Estimación por variables instrumentales (IV)

Gustavo A. García

ggarci24@eafit.edu.co

Econometría II

Programa de Economía

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Motivación
- Causas
- El estimador por variables instrumentales
- Múltiples instrumentos: mínimos cuadrados en dos etapas (MC2E - *2SLS*)
- 2SLS: el caso general
- Expresión general para el estimador 2SLS
- Ejercicio aplicado en R: efectos de la educación de la mujer sobre la fertilidad

Lecturas

- Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2a edición. MA: MIT Press. Cap 5

- Greene, W. H. (2018). *Econometric Analysis*. 8th ed. NY: Pearson. Cap 8

Motivación

Recordemos que el modelo de RLM tiene la siguiente estructura y supuestos:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- Modelo completo: $E(\mathbf{u}) = \mathbf{0}$
- Exogeneidad: $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$
- Perturbaciones esféricas: $Cov(\mathbf{u}) = E(\mathbf{u}\mathbf{u}') = \sigma_u^2\mathbf{I}_n$ (homocedasticidad y no autocorrelación)
- No multicolinealidad perfecta: $\rho(\mathbf{X}_{n\text{x}k}) = k < n$
- Normalidad: $\mathbf{u}_{n\text{x}1} \sim \mathbf{N}(\mathbf{0}_{n\text{x}1}, \sigma_u^2\mathbf{I}_n)$

El estimador por MCO es:

$$\widehat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Ahora, la existencia de correlación entre los regresores y las perturbaciones estocásticas del modelo genera sesgo e inconsistencia en los parámetros estimados por MCO:

- Para que $E(\widehat{\boldsymbol{\beta}}_{MCO}) = \boldsymbol{\beta}$ se requiere que $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$
- Para que $plim(\widehat{\boldsymbol{\beta}}_{MCO}) = \boldsymbol{\beta}$ se requiere que $plim(N^{-1}\mathbf{X}'\mathbf{u}) = 0$

Motivación

Como existe correlación entre $\mathbf{u}$ y $\mathbf{X}$ entonces se genera un problema de identificación, ya que $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \boldsymbol{\beta} + \frac{\partial \mathbf{u}}{\partial \mathbf{X}}$, lo que implica que la estimación por MCO será $\boldsymbol{\beta} + \frac{\partial \mathbf{u}}{\partial \mathbf{X}}$ y no $\boldsymbol{\beta}$

Causas

En econometría aplicada, la endogeneidad puede usualmente venir por las siguientes causas:

- Omisión de variables relevantes
- Errores de medida en los regresores
- Simultaneidad
- Sesgo de selección
- Modelos dinámicos y perturbaciones autocorrelacionadas

Causas: omisión de variables relevantes

Se está interesado en la siguiente esperanza condicional $E(y|\mathbf{x}, q)$, pero q no es observable

No habría problema si q y $\mathbf{x}$ no estuvieran correlacionados, pero ya que q no es observable haría parte del término de error y, por tanto, $\mathbf{x}$ y u estarán correlacionados a través de q, es decir, que $\mathbf{x}$ sería endógena

Ejemplo: El salario de un individuo está dado por la siguiente ecuación:

$$w_i = \beta_1 + \beta_2 Educ_i + \beta_3 Habil_i + u_i$$

donde

$$Educ_i = \alpha_1 + \alpha_2 Habil_i + \epsilon_i$$

w_i : salario; $Educ_i$: nivel educativo; $Habil_i$: habilidad (variable no observable)

Dado que $Habil_i$ es no observable se termina estimando el siguiente modelo:

$$w_i = \beta_1 + \beta_2 Educ_i + \eta_i$$

así que, $\eta_i = \beta_3 Habil_i + u_i$, lo que implica que

$$E(Educ_i \eta_i) = E((\alpha_1 + \alpha_2 Habil_i + \epsilon_i)(\beta_3 Habil_i + u_i)) \neq 0$$

Causas: errores de medida en los regresores

En este caso nos gustaría medir el efecto de una variable x_k^*, pero sólo se puede observar una medida imperfecta de ésta, x_k. Cuando se utiliza como regresor esta medida imperfecta, la medida de error se transmitiría a los errores. En la medida que x_k^* y x_k estén más correlacionadas o no, x_K será o no endógena

Ejemplo: Volviendo sobre la ecuación de salarios del anterior ejemplo:

$$w_i = \beta_1 + \beta_2 Educ_i + u_i$$

La educación $(Educ_i)$ es una variable que no se puede medir completamente y los años de educación es típicamente la mejor variable proxy disponible. Entonces los años de escolaridad $(Schooling_i)$ que es lo que observamos viene determinada por:

$$Schooling_i = Educ_i + \epsilon_i$$

donde ϵ_i es la medida de error. Por simple sustitución se tiene:

$$w_i = \beta_1 + \beta_2 Schooling_i + \eta_i$$

donde $\eta_i = u_i - \beta_2 \epsilon_i$. $Schooling_i$ está claramente correlacionada con η_i, por lo que esta variables es endógena en la ecuación de salarios

Causas: simultaneidad

Este problema ocurre cuando al menos uno de las variables explicativas, es determinada simultáneamente junto con y. Si x_k es determinada parcialmente como función de y, entonces x_k y u estarán correlacionadas

Ejemplo: En el modelo keynesiano de determinación del ingreso, se tiene lo siguiente:

$$C_t = \beta_0 + \beta_1 Y_t + u_t$$

$$Y_t = C_t + I_t$$

C_t : consumo agregado
Y_t : producción agregada
I_t : inversión agregada

Lo anterior implica que

$$Y_t = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_t + \frac{1}{1-\beta_1} u_t$$

Es así que $Cov(Y_t, u_t) = \frac{\sigma^2}{1-\beta_1} \neq 0$. Con lo cual la estimación por MCO del consumo agregado sobre la producción agregada no estima el parámetro β_1 consistentemente. El *feedback* entre u_t y Y_t genera un sesgo de ecuaciones simultáneas en los estimadores MCO

Causas: sesgo de selección

Si la muestra con la que se trabaja es no aleatoria y es seleccionada para un grupo de individuos muy particular se puede incurrir en un sesgo de selección al estimar por MCO

La no aleatoriedad de la muestra se traslada a una forma de sesgo por variables omitidas conocida como sesgo por selección muestra

Ejemplo: Si la muestra seleccionada se basa en los valores que toma la variable dependiente se presenta sesgo de selección. Esto obedece a auto-selección o muestra seleccionada. Se tiene entonces que:

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

$$y = \begin{cases} y^*, & \text{si } y^* > 0, \\ -, & \text{si } y^* \leq 0 \end{cases}$$

Esto implica que $E(y) = E(y^*|y^* > 0) = \mathbf{x}'\boldsymbol{\beta} + E(u|u > -\mathbf{x}'\boldsymbol{\beta})$, con lo que

$$\frac{\partial E(y)}{\partial \mathbf{x}} = \boldsymbol{\beta} + \frac{\partial (E(u|u > -\mathbf{x}'\boldsymbol{\beta}))}{\partial \mathbf{x}} \neq \boldsymbol{\beta}$$

No es posible identificar el vector de parámetros $\boldsymbol{\beta}$ por MCO

Causas: modelos dinámicos y perturbaciones autocorrelacionadas

En el contexto de un modelo dinámico cuyas perturbaciones están correlacionadas se tiene:

$$y_t = \rho y_{t-1} + \mathbf{x}'\boldsymbol{\beta} + u_t$$

$$u_t = \sum_{i=1}^{p} \phi_i u_{t-i} + \epsilon_t$$

En este modelo es posible demostrar que $Cov(y_{t-1}, u_t) \neq 0$, es decir que hay problemas de endogeneidad

Causas

La distinción entre estas posibles formas de endogeneidad no son siempre fácil $\Longrightarrow$ una ecuación puede tener más de una fuente de endogeneidad

Ejemplo: efecto del consumo de alcohol (x) sobre la productividad laboral $(y$, medida por los salarios)

- el consumo de alcohol esta correlacionado con factores inobservables (por ejemplo, el background familiar) que también afecta el salario $\Longrightarrow$ problema de variables omitidas

- el consumo de alcohol depende del salario, pero se puede consumir más o menos alcohol si se tienen más o menos salario $\Longrightarrow$ problema de simultaneidad

El estimador por variables instrumentales

- Consecuencias: la estimación de un modelo por MCO con problemas de endogeneidad lleva a estimadores inconsistentes de todos los β_j

- El método de variables instrumentales (IV) provee una solución general para el problema de variables explicatorias endógenas

- Si un instrumento es disponible, el método de IV puede ser utilizado para corregir el problema de endogeneidad y provee consistentes estimadores de los parámetros estructurales β_j

- Note que, inicialmente nos vamos a concentrar en el caso donde existe una variable explicatoria endógena y un instrumento

El estimador por variables instrumentales

Considere el siguiente modelo lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{1}$$

$$E(u) = 0, Cov(x_j, u) = 0, j = 1, 2, \ldots, k-1$$

donde x_k puede estar correlacionada con u, esto es $x_1, \ldots, x_{k-1}$ son exógenas, pero x_k es potencialmente endógena

Para que el estimador IV sea consistente, es necesario una variable observable, z_1, (el instrumento) que no se encuentre en la ecuación (1) y que satisfaga dos condiciones:

1. z_1 debe estar incorrelacionado con u

$$Cov(z_1, u) = 0 \tag{2}$$

   En otras palabras, z_1 es exógeno o válido. Esto a menudo es referido como una restricción de exclusión

2. El instrumento debe ser informativo o relevante . Esto es, el instrumento z_1 debe estar correlacionado con el regresor endógeno x_k, condicional a todas las variable exógenas en el modelo $(x_2, \ldots, x_{k-1})$. Si se escribe la proyección lineal de x_k sobre todas las variables exógenas:

$$x_k = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \epsilon_k \tag{3}$$

$$E(\epsilon_k) = 0, Cov(x_j, \epsilon_k) = 0, Cov(z_1, \epsilon_k) = 0, j = 1, 2, \ldots, k-1$$

   el principal supuesto en esta proyección lineal es que

$$\theta_1 \neq 0 \tag{4}$$

El estimador por variables instrumentales

- Cuando z_1 satisface (2) y (4) entonces se dice que z_1 es una variable instrumental o un instrumento para x_k

- Si z_1 es un instrumento válido y relevante, y $\theta_1 \neq 0$, z_1 impacta sobre y pero sólo indirectamente a través de la variable x_k

- Ya que $x_1, x_2, \ldots, x_{k-1}$ están incorrelacionadas con u, ellas son sus propias variables instrumentales en la ecuación principal $\Longrightarrow$ la lista completa de variables instrumentales es la misma que la lista de variables exógenas

- La proyección lineal en la ecuación (3) es llamada la ecuación de la forma reducida, mientras que la ecuación (1) es llamada la ecuación estructural

El estimador por variables instrumentales

Los supuestos (2) y (4) (validez y relevancia) permiten identificar los parámetros del modelo

Identificación: es posible escribir los β_j de la ecuación estructural (1) en términos de los momentos poblacionales en las variables observables

Entonces, ya que se ha asumido que $x_2, \ldots, x_{k-1}$ son exógenas, entonces los momentos poblacionales serán:

$$\begin{aligned} E(1 \cdot u) &= 0 \\ E(x_2 u) &= 0 \\ E(x_3 u) &= 0 \\ &\ldots \\ E(x_{k-1} u) &= 0 \end{aligned} \tag{5}$$

Ahora, si todo lo que se tiene son los momentos en (5), los parámetros del modelo no son identificables. La razón es simple, con sólo $k-1$ momentos, no se pueden estimar los k parámetros. Este modelo es por tanto subidentificado (*underidentified*)

El estimador por variables instrumentales

Si el instrumento z_1 es disponible (es observable, válido y relevante), entonces el supuesto de instrumento válido provee el momento adicional, esto es, $E(z_1 u) = 0$. Para ver cómo, la ecuación (1) puede ser escrita en términos matriciales de la forma

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{6}$$

donde $\mathbf{x} = (1, x_2, \ldots, x_k)$. El vector $1 \times K$ de todas las variables exógenas es $\mathbf{z} = (1, x_2, \ldots, x_{k-1}, z_1)$

Los supuestos $Cov(x_j, u) = 0$ y $Cov(z_1, u) = 0$ implica k momentos o condiciones de ortogonalidad, esto es

$$E(\mathbf{z}'u) = 0 \tag{7}$$

De (6) y (7) es posible llegar a que

$$[E(\mathbf{z}'\mathbf{x})]\boldsymbol{\beta} = E(\mathbf{z}'y) \tag{8}$$

lo cual es un sistema de K ecuaciones linealies (note que $\mathbf{z}'$ es $k \times N$, $\mathbf{x}$ es $N \times k$, $\boldsymbol{\beta}$ es $k \times 1$, y y es $N \times 1$)

El estimador por variables instrumentales

La ecuación (8) representa un sistema de k ecuaciones lineales con k parámetros desconocidos $(\beta_1, \beta_2, \ldots, \beta_k)$, por lo tanto, el modelo está exactamente identificado

Este sistema tiene única solución si y sólo si la matriz $E(\mathbf{z}'\mathbf{x})$ tiene rango completo:

$$\text{rango } E(\mathbf{z}'\mathbf{x}) = k \tag{9}$$

esto implica que las columnas son linealmente independientes, que no existe multicolinealidad perfecta o que $E(\mathbf{z}'\mathbf{x})$ es invertible} y este supuesto se llama condición de rango

La solución es

$$\boldsymbol{\beta} = [E(\mathbf{z}'\mathbf{x})]^{-1} E(\mathbf{z}'y) \tag{10}$$

El estimador por variables instrumentales

Mientras β en la ecuación anterior es expresado en función de los momentos poblacionales, se puede usar los momentos muestrales (los datos), para consistente estimar β. Se tiene entonces que el estimador de variables instrumentales es

$$\widehat{\beta}^{IV} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' y_i \right) = (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'\mathbf{Y}) \tag{11}$$

El estimador por variables instrumentales

- Mientras que es claro como la condición de validez nos permite identificar el modelo, el rol de la segunda condición (relevancia) puede parecer menos clara

- La condición de relevancia es necesaria, ya que de otra forma el rango de la matriz $E(\mathbf{z}'\mathbf{x})$ será menos que k y el modelo estará subidentificado

- No se hace la demostración (problema 5.12 en Wooldridge provee algunas ayudas), ya que la intuición es muy clara: si $\theta_1 = 0$ en

$$x_k = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \epsilon_k,$$

  entonces eso equivale a no tener un instrumento, con lo cual el modelo es subidentificado como ya lo hemos visto

El estimador por variables instrumentales

En este punto podemos probar entonces si el estimador IV definido en (11) es consistente bajo los supuestos que se han hecho. No te que

$$\widehat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Y})$$

$$\widehat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'(\mathbf{X}\beta + u))$$

$$\widehat{\beta}^{IV} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'u)$$

y entonces se tiene que

$$plim\ \widehat{\beta}^{IV} = \beta$$

por tanto, consistente: cuando el tamaño de la muestra N va ha infinito, el estimador IV converge en probabilidad al verdadero valor poblacional β

El estimador por variables instrumentales

Procedimiento en dos etapas para obtener el estimador IV

Se tiene el siguiente modelo (ecuación estructural)

$$y = \beta_1 + \beta_2 x_2 + u$$

donde x_2 es endógena y existe un instrumento z_1 disponible (satisfaciendo las condiciones de validez y relevancia). La ecuación de la forma reducida es

$$x_2 = \theta_0 + \theta_1 z_1 + \epsilon$$

Es posible obtener el estimador IV de β_2 por medio del siguiente procedimiento en dos etapas

1. Regrese la variable endógena x_2 sobre el instrumento z_1 usando MCO. Calcule el valor predicho de x_2 $(\widehat{x}_2)$

2. Use $\widehat{x}_2$ como variable explicatoria en la ecuación estructural y estime por MCO. La estimación resultante del coeficiente sobre $\widehat{x}_2$ es el estimador IV de β_2. Este término por tanto esta ''purgado'' de la correlación de la variable endógena con los residuales

Note que si se usa x_2 como su propio instrumento en la primera etapa, estos es $z_1 = x_2$, se obtendrían las estimaciones MCO en la segunda etapa. Por tanto, MCO puede verse como un estimador IV en el que se asume que todas las variables explicatorias son exógenas

El estimador por variables instrumentales

Como se ha mostrado, las condiciones de validez y relevancia del instrumento son igualmente importantes en la identificación de los parámetros estructurales β_j. Sin embargo, existe una importante diferencia entre ellos:

- la condición de relevancia puede ser testeada, por ejemplo calculando un t-statistic asociado a $\hat{\theta}_1$ en la regresión de la ecuación reducida (primera etapa)

- la condición de validez, sin embargo, no puede ser probada, ya que esta condición involucra los errores inobservados u $(Cov(z_1, u) = 0)$. Por tanto, esta condición debe asumirse y es importante entonces relacionar la condición de validez con la teoría económica para que el análisis sea convincente

Múltiples instrumentos: mínimos cuadrados en dos etapas (MC2E - *2SLS*)

- Se ha considerado el estimador simple IV con una variable explicativa endógena y un instrumento $\Longrightarrow$ este es el caso de identificación exacta. Similarmente se puede tener dos variables explicatorias endógenas con dos instrumentos y el modelo seguiría estando exactamente identificado

- Si se tienen menos instrumentos que regresores endógenos, el modelo está subidentificado (*underidentified*)

- Si se tienen más instrumentos que regresores endógenos, el modelo está sobreidentificado (*overidentified*)

- En la práctica es a menudo una buena idea tener más instrumentos que los estrictamente necesarios, ya que instrumentos adicionales pueden ser usados para incrementar la precisión de las estimaciones y para construir tests de validez de las restricciones de sobreidentificación (lo cual da luces sobre la validez de los instrumentos). Pero hay que ser moderado en la inclusión de instrumentos

Múltiples instrumentos: mínimos cuadrados en dos etapas (MC2E - *2SLS*)

Supongamos que se tienen M variables instrumentales para x_k: $z_1, z_2, \ldots, z_M$. Además, que cada uno de estos instrumentos satisface la condición de validez

$$Cov(z_h, u) = 0$$

para todo h. Si cada uno de estos instrumentos tiene alguna correlación parcial con x_k (condición de relevancia), se podría entonces en principio calcular M diferentes estimadores IV. Así que cuál estimador IV debería ser usado?

De acuerdo al teórema 5.3 en Wooldridge (2010) el estimador de mínimos cuadrados en dos etapas (MC2E) o *Two-Stage Least Squares* (2SLS) es el estimador IV más eficiente

El estimador 2SLS es obtenido usando todos los instrumentos simultáneamente en la primera etapa:

$$x_k = \delta_1 + \delta_2 x_2 + \ldots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \theta_2 z_2 + \ldots + \theta_M z_M + \epsilon_k$$

Múltiples instrumentos: mínimos cuadrados en dos etapas (MC2E - *2SLS*)

Por definición, el estimador MCO de la primera etapa construirá la combinación lineal de los instrumentos más altamente correlacionada con x_k

Asumiendo que todos los instrumentos son exógenos, entonces este procedimiento conserva más variación exógena en x_k de lo que conservaría cualquier otra combinación lineal de los instrumentos

Otra forma de decir esto es que los instrumentos producen exógena variación en la predicción de x_k

$$\widehat{x}_k = \hat{\delta}_1 + \hat{\delta}_2 x_2 + \ldots + \hat{\delta}_{k-1} x_{k-1} + \hat{\theta}_1 z_1 + \hat{\theta}_2 z_2 + \ldots + \hat{\theta}_M z_M$$

y la estimación MCO en la primera etapa asegura que haya tanta variación exógena como sea posible

Con pocos instrumentos podría existir menos variación exógena en $\widehat{x}_k$, por tanto los estimadores podrían no ser eficientes

Múltiples instrumentos: mínimos cuadrados en dos etapas (MC2E - *2SLS*)

Surge entonces la pregunta: cómo es la condición de relevancia en este caso donde existen más instrumentos que regresores endógenos? La respuesta es que al menos uno de los θ_j en la primera etapa debe ser diferente de cero para que el modelo esté identificado

Dado todo lo anterior, se podría estar tentado incluir tantos instrumentos como fuera posible, ya que más instrumentos mejora la eficiencia de los estimadores 2SLS

Sin embargo, se sabe que teniendo muchos instrumentos, relativo al tamaño de la muestra, se generan potencialmente importantes sesgos, más aún si algunos/muchos/todos los instrumentos son débilmente correlacionados con las variables explicativas endógenas

Usando muchos instrumentos (débiles) tiende a sesgar los estimadores 2SLS hacia los estimadores MCO

El concejo sobre cómo proceder en el ejercicio empírico es usar un modelo moderadamente sobreidenficado $\Longrightarrow$ menos eficiencia pero con menos sesgo

2SLS: el caso general

Ahora se discute el caso en el que existen varias variables explicativas endógenas, pero los mecanismos principales cuando se tiene sólo un regresor endógeno se mantienen

Las condiciones de validez y relevancia en el caso general, donde varios elementos de $\mathbf{x}$ pueden estar correlacionados con u, son como sigue:

$$E(\mathbf{z}'u) = 0 \qquad \text{(Validez)}$$

$$\text{rango } \ E(\mathbf{z}'\mathbf{x}) = k \qquad \text{(Relevancia)}$$

donde $\mathbf{z}$ es $1 \times L$ y el rango $(\mathbf{z}'\mathbf{z}) = L$, descartando la colinealidad entre los instrumentos. En esta notación, cualquier elemento exógeno de $\mathbf{x}$, incluyendo la constante, están incluidos en $\mathbf{z}$

2SLS: el caso general

La condición de validez es fácil de entender, pero la condición de relevancia quizás no

Claramente para que la condición de relevancia (definida como condición de rango) se mantenga, se necesita al menos tantos instrumentos como variables explicatorias existan: $L \geq k \Longrightarrow$ esta condición es conocida como condición de orden

Sin embargo, aunque la condición de orden es necesaria, no es suficiente para que el rango $E(\mathbf{z}'\mathbf{x}) = k$

También es necesario que los elementos de $\mathbf{z}$ estén correlacionados con los elementos de $\mathbf{x}$

Probar formalmente la condición de rango es tedioso y algo complicado, así que no nos meteremos con esto (Wooldridge tampoco lo hace). Lo que si es útil es mirar cuidadosamente los resultados de la primera etapa

Se va a tener tantas regresiones en el primera etapa como variables explicatorias endógenas haya, y es necesario (no suficiente) al menos un coeficiente significante en los instrumentos en cada regresión de la forma reducida para que el modelo este identificado

2SLS: el caso general

Para ver lo anterior, consideremos el siguiente modelo

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

donde x_3 y x_4 son endógenos y, por tanto, se necesitan al menos 2 instrumentos, z_1 y z_2

Estos instrumentos entran en las ecuaciones de la forma reducida para x_3 y x_4

$$x_3 = \pi_1 + \pi_2 x_2 + \pi_3 z_1 + \pi_4 z_2 + \epsilon_1$$

$$x_4 = \gamma_1 + \gamma_2 x_2 + \gamma_3 z_1 + \gamma_4 z_2 + \epsilon_2$$

- Si $\pi_3 = 0$, $\pi_4 \neq 0$, $\gamma_3 = 0$, $\gamma_4 \neq 0$: la ecuación estructural no estará identificada, ya que el instrumento z_1 no es relevante en ambas ecuaciones y, por tanto, efectivamente sólo se tendrá un instrumento
- Si $\pi_3 = 0$, $\pi_4 \neq 0$, $\gamma_3 \neq 0$, $\gamma_4 = 0$: la ecuación estructural estará identificada, ya que el instrumento z_1 es relevante en la ecuación determinando x_4, mientras que z_2 es relevante para x_3

2SLS: el caso general

- Desde un punto de vista práctico, es posible identificar rápidamente si la identificación falla. Si el modelo no está identificado, debido a que tiene pocos instrumentos o por que los instrumentos son colineales, un software como R o Stata reportará esto y no estimará nada

- Si los instrumentos tiene una correlación muy débil con las variables explicatorias endógenas, los coeficientes de los instrumentos en la primera etapa pueden ser insignificantes, y los errores estándar 2SLS serán muy grandes $\Longrightarrow$ en otras palabras, no estaríamos aprendiendo nada del modelo planteado

Expresión general para el estimador 2SLS

El álgebra del estimador 2SLS es más complicado que el del estimador IV. Usando álgebra matricial nos ayuda a entender los mecanismos generales

Recordemos que el estimador IV tiene la siguiente estructura

$$\widehat{\beta}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Y})$$

Es fácil mostrar que esta expresión puede ser expresada como

$$\widehat{\beta}^{IV} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}(\widehat{\mathbf{X}}'\mathbf{Y})$$

es decir, MCO usando la predicción en lugar de los valores reales de las variables explicatorias (para las variables exógenas en $\mathbf{X}$, los valores predicho y reales coinciden)

La misma expresión se mantiene para los estimadores 2SLS

$$\widehat{\beta}^{2SLS} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}(\widehat{\mathbf{X}}'\mathbf{Y})$$

Expresión general para el estimador 2SLS

Sin embargo, ya que el modelo esta sobreidenficado (*overidentified*) la expresión del estimador 2SLS no es igual a la del estimador IV

Para ver qué se obtiene si se escribe el estimador 2SLS en términos de las matrices $\mathbf{Z}$ y $\mathbf{X}$, note que

$$\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P}_Z\mathbf{X}$$

donde $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ (conocida como la matriz de proyección) es idempotente $(\mathbf{P}_Z\mathbf{P}_Z = \mathbf{P}_Z)$ y simétrica $(\mathbf{P}'_Z = \mathbf{P}_Z)$. Por tanto,

$$\widehat{\mathbf{X}}'\mathbf{X} = (\mathbf{P}_Z\mathbf{X})'\mathbf{X} = \mathbf{X}'\mathbf{P}_Z\mathbf{X} = \mathbf{X}'\mathbf{P}_Z\mathbf{P}_Z\mathbf{X} = \mathbf{X}'\mathbf{P}'_Z\mathbf{P}_Z\mathbf{X} = (\mathbf{P}_Z\mathbf{X})'\mathbf{P}_Z\mathbf{X} = \widehat{\mathbf{X}}'\widehat{\mathbf{X}}$$

Incorporando esta expresión en el estimador 2SLS se tiene que la expresión de éste estimador es

$$\widehat{\beta}^{2SLS} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}(\widehat{\mathbf{X}}'\mathbf{Y})$$

que es el estimador IV cuando usa como instrumento $\widehat{\mathbf{X}}$

Expresión general para el estimador 2SLS

El estimador 2SLS puede ser obtenido de los siguientes pasos

1. Obtenga los valores estimados x_k de la regresión

$$x_K \text{ sobre } 1, x_1, \ldots, x_{k-1}, z_1, \ldots, z_M$$

   La anterior ecuación es llamada regresión de la primera etapa

2. Corra por MCO la regresión

$$y \text{ sobre } 1, x_1, \ldots, x_{K-1}, \widehat{x}_k$$

   La anterior ecuación es llamada regresión de la segunda etapa, y con esta se obtienen los $\widehat{\beta}_j$

En términos prácticos, es mejor utilizar un software con un comando o paquete 2SLS, en lugar de hacerlo manualmente. Esto por dos principales razones

- errores en la especificaciones en alguna de las dos etapas
- los errores estándar MCO con incorrectos cuando se tienen variables explicativas predichas

Ejercicio aplicado: efectos de la educación de la mujer sobre la fertilidad

Las mujeres más educadas tienen menos hijos?

Numerosos estudios indican que la educación de las mujeres tiene un efecto negativo sobre la fertilidad. Varias son las posibles explicaciones:

- la escolarización aumenta el coste de oportunidad de tener un hijo
- aumenta la eficiencia del control de fertilidad
- simplemente reduce la preferencia por los hijos

En este ejercicio vamos a estudiar este posible efecto negativo. El siguiente ejercicio se basa en el paper: McCrary, J y Royer, H. (2011). "The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth". *American Economic Review*, 101: 158-195.

Los datos para este ejercicio proviene de la *General Social Survey (GSS)* de los Estados Unidos. En los siguientes links se encuentran los datos, la descripción dellada de los datos y el código utilizado en R:

- Datos
- Descripción de la información
- Código en R

Ejercicio aplicado: efectos de la educación de la mujer sobre la fertilidad

Cargando las librerías

```
library(devtools); library(haven); library(dplyr); library(cragg); library(AER); library(tidyverse); library(stargazer); library(modelsummary)
library(gt); library(knitr); library(kableExtra); library(tibble)
```

Leyendo los datos y procesando la información

```
setwd("C:/Users/ggarci24/OneDrive - Universidad EAFIT/EAFIT/Cursos EAFIT/Econometria II/R/Tema 9")

data <- read_dta("GSS2012_2018.DTA") %>% # Leyendo el archivo .dta
  select(year, age, sex, race, educ, childs, paeduc, maeduc, wrkstat, marital) %>% # Seleccionando variables de la base
  filter(sex==2, year>=2014 & year<=2018, age>=35 & age<=55) %>% # Filtrando la base para mujeres y años
  mutate(age2 = age*age, afroa = case_when(race == 1 ~ 1,
                                           race == 2 ~ 0),
                         working = case_when(wrkstat >= 1 & wrkstat<= 2 ~ 1,
                                             wrkstat >= 3 & wrkstat<= 8 ~ 0),
                         casado = case_when(marital == 1 ~ 1,
                                            marital != 1 ~ 0)) %>% # Creando variables
  drop_na() # Borrando missings de toda la base
```

```
data[,c("race","afroa")]
```

```
# A tibble: 935 x 2
        race afroa
   <dbl+lbl> <dbl>
 1 1 [white]     1
 2 1 [white]     1
 3 1 [white]     1
 4 1 [white]     1
 5 2 [black]     0
 6 1 [white]     1
 7 1 [white]     1
 8 1 [white]     1
 9 1 [white]     1
10 2 [black]     0
# ... with 925 more rows
```

```
data[,c("wrkstat","working")]
```

```
# A tibble: 935 x 2
                  wrkstat working
                <dbl+lbl>   <dbl>
 1 7 [keeping house]           0
 2 1 [working fulltime]        1
 3 3 [temp not working]        0
 4 1 [working fulltime]        1
 5 1 [working fulltime]        1
 6 1 [working fulltime]        1
 7 1 [working fulltime]        1
 8 1 [working fulltime]        1
 9 1 [working fulltime]        1
10 1 [working fulltime]        1
# ... with 925 more rows
```

```
data[,c("marital","casado")]
```

```
# A tibble: 935 x 2
              marital casado
            <dbl+lbl>  <dbl>
 1 5 [never married]       0
 2 5 [never married]       0
 3 5 [never married]       0
 4 4 [separated]           0
 5 5 [never married]       0
 6 1 [married]             1
 7 5 [never married]       0
 8 1 [married]             1
 9 1 [married]             1
10 5 [never married]       0
# ... with 925 more rows
```

Ejercicio aplicado: efectos de la educación de la mujer sobre la fertilidad

Estimación por OLS

```
ols <- lm(childs ~ educ+age+I(age2)+casado+afroa+working, data=data)


summary(ols)
```

```
Call:
lm(formula = childs ~ educ + age + I(age2) + casado + afroa +
    working, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6795 -0.8997 -0.0462  0.7965  5.6386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.231814   2.675422  -0.087  0.93097
educ        -0.136673   0.015867  -8.614  < 2e-16 ***
age          0.204314   0.119975   1.703  0.08891 .
I(age2)     -0.002341   0.001328  -1.763  0.07820 .
casado       0.471561   0.091202   5.171 2.86e-07 ***
afroa       -0.196362   0.124672  -1.575  0.11559
working     -0.259425   0.098934  -2.622  0.00888 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.345 on 928 degrees of freedom
Multiple R-squared:  0.1132,    Adjusted R-squared:  0.1075
F-statistic: 19.75 on 6 and 928 DF,  p-value: < 2.2e-16
```

Estimación por 2SLS

```
iv <- ivreg(childs ~ educ+age+I(age2)+casado+afroa+working |
                            age+I(age2)+casado+afroa+working+paeduc+maeduc,

summary(iv)
```

```
Call:
ivreg(formula = childs ~ educ + age + I(age2) + casado + afroa +
    working | age + I(age2) + casado + afroa + working + paeduc +
    maeduc, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-4.01831 -0.89164 -0.05782  0.79643  5.53646

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.176497   2.699724   0.065   0.9479
educ        -0.179672   0.032775  -5.482 5.42e-08 ***
age          0.212482   0.120572   1.762   0.0784 .
I(age2)     -0.002446   0.001335  -1.832   0.0672 .
casado       0.506371   0.094453   5.361 1.04e-07 ***
afroa       -0.200423   0.125194  -1.601   0.1097
working     -0.196300   0.107859  -1.820   0.0691 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.351 on 928 degrees of freedom
Multiple R-Squared: 0.1062,    Adjusted R-squared: 0.1004
Wald test: 12.33 on 6 and 928 DF,  p-value: 2.171e-13
```

Ejercicio aplicado: efectos de la educación de la mujer sobre la fertilidad

Utilizamos el paquete `modelsummary` para generar tablas editadas (Word, tex, text, png, html...)

```
modelos <- list("OLS" = lm(childs ~ educ+age+I(age2)+casado+afroa+working, data=data),
                "IV"  = ivreg(childs ~ educ+age+I(age2)+casado+afroa+working |
                              age+I(age2)+casado+afroa+working+paeduc+maeduc, data=data))
cm <- c( '(Intercept)' = 'Constante', 'educ' = 'Educación', 'age' = 'Edad', 'I(age2)' = 'Edad2', 'casado' = 'Casado', 'afroa' = 'Afroamericano', 'work
cap <- 'Tabla 1. Determinantes de la fertilidad'
modelsummary(modelos, output = 'gt', coef_map = cm, stars = c('*'=.1, '**'=.05, '***'=.01), statistic = "std.error", title = cap,
             gof_omit = 'IC|Log', coef_omit = "[^educ]") %>%
    tab_style(style = cell_text(size = 'medium'), locations = cells_body(rows = 1:6)) %>%
    tab_style(style = cell_text(color = 'red'), locations = cells_body(rows = 1)) %>%
    tab_source_note(source_note = "Nota: Errores estándar en paréntesis") %>%
    tab_style(style = cell_text(color = "black", size = "x-small"), locations = cells_source_notes())
```

Tabla 1. Determinantes de la fertilidad

	OLS	IV
Educación	-0.137***	-0.180***
	(0.016)	(0.033)
Num.Obs.	935	935
R2	0.113	0.106
R2 Adj.	0.107	0.100
F	19.747	

* p < 0.1, ** p < 0.05, *** p < 0.01

Nota: Errores estándar en paréntesis