Econometría Aplicada Avanzada
Estimación con Variables Instrumentales

César Mora Ruiz

Q-Lab
PUCP

Enero de 2024

Estructura de la clase

- El problema de endogeneidad
- Causas de la endogeneidad
- El uso de variables instrumentales
- Mínimos Cuadrados en dos etapas (MC2E)
- Validez de los instrumentos
- Aplicación práctica en Stata

El problema de endogeneidad

El problema de endogeneidad

Consideremos el modelo general de regresión con “k” variables explicativas:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + u_i$$

Suponiendo que la variable X_j es una que no cumple con el supuesto de exogeneidad, entonces se cumplirá lo siguiente:

- $cov(X_{ji}, u_i) \neq 0$
- $\mathrm{E}(u_i | X_{ji}) \neq 0$
- $\mathrm{E}(X_{ji} u_i) \neq 0$

El problema de endogeneidad

Y considerando el modelo escrito de forma compacta (matricial):

$$Y = X\beta + U$$

La estimación del vector de coeficientes $\widehat{\beta}$ queda determinada por:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Y al reemplazar la definición de Y, en esta expresión, entonces:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
\hat{\beta} &= (X'X)^{-1}X'(X\beta + U) \\
\hat{\beta} &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'U \\
\hat{\beta} &= \beta + (X'X)^{-1}X'U
\end{aligned}$$

El problema de endogeneidad

Al tomar valor esperado a dicha expresión, entonces:

$$E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'U]$$
$$E(\hat{\beta}) = \beta + (X'X)^{-1}\underbrace{E(X'U)}_{\neq 0}$$

- Se logra concluir que ante la presencia de endogeneidad, el estimador $\widehat{\beta}$ ya no es insesgado.

- En ese sentido, si al menos una variable explicativa del modelo tiene el problema de endogeneidad, entonces la estimación del coeficiente será sesgada.

Las causas de la endogeneidad

Causas de la endogeneidad

- La endogeneidad es un problema que suele estar presente en la mayoría de aplicaciones prácticas de la Econometría.

- Por ejemplo, considerando el siguiente modelo:

$$Salarios_i = \alpha + \beta_1(educación)_i + \beta_2(sexo)_i + u_i$$

La variable explicativa de ***educación*** es endógena, pues depende de diversos factores asociados al nivel socioeconómico de la persona. Entonces, si estimamos el vector β a través de MCO sin considerar este problema, vamos a obtener un estimador sesgado

Causas de la endogeneidad

Entre las principales causas de la endogeneidad podemos enumerar:

1. Error en la medición de variables
2. Causalidad simultánea
3. Variables omitidas correlacionadas con otras explicativas
4. Especificación incorrecta de la forma funcional

1. Error de medición

- Este problema sucede cuando no se puede observar una variable directamente, y se aproxima mediante otra.
- Por ejemplo, si quisiéramos utilizar a la "habilidad matemática" como una variable explicativa, pero esta no es medible, e intentamos aproximarla utilizando el número de años de educación de la persona.

- Una representación del modelo real que quisiéramos estimar sería:

$$Y_i = X_i^* \beta + v_i, \qquad asumiendo\ que \quad v_i \sim N(0, \sigma_v^2)$$

pero no tenemos información directa sobre X_i^* , y aproximamos dicha variable mediante X_i :

$$X_i = X_i^* + \varepsilon_i$$
$$X_i - \varepsilon_i = X_i^*, \qquad asumiendo\ que: \ \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

El modelo real se convierte en:

$$Y_i = \beta X_i^* + v_i$$
$$Y_i = \beta (X_i - \varepsilon_i) + v_i$$
$$Y_i = \beta X_i + (v_i - \beta \varepsilon_i)$$
$$Y_i = \beta X_i + (u_i)$$

1. Error de medición

- Esto genera un problema ya que $u_i = (v_i - \beta\varepsilon_i)$, y sucederá lo siguiente:

$$\Rightarrow u_i = (v_i - \beta\varepsilon_i)$$
$$\Rightarrow \mathrm{cov}(X_i, u_i) = \mathrm{cov}(X_i, v_i - \beta\varepsilon_i)$$
$$\mathrm{cov}(X_i, u_i) = \mathrm{cov}(X_i^* + \varepsilon_i, v_i - \beta\varepsilon_i)$$
$$\mathrm{cov}(X_i, u_i)$$
$$= \mathrm{cov}(X_i^*, v_i) + \mathrm{cov}(X_i^*, -\beta\varepsilon_i) + \mathrm{cov}(\varepsilon_i, v_i) + \mathrm{cov}(\varepsilon_i, -\beta\varepsilon_i)$$
$$\mathrm{cov}(X_i, u_i) = -\beta\sigma_{\varepsilon_i}^2 \neq 0$$

- Desembocando claramente en endogeneidad del vector de variables explicativas

2. Causalidad simultánea

- Sucede cuanto tenenos un sistema de ecuaciones, y en una ecuación una variable es dependiente, mientras que en otra(s) ecuación(es), la misma variable es explicativa.

- Consideremos el siguiente modelo como ejemplo:

$$Y_t = C_t + I_t \;\; \dots (i)$$
$$C_t = \beta_1 + \beta_2 Y_t + u_t \;\; \dots (ii)$$

- Reemplazando (ii) en (i) obtenemos:

$$\Rightarrow Y_t = (\beta_1 + \beta_2 Y_t + u_t) + I_t$$
$$(1-\beta_2)Y_t = (\beta_1 + u_t) + I_t$$
$$Y_t = \frac{(\beta_1 + I_t) + u_t}{(1-\beta_2)}$$

- Por lo que $\operatorname{cov}(Y_t, u_t) = \dfrac{\sigma^2}{1-\beta_2} \neq 0$ genera que no se cumpla el supuesto de exogeneidad para la ecuación (ii)

3. Variables omitidas correlacionadas

- Considere el siguiente modelo completo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Sin embargo, supongamos que omitimos X_2 del modelo y solo estimamos el modelo reducido:

$$Y_i = \beta_0 + \beta_1 X_{1i} + (\beta_2 X_{2i} + u_i)$$
$$Y_i = \beta_0 + \beta_1 X_{1i} + w_i$$

- En este caso si es que la variable omitida X_2 está correlacionada con la observada entonces:

$$\text{cov}(X_{1i}, w_i) = \beta_2 \,\text{cov}(X_{1i}, X_{2i}) \neq 0$$

4. Especificación incorrecta de la forma funcional

Z_i

- Considere que el modelo correcto es:

$$Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + u_i, \qquad donde \quad E(u|X, X^2) = 0$$

- Sin embargo, se estima la siguiente especificación incompleta:

$$Y_i = \beta_0 + \beta_1 X_i + v_i, \qquad donde \quad v_i = \beta_2 {X_i}^2 + u_i$$

- Entonces no se cumplirá el supuesto: $\mathrm{cov}(X_i, v_i) = 0$

Ya que: $cov(X_i, v_i) = cov(X_i, \beta_2 {X_i}^2 + u_i) \neq 0$

Y se presentará un **sesgo de estimación.**

El uso de variables instrumentales

Variables instrumentales - Definición

Se puede definir una variable llamada **variable instrumental** o **instrumento** definido como Z_i tal que cumpla dos condiciones:

1. **Exogeneidad:** No debe estar correlacionada con u_i

2. **Relevancia:** Debe estar correlacionada con el vector X_i

Validez del Instrumento:
Entonces, Z_i será considerado como **instrumento válido si:**

1. **Exogeneidad:** $\operatorname{cov}(Z_i, u_i) = 0$

2. **Relevancia:** $\operatorname{cov}(Z_i, X_i) \neq 0$

- Lo anterior, también aplica para el caso en el que Z_i sea un vector de variables instrumentales

Regresión por variables instrumentales

Considerando el modelo:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + u_i$$

$$Y_i = X\beta + u_i$$

Y definiendo X_i matricialmente:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

Regresión por variables instrumentales

- Procedemos a definir Z_i como una matriz de variables instrumentales, con un solo instrumento, tal que:

$$Z = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k-1,1} & z_{11} \\ 1 & x_{12} & x_{22} & \cdots & x_{k-1,2} & z_{12} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{k-1,n} & z_{1n} \end{bmatrix}$$

- En la que se cumplirán las dos condiciones previamente presentadas para **un instrumento válido**

$$\operatorname{cov}(Z_i, u_i) = 0$$

$$\operatorname{cov}(Z_i, X_i) \neq 0$$

Regresión por variables instrumentales

- En el modelo de forma compacta $Y_i = X\beta + u_i$ procedemos a premultiplicar por la matriz Z'

$$
\begin{aligned}
&Z'Y = Z'X\beta + Z'u \\
&E(Z'Y) = E(Z'X\beta) + E(Z'u) \\
&E(Z'Y) = E(Z'X)\beta
\end{aligned}
$$

$$E(Z'X)^{-1}E(Z'Y) = \beta_{VI}$$

- Y finalmente: $\beta_{VI} = (Z'X)^{-1}(Z'Y)$

- Es posible demostrar que este estimador es insesgado

Insesgadez del estimador de V.I.

- Para demostrar la insesgadez, tomamos el estimador y reemplazamos la variable Y_i :

$$\beta_{VI} = (Z'X)^{-1}(Z'Y)$$
$$\beta_{VI} = (Z'X)^{-1}[Z'(X\beta + u)]$$
$$\beta_{VI} = (Z'X)^{-1}[Z'X\beta + (Z'u)]$$
$$\beta_{VI} = (Z'X)^{-1}(Z'X)\beta + (Z'X)^{-1}(Z'u)$$
$$\beta_{VI} = \beta + (Z'X)^{-1}(Z'u)$$

- Al tomar el valor esperado:

$$E[\beta_{VI}] = E[\beta] + E[(Z'X)^{-1}(Z'u)]$$
$$E[\beta_{VI}] = \beta + (Z'X)^{-1}\boldsymbol{E(Z'u)}$$

$$E[\beta_{VI}] = \beta$$

Regresión por variables instrumentales

- Una manera alternativa de obtener el estimador de variables instrumentales, utilizando términos de covarianzas entre variables, es la siguiente.

- Partimos del modelo lineal simple $Y_i = X\beta + u_i$

- Y considerando que el instrumento cumple con $\text{cov}(Z_i, u_i) = 0$ y con $\text{cov}(Z_i, X_i) \neq 0$, entonces:

$$cov(Z_i, Y_i) = cov(Z_i, X\beta + u_i)$$
$$cov(Z_i, Y_i) = \beta cov(Z_i, X) + \boldsymbol{cov(Z, u)}$$

$$\beta = \frac{cov(Z_i, Y)}{cov(Z_i, X)}$$

- Volveremos posteriormente para analizar esta expresión, que nos permitirá identificar la relevancia del instrumento

Mínimos cuadrados en dos etapas

Mínimos cuadrados en dos etapas (MC2E)

- Ahora exploraremos la metodología de Mínimos cuadrados en dos etapas para obtener el mismo estimador insesgado

- Asumimos nuevamente que la variable explicativa endógena es X_{ki}, y el instrumento que utilizaremos es Z_1.

- El modelo a estimar es: $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + u_i$

Primera etapa:

$$X_{ki} = \delta_0 + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \theta Z_1 + \varepsilon_i$$

- De donde obtenemos el estimado:

$$\hat{X}_{ki} = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \ldots + \hat{\delta}_{k-1} X_{k-1} + \hat{\theta} Z_1$$

Mínimos cuadrados en dos etapas (MC2E)

Segunda etapa:

- Estimamos el modelo $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + u_i$

  pero utilizando el estimado $\hat{\boldsymbol{X}}_{\boldsymbol{ki}}$ en vez de $X_{k,i}$

- De este último procedimiento, obtendremos el estimador de interés, el cual tiene la forma:

$$\hat{\beta}^{MC2E} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

Donde:

$$\hat{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k-1,1} & \hat{x}_{k1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{k-1,n} & \hat{x}_{kn} \end{bmatrix}$$

MC2E: una endógena y "m" instrumentos

Primera etapa:

- Realizar la estimación para $X_{ki} = \delta_0 + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \theta_1 Z_1 + \ldots + \theta_m Z_m + \varepsilon_i$

- El estimador será: $\hat{X}_{ki} = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \ldots + \hat{\delta}_{k-1} X_{k-1} + \hat{\theta}_1 Z_1 + \ldots + \hat{\theta}_m Z_m$

Segunda etapa:

- En la ecuación original a estimar, reemplazaremos el estimado $\hat{X}_{ki}$ de la primera etapa con los "m" instrumentos y las "k-1" explicativas exógenas.

- El estimado será:

$$\hat{\beta}^{MC2E} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

MC2E: “L” endógenas y “m” instrumentos

- En el caso más general, asumiremos que de las “K” variables explicativas, “L” son endógenas, y el resto “r” son exógenas. De este modo, entonces **L+r=K**

- Representaremos el modelo de manera alternativa como:

$$Y_i = \beta_0 + \boldsymbol{\beta_1 X_{1,i}} + \ldots + \boldsymbol{\beta_L X_{L,i}} + \lambda_1 w_{1,i} + \cdots + \lambda_r w_{r,i} + u_i$$

Primera etapa:

- Suponiendo que contamos con “m” instrumentos, entonces la estimación de la matriz X_{ij} con **L** endógenas quedará denotada por:

$$X_{j,i} = \delta_0 + \delta_1 w_1 + \cdots + \delta_r w_r + \theta_1 Z_1 + \cdots + \theta_m Z_m + \epsilon_i$$

Donde:
- j=1,2,…L
- i=1,2…n

MC2E: “L” endógenas y “m” instrumentos

Primera etapa:

- Entonces obtendremos:

$$\hat{X}_{j,i} = \widehat{\delta_0} + \widehat{\delta_1} w_1 + \cdots + \widehat{\delta_r} w_r + \widehat{\boldsymbol{\theta_1}} \boldsymbol{Z_1} + \cdots + \widehat{\boldsymbol{\theta_m}} \boldsymbol{Z_m}$$

Segunda etapa:

- En la ecuación original a estimar, reemplazaremos la matriz $\hat{X}_{j,i}$, la cual contiene a todas las variables explicativas estimadas en la primera etapa con los “m” instrumentos y las “r” explicativas exógenas:

$$\hat{\beta}^{MC2E} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

Identificación del modelo

Dependiendo del número de explicativas endógenas y variables instrumentales disponibles, el modelo podrá estar:

- **Exactamente identificado** si **m=L**
- **Sobreidentificado** si **m>L:** En este caso es necesario testear la validez de los instrumentos, para solo quedarnos con los relevantes
- **Sub-identificado** si **m<L:** hay insuficientes instrumentos, por lo que no se podrá estimar el vector de coeficientes.

Análisis de validez de los instrumentos

Validez de los instrumentos

Relevancia:

- Para evaluar la exogeneidad contamos con pruebas tales como el **Test J de restricciones de sobreidentificación**, y el **Test de Hausman**
- En ambos casos se evalúa la hipótesis nula de exogeneidad (es decir que todos los instrumentos son exógenos) contra la hipótesis alternativa de que al menos un instrumento es endógeno

Validez de los instrumentos

Exogeneidad: $\text{cov}(Z_i, X_i) \neq 0$

- Debe existir relación entre los instrumentos y las explicativas endógenas, pues recordar que:

$$\hat{X}_{j,i} = \widehat{\delta_0} + \widehat{\delta_1} w_1 + \cdots + \widehat{\delta_r} w_r + \boldsymbol{\widehat{\theta_1} Z_1 + \cdots + \widehat{\theta_m} Z_m}$$

- Los instrumentos son **relevantes**, si los coeficientes de $\hat{\theta}$ son distintos a cero $(\theta \neq 0)$
- Los instrumentos son **débiles,** si los coeficientes de $\hat{\theta}$ son cercanos a cero $(\theta \approx 0)$
- Recordar que $\beta = \frac{cov(Z_i, Y)}{cov(Z_i, X)}$, por lo que si $\text{cov}(Z_i, X_i) \approx 0$, entonces la estimación de β no será posible

Consecuencias de los instrumentos débiles

1. Teniendo en cuenta que $Var(\beta_{VI}) = \frac{Var(u_i)}{N*Var(X_{k,i})cov(Z_i,X_i)}$, entonces si el instrumento es débil, $\mathrm{cov}(Z_i, X_i) \approx 0$, se refleja que la $Var(\beta_{VI})$, será mayor, afectando la eficiencia del estimador.

2. Sesgo asintótico en las estimaciones:

$$plim\beta_1^{VI} = \frac{cov(z,y)}{cov(z,x)} = \frac{cov(z,\beta_0 + \beta_1 x + u)}{cov(z,x)}$$

$$plim\beta_1^{VI} = \beta_1 + \frac{\boldsymbol{cov(z,u)}}{\boldsymbol{cov(z,x)}}$$

Aplicación práctica en Stata

Aplicación práctica en Stata

- Utilizaremos un ejemplo inspirado en el paper de David Card **"Using geographic variation in college proximity to estimate the return of schooling" (1993)**

- En este documento el autor busca identificar el impacto de los años de educación sobre los salarios de los individuos:

$$\log(W_i) = \beta_0 + \beta_1 * (educación_i) + \lambda X_i + u_i$$

- Recordemos que el nivel educativo es una variable endógena que depende de otras variables (ingreso de la familia, nivel educativo de los padres, etc)

- En ese sentido hace falta encontrar una variable instrumental **independiente y relevante** para tratar este problema

- Además dicha VI **solamente debe afectar a los salarios a través de su efecto sobre la escolaridad**, mas no directamente (supuesto de exogeneidad)

Aplicación práctica en Stata

- Card utiliza a la **proximidad** de una escuela en la region en la que creció la persona como instrumento de la variable explicativa endógena de **educación**, pues se aprecia que dichas personas, en promedio, tienen mayores niveles educativos que las que crecieron en lugares con escuelas lejanas.

- Una mayor proximidad del centro de estudios **no garantiza** el incremento de la escolaridad, pero sí incrementa la probabilidad de asistir a clases en comparación a un estudiante que vivió en una comunidad con escuela más alejada, y debió invertir mayor tiempo (y dinero) transportándose a la escuela.

Condiciones:

1. **Independencia:**

La presencia de la escuela no está asociada directamente con los salarios. Solo afecta a estos a través de la variable endógena de escolaridad

2. Relevancia:

La presencia de la escuela afecta la probabilidad de incrementar los años de escolaridad

Aplicación práctica en Stata

Regresión MCO: $\log(W_i) = \beta_0 + \beta_1 * (educación_i) + u_i$

```
. reg lwage educ,robust

Linear regression                               Number of obs     =      3,010
                                                F(1, 3008)        =     321.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0987
                                                Root MSE          =     .42139

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0520942   .0029069    17.92   0.000     .0463946    .0577939
       _cons |   5.570882   .0390935   142.50   0.000      5.49423    5.647535
------------------------------------------------------------------------------
```

Aplicación práctica en Stata

Regresión con variables instrumentales (mostrando cada etapa)

Primera etapa: $educación_i = \alpha_0 + \alpha_1(nearc4) + \epsilon_i$, de donde obtendremos $\widehat{educación}$

```
. ivregress 2sls lwage (educ=nearc4), first robust

First-stage regressions
-----------------------

                                                Number of obs     =      3,010
                                                F(   1,   3008)   =      60.37
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0208
                                                Adj R-squared     =     0.0205
                                                Root MSE          =     2.6494

------------------------------------------------------------------------------
             |               Robust
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc4 |    .829019   .1066941     7.77   0.000     .6198182     1.03822
       _cons |   12.69801   .0902199   140.75   0.000     12.52112    12.87491
------------------------------------------------------------------------------
```

Aplicación práctica en Stata

Regresión con variables instrumentales (mostrando cada etapa)

Segunda etapa: $\log(W_i) = \beta_0 + \beta_1 * (\widehat{educación_i}) + u_i$

```
Instrumental variables (2SLS) regression          Number of obs   =      3,010
                                                   Wald chi2(1)    =      51.78
                                                   Prob > chi2     =     0.0000
                                                   R-squared       =          .
                                                   Root MSE        =     .55667

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1880626   .0261339     7.20   0.000     .1368412    .2392841
       _cons |   3.767472   .3466268    10.87   0.000     3.088096    4.446848
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   nearc4
```