste porque $\sum z_i \hat{u}_i = 0$. Al estimar $\hat{u}_i$ perdemos K grados de libertad por lo tanto $J_N \xrightarrow{d}_{H_0} \chi^2_{L-K}$.

Discusión

1. Para modelos homoscedásticos, el test se puede realizar con el estadístico $LM_N = N\hat{R}^2$ donde $\hat{R}^2$ es el R^2 estimado en la regresión de $\hat{u}_i$ sobre z_i (incluye las variables exógenas incluidas). Bajo la hipótesis nula $LM_N = N\hat{R}^2 \xrightarrow{d}_{H_0} \chi^2_{L-K}$.

Discusión

1. Para modelos homoscedásticos, el test se puede realizar con el estadístico $LM_N = N\hat{R}^2$ donde $\hat{R}^2$ es el R^2 estimado en la regresión de $\hat{u}_i$ sobre z_i (incluye las variables exógenas incluidas). Bajo la hipótesis nula $LM_N = N\hat{R}^2 \xrightarrow{d}_{H_0} \chi^2_{L-K}$.
2. Interpretación alternativa: Intuitivamente, el test compara las estimaciones de los modelos exactamente identificados. Rechazamos en caso que las estimaciones de los modelos exactamente identificados sean distintas. A veces no rechazamos porque los s.e. son muy grandes.

Discusión

1. Para modelos homoscedásticos, el test se puede realizar con el estadístico $LM_N = N\hat{R}^2$ donde $\hat{R}^2$ es el R^2 estimado en la regresión de $\hat{u}_i$ sobre z_i (incluye las variables exógenas incluidas). Bajo la hipótesis nula $LM_N = N\hat{R}^2 \xrightarrow{d}_{H_0} \chi^2_{L-K}$.
2. Interpretación alternativa: Intuitivamente, el test compara las estimaciones de los modelos exactamente identificados. Rechazamos en caso que las estimaciones de los modelos exactamente identificados sean distintas. A veces no rechazamos porque los s.e. son muy grandes.
3. El test de sobreidentificación no suele ser muy útil en la práctica porque los s.e. son muy grandes y es difícil rechazar H_0. El test tiene poco poder.

Discusión

1. Para modelos homoscedásticos, el test se puede realizar con el estadístico $LM_N = N\hat{R}^2$ donde $\hat{R}^2$ es el R^2 estimado en la regresión de $\hat{u}_i$ sobre z_i (incluye las variables exógenas incluidas). Bajo la hipótesis nula $LM_N = N\hat{R}^2 \xrightarrow{d}_{H_0} \chi^2_{L-K}$.
2. Interpretación alternativa: Intuitivamente, el test compara las estimaciones de los modelos exactamente identificados. Rechazamos en caso que las estimaciones de los modelos exactamente identificados sean distintas. A veces no rechazamos porque los s.e. son muy grandes.
3. El test de sobreidentificación no suele ser muy útil en la práctica porque los s.e. son muy grandes y es difícil rechazar H_0. El test tiene poco poder.
4. En Stata utilizar estat overid

Testear competencia imperfecta en el mercado de pescado de Fulton, NY

- Modelo

$$\begin{aligned} ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t \\ + \alpha\, lavgprc_t + u_t, \end{aligned}$$

  donde *lavgprc*: Logaritmo del precio promedio del día ($ por libra), *ltotqty*: Logaritmo de ventas totales del día en libras, $mon = 1$ para Lunes, $tues = 1$ para Martes, $wed = 1$ para Miércoles, $thurs = 1$ para Jueves.

- Instrumentos: *speed*2: Mínimo en los últimos 2 días de la velocidad media del viento, *speed*3: Mínimo hace 3 días de la velocidad media del viento, *wave*2: Máximo en los últimos 2 días de la altura promedio de las olas, *wave*3: Máximo hace 3 y 4 días de la altura promedio de las olas.

Test de sobreidentificación con *wave2* y *wave3*

```
.         estat overid

  Test of overidentifying restrictions:

  Score chi2(1)           =  .026179  (p = 0.8715)
```

Test de sobreidentificación a mano

```
.     ivregress 2sls ltotqty (lavgprc=wave2 wave3) mon tues wed thurs, robust
.         predict uhat, resid
.         reg uhat wave2 wave3 mon tues wed thurs, robust

Linear regression                                   Number of obs =      97
                                                    F(  6,    90) =    0.00
                                                    Prob > F      =  1.0000
                                                    R-squared     =  0.0003
                                                    Root MSE      =  .70921

------------------------------------------------------------------------------
             |               Robust
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       wave2 |  -.0048563   .0485799    -0.10   0.921    -.1013687    .0916562
       wave3 |   .0058697   .0390458     0.15   0.881    -.0717017    .0834411
         mon |   .0028326   .2455883     0.01   0.991    -.4850714    .4907366
        tues |  -.0025088   .2095872    -0.01   0.990    -.4188904    .4138728
         wed |  -.0056111   .2224295    -0.03   0.980     -.447506    .4362839
       thurs |  -.0032378   .1677995    -0.02   0.985    -.3366008    .3301251
       _cons |  -.0029474   .2657641    -0.01   0.991    -.5309343    .5250395
------------------------------------------------------------------------------
```

$$LM_N = N \times \hat{R}^2 = 97 \times 0.0003 = 0.0291$$

Test de sobreidentificación: intuición

```
.     ivregress 2sls ltotqty (lavgprc=wave2) mon tues wed thurs, robust

Instrumental variables (2SLS) regression              Number of obs =      97
                                                      Wald chi2(5)  =   26.21
                                                      Prob > chi2   =  0.0001
                                                      R-squared     =  0.1891
                                                      Root MSE      =  .68503

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.8410206   .3827023    -2.20   0.028    -1.591103   -.0909378
         mon |  -.3071333   .2374899    -1.29   0.196     -.772605    .1583385
        tues |  -.6848969   .2011967    -3.40   0.001    -1.079235   -.2905587
         wed |  -.5194648   .2131435    -2.44   0.015    -.9372184   -.1017111
       thurs |   .0971396   .1670053     0.58   0.561    -.2301848     .424464
       _cons |   8.157156   .1650797    49.41   0.000     7.833606    8.480706
------------------------------------------------------------------------------
Instrumented:  lavgprc
Instruments:   mon tues wed thurs wave2
```

Test de sobreidentificación: intuición

```
.         ivregress 2sls ltotqty (lavgprc=wave3) mon tues wed thurs, robust

Instrumental variables (2SLS) regression              Number of obs =      97
                                                      Wald chi2(5)  =   27.69
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.2013
                                                      Root MSE      =  .67983

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.7610668   .4245699    -1.79   0.073    -1.593209    .0710749
         mon |  -.3080921   .2375451    -1.30   0.195    -.7736719    .1574877
        tues |  -.6843645   .1993474    -3.43   0.001    -1.075078   -.2936508
         wed |  -.5231113    .212618    -2.46   0.014    -.9398351   -.1063876
       thurs |     .08958   .1649782     0.54   0.587    -.2337713    .4129313
       _cons |   8.179184   .1779939    45.95   0.000     7.830322    8.528045
------------------------------------------------------------------------------
Instrumented:  lavgprc
Instruments:   mon tues wed thurs wave3
```

Contraste de endogeneidad (Durbin-Wu-Hausman)

- Queremos contrastar que la variable x_K es endógena:

$$H_0 : \mathbb{E}(x_K u) = 0,$$
$$H_1 : \mathbb{E}(x_K u) \neq 0.$$

Contraste de endogeneidad (Durbin-Wu-Hausman)

- Queremos contrastar que la variable x_K es endógena:

$$H_0 : \mathbb{E}(x_K u) = 0,$$
$$H_1 : \mathbb{E}(x_K u) \neq 0.$$

- Idea: Hausman (1978) propuso comparar el estimador por MCO, consistente bajo H_0, con el estimador IV, consistente bajo H_0 y H_1. Se puede demostrar que

$$\sqrt{N}(\hat{\beta}_{IV} - \hat{\beta}_{MCO}) \xrightarrow{d} \text{Normal}(0, \text{AVar}(\sqrt{N}\,\hat{\beta}_{IV}) - \text{AVar}(\sqrt{N}\,\hat{\beta}_{MCO})),$$

para construir el test estadístico

$$H_N = (\hat{\beta}_{1,IV} - \hat{\beta}_{1,MCO})'(\text{AVar}(\hat{\beta}_{1,IV}) - \text{AVar}(\hat{\beta}_{1,MCO}))^{-1}(\hat{\beta}_{1,IV} - \hat{\beta}_{1,MCO}) \xrightarrow{d}_{H_0} \chi^2_{L_1}.$$

Contraste de endogeneidad (Durbin-Wu-Hausman)

- Queremos contrastar que la variable x_K es endógena:

$$H_0 : \mathbb{E}(x_K u) = 0,$$
$$H_1 : \mathbb{E}(x_K u) \neq 0.$$

- Idea: Hausman (1978) propuso comparar el estimador por MCO, consistente bajo H_0, con el estimador IV, consistente bajo H_0 y H_1. Se puede demostrar que

$$\sqrt{N}(\hat{\beta}_{IV} - \hat{\beta}_{MCO}) \xrightarrow{d} \text{Normal}(0, \text{AVar}(\sqrt{N}\,\hat{\beta}_{IV}) - \text{AVar}(\sqrt{N}\,\hat{\beta}_{MCO})),$$

para construir el test estadístico

$$H_N = (\hat{\beta}_{1,IV} - \hat{\beta}_{1,MCO})'(\text{AVar}(\hat{\beta}_{1,IV}) - \text{AVar}(\hat{\beta}_{1,MCO}))^{-1}(\hat{\beta}_{1,IV} - \hat{\beta}_{1,MCO}) \xrightarrow{d}_{H_0} \chi^2_{L_1}.$$

- En Stata se puede utilizar estat endog.

Discusión

1. El test se puede realizar con el siguiente procedimiento:
   1.1 Estima la primera etapa por MCO y guarda los residuos,
   1.2 Estima la ecuación de interés por MCO controlando por los residuos obtenidos en *1)*,
   1.3 Hacer un contraste de Wald sobre los residuos.

Discusión

1. El test se puede realizar con el siguiente procedimiento:
   1.1 Estima la primera etapa por MCO y guarda los residuos,
   1.2 Estima la ecuación de interés por MCO controlando por los residuos obtenidos en *1)*,
   1.3 Hacer un contraste de Wald sobre los residuos.
2. Es una forma alternativa de estimar los parámetros en forma consistente: enfoque de funciones de control. Una vez que controlamos por los residuos de la primera etapa, las variables explicativas no están correladas con los inobservados y podemos estimar por MCO.

Discusión

1. El test se puede realizar con el siguiente procedimiento:
   1.1 Estima la primera etapa por MCO y guarda los residuos,
   1.2 Estima la ecuación de interés por MCO controlando por los residuos obtenidos en *1)*,
   1.3 Hacer un contraste de Wald sobre los residuos.
2. Es una forma alternativa de estimar los parámetros en forma consistente: enfoque de funciones de control. Una vez que controlamos por los residuos de la primera etapa, las variables explicativas no están correladas con los inobservados y podemos estimar por MCO.
3. El test de endogeneidad no suele ser muy útil en la práctica porque los s.e. son muy grandes y es difícil rechazar H_0. El test tiene bajo poder.

Testear competencia imperfecta en el mercado de pescado de Fulton, NY

- Modelo

$$\begin{aligned} ltotqty_t = \beta_0 + \beta_1\, mon_t + \beta_2\, tues_t + \beta_3\, wed_t + \beta_4\, thurs_t \\ + \alpha\, lavgprc_t + u_t, \end{aligned}$$

  donde *lavgprc*: Logaritmo del precio promedio del día ($ por libra), *ltotqty*: Logaritmo de ventas totales del día en libras, $mon = 1$ para Lunes, $tues = 1$ para Martes, $wed = 1$ para Miércoles, $thurs = 1$ para Jueves.

- Instrumentos: *speed*2: Mínimo en los últimos 2 días de la velocidad media del viento, *speed*3: Mínimo hace 3 días de la velocidad media del viento, *wave*2: Máximo en los últimos 2 días de la altura promedio de las olas, *wave*3: Máximo hace 3 y 4 días de la altura promedio de las olas.

Test de endogeneidad con *wave2* y *wave3*

```
.     estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Robust score chi2(1)            =  1.10541  (p = 0.2931)
  Robust regression F(1,90)       =  1.10986  (p = 0.2949)
```

Test de endogeneidad a mano

```
.     reg lavgprc wave2 wave3 mon tues wed thurs,  robust

Linear regression                               Number of obs =      97
                                                F(  6,    90) =    9.31
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3041
                                                Root MSE      =  .34856

------------------------------------------------------------------------------
             |               Robust
     lavgprc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       wave2 |   .0944805   .0180429     5.24   0.000     .0586352    .1303258
       wave3 |    .052566   .0168191     3.13   0.002     .0191519    .0859801
         mon |  -.0120799   .1148977    -0.11   0.917    -.2403442    .2161844
        tues |  -.0089758   .1221166    -0.07   0.942    -.2515817     .23363
         wed |   .0505471   .1113689     0.45   0.651    -.1707068    .2718009
       thurs |   .1241913   .1030767     1.20   0.231    -.0805885     .328971
       _cons |  -1.022801    .136276    -7.51   0.000    -1.293537   -.7520651
------------------------------------------------------------------------------

.     predict uhat, resid
```

Test de endogeneidad a mano

```
.     reg ltotqty lavgprc mon tues wed thurs uhat,  robust

Linear regression                               Number of obs =      97
                                                F(  6,    90) =    7.27
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2268
                                                Root MSE      =  .69442

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.8158181   .3172522    -2.57   0.012    -1.446095   -.1855414
         mon |  -.3074355   .2479658    -1.24   0.218    -.8000628    .1851918
        tues |  -.6847291   .1993719    -3.43   0.001    -1.080816    -.288642
         wed |  -.5206142   .2130994    -2.44   0.017    -.9439733   -.0972551
       thurs |   .0947567   .1658049     0.57   0.569    -.2346437    .4241572
        uhat |   .4147442   .3936833     1.05   0.295    -.3673763    1.196865
       _cons |   8.164099   .1546984    52.77   0.000     7.856764    8.471435
------------------------------------------------------------------------------
```

Mínimos cuadrados en 2 etapas

```
Instrumental variables (2SLS) regression              Number of obs =      97
                                                      Wald chi2(5)  =   29.85
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.1933
                                                      Root MSE      =  .68324

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.8158181   .3234293    -2.52   0.012    -1.449728   -.1819083
         mon |  -.3074355   .2374609    -1.29   0.195    -.7728503    .1579793
        tues |  -.6847291   .2005469    -3.41   0.001    -1.077794   -.2916644
         wed |  -.5206142   .2126399    -2.45   0.014    -.9373808   -.1038476
       thurs |   .0947567   .1647731     0.58   0.565    -.2281926     .417706
       _cons |   8.164099   .1569425    52.02   0.000     7.856497    8.471701
------------------------------------------------------------------------------
Instrumented:  lavgprc
Instruments:   mon tues wed thurs wave2 wave3
```

Test de endogeneidad a mano

```
. test uhat

 ( 1)  uhat = 0

       F(  1,    90) =    1.11
            Prob > F =    0.2949
```