Variables Instrumentales

Juan Manuel del Pozo Segura

NFER, PUCP

jmdelpozo@pucp.pe

PUCP Q Lab - Junio 2023

Outline

1. Motivación
2. La endogeneidad y el instrumento
   - La lógica del modelo de regresión
   - La falla del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$
   - Entendiendo a la variable instrumental
   - Ilustración: Angrist y Evans (1996)
   - Cómo los instrumentos suelen fallar
3. Efectos de tratamiento homogéneos
   - Estimadores de variables instrumentales
   - Ilustración: Angrist y Krueger (1991)
4. Efectos de tratamiento heterogéneos
   - El estimador de Wald
   - Ilustración: Angrist (1990)
5. Aspectos técnicos y problemas
   - El problema de instrumentos débiles
   - Dos (o tres) tests importantes
6. Conclusión

Qué es endogeneidad

- El concepto de **endogeneidad** es una de las contribuciones fundamentales de la Econometría a la Estadística
- En términos generales, **endogeneidad** se refiere a la correlación entre el término de error y alguna de las variables independientes en nuestro modelo de regresión
  - En la literatura de evaluación de programas, a la endogeneidad también se le conoce como **selección en no observables**
  - Los individuos pueden auto-seleccionarse en un programa por razones que no observamos y que están correlacionadas con la variable de resultado
- Sea cual sea el contexto donde ocurra, endogeneidad es una gran amenaza en la estimación: genera sesgo e inconsistencia de nuestro estimador, lineal o no lineal

Qué es endogeneidad

- Por tanto, es fundamental encontrar métodos que permitan solucionar este problema
- Existen diferentes métodos
  1. Encontrar datos con una dimensión adicional de los datos aparte de la individual, usualmente tiempo, y estimar modelos panel de efectos fijos
  2. Usar el método de **variables instrumentales** (VI), lo que requiere *encontrar* un instrumento
- En esta clase, nos centraremos en este segundo método

Outline

1. Motivación
2. La endogeneidad y el instrumento
   - La lógica del modelo de regresión
   - La falla del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$
   - Entendiendo a la variable instrumental
   - Ilustración: Angrist y Evans (1996)
   - Cómo los instrumentos suelen fallar
3. Efectos de tratamiento homogéneos
   - Estimadores de variables instrumentales
   - Ilustración: Angrist y Krueger (1991)
4. Efectos de tratamiento heterogéneos
   - El estimador de Wald
   - Ilustración: Angrist (1990)
5. Aspectos técnicos y problemas
   - El problema de instrumentos débiles
   - Dos (o tres) tests importantes
6. Conclusión

¿Qué queremos que nuestro modelo explique?

- Para responder a esta pregunta usamos una herramienta fundamental: **la función de expectativa condicional (CEF)**

$$E(y|\mathbf{x}) = f(\mathbf{x}, \beta)$$

La CEF, también llamada **el modelo poblacional**, es la porción que nos interesa explicar de nuestro outcome

- Una vez definida la CEF, podemos expresar el **modelo en error-form**

$$E(y|\mathbf{x}) = f(\mathbf{x}, \beta) \rightarrow y = E(y|\mathbf{x}) + \varepsilon$$

La dependiente, y, puede separarse en 2 partes

1. una parte sistemática dada por la CEF $E(y|\mathbf{x})$
2. un error ε, que es una *bolsa* donde está todo lo que determina y pero que *no* forma parte de las variables en $\mathbf{x}$

- En una CEF bien definida el error ε tiene una característica *fundamental*: su promedio no está relacionado con $\mathbf{x}$, ε es mean independent de $\mathbf{x}$

$$\boxed{E(\varepsilon|\mathbf{x}) = 0}$$

A esta condición se le conoce como la **Zero Conditional Mean (ZCM)**

La condición poblacional más importante

- La CEF *puede tomar cualquier forma*, pero asumiremos que depende **linealmente** de un vector de parámetros $\beta_{K\times 1}$

$$E(y|\mathbf{x}) = f(\mathbf{x},\beta) \quad \underbrace{\rightarrow}_{\text{asumiendo linealidad}} \quad E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K \equiv \mathbf{x}'\beta$$

- En nuestro **modelo poblacional** $E(y|\mathbf{x}) = \mathbf{x}'\beta$:
  - el vector de parámetros poblacionales, $\beta_{(1+K)\times 1}$, *no* son observados
  - necesitamos un **estimador**, $\hat{\beta}$, que nos provea valores *fiables* de β usando info de una muestra que *sí* observamos. Siguiendo el curso, comenzaremos con el estimador MCO
- Para que $\hat{\beta}$ sea fiable, es necesario que la ZCM se cumpla, i.e. que ε sea **mean independent** de $\mathbf{x}$

$$\boxed{E(\varepsilon|\mathbf{x}) = 0}$$

¿Qué implica la ZCM en nuestro modelo poblacional?

- Volvamos a la CEF $E(y|\mathbf{x}) = \mathbf{x}'\beta$
- Si se cumple la ZCM $E(\varepsilon|\mathbf{x}) = 0$, los parámetros poblacionales β, llamados **efectos parciales**, que se interpretan como

$$\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}} \underbrace{=}_{\text{asumiendo linealidad}} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \mathbf{x}} = \boldsymbol{\beta}$$

  nos proveen el efecto que sólo y *únicamente* puede atribuirse a cada variable $x \in \mathbf{x}$ y *no* a factores no observados en ε
- En este caso, decimos que los β miden el **efecto causal** de cada una de las variables en $\mathbf{x}$
- Por tanto el cumplimiento de la ZCM implica que nuestro modelo poblacional es causal

$$\boxed{\text{validez de la ZCM}\rightarrow\text{validez de la CEF} \rightarrow \text{modelo causal}}$$

¿Qué implica la ZCM para el estimador?

- Un estimador válido debe incorporar en su derivación el supuesto de ZCM
- La ZCM $E(\varepsilon|\mathbf{x}) = 0$ junto con la LIE $E[E(y|\mathbf{x})] = E(y)$ implican 2 resultados

1. El error ε tiene promedio incondicional 0

$$E(\varepsilon|\mathbf{x}) = 0 \underbrace{\rightarrow}_{\text{tomando expectativas}} E[E(\varepsilon|\mathbf{x})] = E(0) \underbrace{\Leftrightarrow}_{\text{por la LIE}} E(\varepsilon) = 0$$

2. La covarianza entre ε y las $\mathbf{x}$ es 0

$$E(\varepsilon|\mathbf{x}) = 0 \text{ implica } Cov(\mathbf{x}, \varepsilon) = 0 \Leftrightarrow E(\mathbf{x}\varepsilon) - E(\mathbf{x}) \underbrace{E(\varepsilon)}_{=0} = 0 \Leftrightarrow Cov(\mathbf{x}, \varepsilon) = \boxed{E(\mathbf{x}\varepsilon) = 0}$$

Para derivar el estimador *no* estamos recurriendo a la muestra sino sólo a supuestos poblacionales

¿Qué implica la ZCM para el estimador?

- Para ver el rol de estos 2 resultados, volvamos al modelo *lineal* en error-form

$$y = \mathbf{x}^{'}\beta + \varepsilon$$

 pre-multipliquémoslo por $\mathbf{x}$ y tomemos las expectativas

$$y = \mathbf{x}^{'}\beta + \varepsilon \Leftrightarrow \mathbf{x}y = \mathbf{x}\mathbf{x}^{'}\beta + \mathbf{x}\varepsilon \Leftrightarrow E(\mathbf{x}y) = E\left(\mathbf{x}\mathbf{x}^{'}\beta\right) + E(\mathbf{x}\varepsilon) \Leftrightarrow E(\mathbf{x}y) = E\left(\mathbf{x}\mathbf{x}^{'}\right)\beta + E(\mathbf{x}\varepsilon)$$

acabamos de ver que $E(\varepsilon|\mathbf{x}) = 0$ implica $E(\mathbf{x}\varepsilon) = 0$, entonces

$$E(\mathbf{x}y) = E\left(\mathbf{x}\mathbf{x}^{'}\right)\beta \Leftrightarrow \beta_{MCO} = \left[E\left(\mathbf{x}\mathbf{x}^{'}\right)\right]^{-1} E(\mathbf{x}y)$$

- Esta es la expresión **poblacional del estimador MCO** β_{MCO}
  - Los efectos parciales β son una función de promedios **poblacionales**: $E\left(\mathbf{x}\mathbf{x}^{'}\right)$, $E(\mathbf{x}y)$
  - β_{MCO} está **identificado** si la matriz $\left[E\left(\mathbf{x}\mathbf{x}^{'}\right)\right]^{-1}$ existe. En ese caso β tiene un valor único
- Si la ZCM es falsa, la expresión β_{MCO} no se corresponde con $\left[E\left(\mathbf{x}\mathbf{x}^{'}\right)\right]^{-1} E(\mathbf{x}y)$

El estimador MCO

- β_{MCO} depende de $E\left(\mathbf{x}\mathbf{x}^{'}\right), E(\mathbf{x}y)$ y por tanto no es directamente observable.
- Pero tenemos una **muestra**, pues observamos $\{(y_i, \mathbf{x}_i) : i = 1, ..., N\}$ ¿Cómo vincular $(y_i, \mathbf{x}_i)$ en la muestra con $E\left(\mathbf{x}\mathbf{x}^{'}\right), E(\mathbf{x}y)$ en la población?
- El **principio de analogía** de Manski (Goldberger 1968, Manski 1968) nos dice que los momentos poblacionales pueden estimarse consistentemente a partir de momentos muestrales

$$\beta_{MCO} = \left[E\left(\mathbf{x}\mathbf{x}^{'}\right)\right]^{-1} E(\mathbf{x}y) \underbrace{\rightarrow}_{\text{princip. analog.}} \boxed{\hat{\beta}_{MCO} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{'}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{'}\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i}$$

Esta es la expresión **muestral del estimador MCO** $\hat{\beta}_{MCO}$, tras reemplazar

- promedios poblacionales $E\left(\mathbf{x}\mathbf{x}^{'}\right)$ y $E(\mathbf{x}y)$ que no observamos, por
- promedios muestrales, $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{'}$ y $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i y_i$ que *sí* observamos

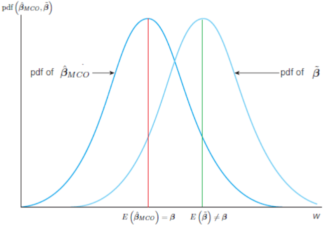
La ZCMA y el insesgamiento

- Hasta ahora hemos dicho que la ZCM $E(\varepsilon|\mathbf{x}) = 0$ permite que
  1. nuestra CEF $E(y|\mathbf{x}) = \mathbf{x}'\beta$ sea válida
  2. $E(\varepsilon|\mathbf{x}) = 0 \Rightarrow Cov(\mathbf{x}, \varepsilon) = 0$, lo que nos permitió derivar β_{MCO}

  por el principio de analogía, a partir de β_{MCO} obtuvimos $\hat{\beta}_{MCO}$
- Si este supuesto se cumple, entonces decimos nuestro estimador es **insesgado**

$$E\left(\hat{\beta}_{MCO}\right) = \beta$$

  i.e. el promedio de los infinitos estimados $\hat{\beta}^1_{MCO}, \hat{\beta}^2_{MCO}, ..., \hat{\beta}^N_{MCO}$ que podemos calcular con diferentes muestras de tamaño N es igual al β poblacional (verdadero)



Pero qué significa realmente insesgamiento

- Pero rara vez verán que textos «serios» hablan de insesgamiento. Si bien es una propiedad deseable, no se considera en los tratamientos formales pues
  1. *Sólo* puede derivarse matemáticamente para modelos lineales. Algunos estimadores no lineales, y el **estimador IV,** no puede analizarse bajo el concepto de insesgamiento
  2. El supuesto ZCM es **muy fuerte**: es muy fácil de que falle: Basta que *sólo* 1 de los componentes en $\mathbf{x}$ ocurra que $E(\varepsilon|\mathbf{x}) \neq 0$ para que *todos* los $\hat{\beta}$ sean sesgados.
- E.g. si $\mathbf{x} = (x_1, x_2, x_3)$ y
  - $E(\varepsilon|x_1) = E(\varepsilon|x_2) = 0$
  - $E(\varepsilon|x_3) \neq 0$

  entonces $\hat{\beta}_3$ es sesgado ¡y también $\hat{\beta}_2$ y $\hat{\beta}_1$!

$E(\varepsilon|\mathbf{x}) = 0$ puede fallar de formas *no* obvias

- $E(\varepsilon|\mathbf{x}) = 0$ significa que, en promedio, no existe ningún tipo de relación *lineal o no lineal* entre *ninguna* de las $\mathbf{x}$ y ε . Esto implica que
  1. nuestro modelo incluye *todas* las relaciones lineales y *no lineales*. Esto se violaría si, e.g., declaramos que nuestro modelo es $y = \beta_0 + \beta_x x + \varepsilon$ pero, en realidad, el modelo verdadero es $y = \beta_0 + \beta_{x1} x + \boxed{\beta_{x2} x^2 + \beta_{x3} x^3} + \varepsilon$. La omisión de x^2, x^3 lleva a la falla de ZCM
  2. nuestro modelo especifica *exactamente a* la variable dependiente. Esto se violaría si, e.g. en nuestro modelo usamos el *salario* en vez de $\ln$ *salario* como dependiente si es el *log* en realidad pertenece al modelo poblacional.
- Queda claro que el supuesto de ZCM $E(\varepsilon|\mathbf{x}) = 0$ es más restrictivo de lo que parece, pues implica que
  1. formas funcionales incorrectas y/o
  2. relaciones no lineales no capturadas en nuestro modelo

  crean problemas de sesgo

De insesgamiento a consistencia

- Por tanto, es razonable *relajar* el supuesto de $E(\varepsilon|\mathbf{x}) = 0$: *repensaremos* los resultados en términos de **consistencia** en vez de insesgamiento
  - Ahora queremos que el estimador $\hat{\beta}$ se *acerque más y más a* β a medida que el N que se usa para calcularlo aumenta.
  - En este caso, el estimador $\hat{\beta}$ **converge** a β, y en términos de notación

$$\text{plim}\hat{\beta} - \beta = \mathbf{0} \Leftrightarrow \text{plim}\hat{\beta} = \beta$$

- La condición básica **de consistencia** ya la vimos, y se desprendió de la ZCM

$$E(\varepsilon|\mathbf{x}) = 0 \underbrace{\rightarrow}_{\text{cambia a}} Cov(\varepsilon, \mathbf{x}) = E(\mathbf{x}\varepsilon) - E(\mathbf{x})E(\varepsilon) = 0 \Leftrightarrow \boxed{E(\mathbf{x}\varepsilon) = 0}$$

i.e. en vez de ZCM ahora simplemente asumimos *directamente* 0 covarianzas (o 0 correlación)

Qué permite $Cov(\varepsilon, \mathbf{x}) = 0$

- Al pensar en $E(\mathbf{x}\varepsilon) = 0$ en vez de $E(\varepsilon|\mathbf{x}) = 0$ reducimos la lista de posibles problemas en nuestra estimación
  - $E(\varepsilon|\mathbf{x}) = 0$ implica pensar en relacionales lineales y no lineales entre cada una de las x y ε, las cuales *no* son obvias
  - $Cov(\varepsilon, \mathbf{x}) = 0 \Leftrightarrow E(\mathbf{x}\varepsilon) = 0$ implica pensar sólo en relacionales *lineales* entre cada una de las x *y* ε

  Es decir, asumir $Cov(\varepsilon, \mathbf{x}) = 0 \Leftrightarrow E(\mathbf{x}\varepsilon) = 0$ implica que ε puede estar correlacionada con funciones no lineales de las variables en $\mathbf{x}$ y que no han sido incluidas, tales como cuadráticas, cúbicas interacciones, etc. ¡Y aún así nuestro estimador será consistente! (pero sesgado)

¿Cuán razonable es $Cov(\varepsilon, \mathbf{x}) = 0$ en el análisis aplicado?

- Entonces, debido a que pasar de $E(y|\mathbf{x}) = 0$ a $Cov(\varepsilon, \mathbf{x}) = 0$ requiere sólo pensar en relaciones lineales entre $\mathbf{x}$ y ε ¿la condición $Cov(\varepsilon, \mathbf{x}) = 0$ queda asegurada? !No!
- La condición de consistencia $Cov(\varepsilon, \mathbf{x}) = 0$ todavía puede (y probablemente falle) fallar por 3 motivos comunes
  - **Variable omitida**: e.g. habilidad (que determina salarios y que no se observa, y por tanto está en ε) está correlacionada con al menos 1 x_j (educación) en $\mathbf{x}$. Esto es el problema más común debido a *limitaciones de datos*
  - Errores de medición: nuestra variable de interés es x_j^* pero sólo obtenemos una medida imperfecta de esta variable, llamada x_j
  - Simultaneidad: cuando al menos 1 de las variables explicativas se determina simultáneamente junto con y
- Es de esperar que alguno de estos 3 problemas, con datos no experimentales, ***siempre esté presente***

Salvo en casos experimentales, cuando estimamos un modelo lineal con 1 cross section, es razonable asumir que endogeneidad **siempre** estará presente

Outline

1. Motivación
2. La endogeneidad y el instrumento
   - La lógica del modelo de regresión
   - La falla del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$
   - Entendiendo a la variable instrumental
   - Ilustración: Angrist y Evans (1996)
   - Cómo los instrumentos suelen fallar
3. Efectos de tratamiento homogéneos
   - Estimadores de variables instrumentales
   - Ilustración: Angrist y Krueger (1991)
4. Efectos de tratamiento heterogéneos
   - El estimador de Wald
   - Ilustración: Angrist (1990)
5. Aspectos técnicos y problemas
   - El problema de instrumentos débiles
   - Dos (o tres) tests importantes
6. Conclusión

El modelo de retornos a la educación

- Comencemos ilustrando un problema de economía laboral: la inconsistencia de los retornos a la educación
- Digamos que el verdadero modelo de retornos a la educación (x) está dado por

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

  pero digamos que a, la habilidad *individual*, no se observa porque
  - no existe tal variable en nuestros datos, o
  - no es *realmente* medible
- Debido a que habilidad no se observa, a entra en la «bolsa» del error

$$y = \beta_0 + \beta_1 x + \gamma a + u \quad \underbrace{\rightarrow}_{\text{debido a que } a \text{ es no observado}} \quad \boxed{y = \beta_0 + \beta_1 x + \underbrace{\varepsilon}_{\equiv \gamma a + u}}$$

  A diferencia de la primera ecuación, la segunda (**ecuación corta)** *sí podemos* estimar.

El (in)cumplimiento del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$

- Dada la ecuación estimable $y = \beta_0 + \beta_1 x + \varepsilon$, ¿qué se necesita para la consistencia del estimador MCO de β_1 y β_0, $\hat{\beta}_{1,MCO}, \hat{\beta}_{0,MCO}$? Recuérdese que la condición de *consistencia* del estimador MCO es

$$E(\mathbf{x}\varepsilon) = 0 \underbrace{\rightarrow}_{\text{en este caso}} E(x\varepsilon) = E(x(\gamma a + u)) = \gamma E(xa) + E(xu) = 0$$

  Entonces requerimos que

$$E(x\varepsilon) = 0 \Leftrightarrow E(xa) = 0 \text{ y } E(xu) = 0$$

  - Centrémonos en el segundo término. Podemos asumir que en $y = \beta_0 + \beta_1 x + \gamma a + u$, el error u *no* contiene ningún elemento que explique los salarios y que esté correlacionado con x. Por tanto

$$E(xu) = 0$$

  - ¿Pero qué hay del primer término? Evidentemente, la habilidad *sí* explica el salario y está correlacionada con la educación. Por tanto

$$E(xa) \neq 0$$

El (in)cumplimiento del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$

- Por tanto, educación, x, es una variable **endógena** pues está correlacionada con el error compuesto vía la correlación que tiene con habilidad

$$E(x\varepsilon) = \gamma \underbrace{E(xa)}_{\neq 0} + \underbrace{E(xu)}_{=0} \neq 0$$

  La condición de consistencia *no* se cumple en este caso y los estimadores de los retornos a la educación son inconsistentes

- Las implicancias en términos de violación del supuesto de $Cov(\varepsilon, \mathbf{x}) = 0$ son tan (o más) graves que la violación de la ZCM. E.g. si $\mathbf{x} = (x_1, x_2, x_3)$ y
  - $Cov(\varepsilon, x_1) = Cov(\varepsilon, x_2) = 0$
  - $Cov(\varepsilon, x_3) \neq 0$

  Entonces, aun cuando *sólo* x_3 origina la falla de $Cov(\varepsilon, \mathbf{x}) = 0$, *todos los estimadores,* $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$ *serán inconsistentes*

¿Qué significa para nuestro estimador MCO?

- Recuérdese que estamos interesados en $\hat{\beta}_{1,MCO}$ en el modelo de regresión corto $y = \beta_0 + \beta_1 x + \varepsilon$. La fórmula del estimador MCO en este **modelo de regresión simple** es sencilla

$$\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) \underline{y_i}}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

  pero sabemos que el modelo *verdadero* es el modelo largo
  $\underline{y_i} = \beta_0 + \beta_1 x_i + \gamma a_i + u_i$

$$\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) (\beta_0 + \beta_1 x_i + \gamma a_i + u_i)}{\sum_{i=1}^{N} (x_i - \overline{x})^2} = \beta_1 + \gamma \frac{\sum_{i=1}^{N} (x_i - \overline{x}) a_i}{\sum_{i=1}^{N} (x_i - \overline{x})^2} + \frac{\sum_{i=1}^{N} (x_i - \overline{x}) u_i}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

- Para analizar la *consistencia* de $\hat{\beta}_{1,MCO}$ debemos tomar el probability limit (*plim*). Operando obtenemos la expresión para la magnitud de inconsistencia de nuestro estimador a partir del modelo corto

$$\boxed{\text{Inconsistencia : plim}\left(\hat{\beta}_{1,MCO}\right) - \beta_1 = \gamma \times \frac{Cov(x,a)}{Var(x)}}$$

  Entonces, si $\gamma \frac{Cov(x,a)}{Var(x)} = 0$ nuestro estimador es consistente

  $\text{plim}\left(\hat{\beta}_{1,MCO}\right) - \beta_1 = 0 \Leftrightarrow \text{plim}\left(\hat{\beta}_{1,MCO}\right) = \beta_1$ ▸ Derivación

Qué nos dice la expresión de la inconsistencia

- La expresión $\text{Inconsistencia} = \gamma \times \frac{Cov(x,a)}{Var(x)}$ nos dice que la inconsistencia por estimar $\hat{\beta}_{1,MCO}$ a partir del modelo *corto* $y = \beta_0 + \beta_1 x + \varepsilon$ depende de
  1. γ, el efecto *parcial* de a (la variable omitida) sobre y en el modelo original (que no podemos estimar porque no observamos a) $y = \beta_0 + \beta_1 x + \gamma a + u$
  2. $\frac{Cov(x,a)}{Var(x)} \equiv \theta_1$, el efecto marginal de x sobre a (la variable omitida), igual a la pendiente de la regresión poblacional (que tampoco podemos estimar)

$$a = \theta_0 + \theta_1 x + v$$

- Esto significa que *no existirá* una inconsistencia por variable omitida si se cumple ***sólo 1*** de las siguientes 2 condiciones
  - $\gamma = 0$: la variable omitida a *no* tiene efecto *parcial sobre* y, y entonces $\text{Inconsistencia} = 0 \times \frac{Cov(x,a)}{Var(x)} = 0$, ***ó***
  - $\theta_1 = 0$: x no tiene efecto sobre a , y entonces $\text{Inconsistencia} = \gamma \times 0 = 0$

El sesgo de variables omitidas ocurre bajo condiciones más estrictas que simplemente dejar de incluir una variable en el modelo

La interpretación de la expresión en nuestro caso

- Nótese que también la expresión $\text{Inconsistencia} = \gamma \frac{Cov(x,a)}{Var(x)}$, nos permite cuantificar la magnitud del sesgo. Para ello necesitamos
  1. medir γ, el efecto parcial de a sobre y en $y = \beta_0 + \beta_1 x + \gamma a + u$
  2. medir $\frac{Cov(x,a)}{Var(x)} \equiv \theta_1$, el efecto marginal de x sobre a en $a = \theta_0 + \theta_1 x + v$
- De todos modos, en nuestro caso es de esperar que
  1. la habilidad a tenga un efecto *parcial* (positivo) sobre los salarios y, por lo que $\gamma \neq 0$
  2. la educación x tenga un efecto marginal (positivo) sobre la habilidad a, por lo que $Cov(x, a) = E(xa) \neq 0$ y entonces $\theta_1 \neq 0$

  Por tanto

$$\gamma \frac{Cov(x, a)}{Var(x)} \equiv \gamma \theta_1 \neq 0$$

  entonces nuestro estimador MCO es inconsistente: ¡$\hat{\beta}_{1,MCO}$ se acerca cada vez más y más a un valor poblacional erróneo!

Outline

Una ilustración del proceso de endogeneidad

- Para comprender el estimador de variables instrumentales, es útil usar un **DAG**. Este muestra la cadena de efectos causales de nuestro modelo
- En nuestro modelo corto $y = \beta_0 + \beta_1 x + \varepsilon$ hemos visto que $Cov(x, \varepsilon) = E(x\varepsilon) = \gamma \underbrace{E(xa)}_{\neq 0} + \underbrace{E(xu)}_{=0} \neq 0$. Esto puede expresarse como

$$
\begin{tikzcd}
x \arrow[r] & {} \arrow[r] & {} \arrow[r] & y \\
& \varepsilon = \gamma a + u \arrow[ul] \arrow[urr] & &
\end{tikzcd}
$$

  Estamos interesados en el efecto directo de x sobre y, $x \rightarrow y$
- ¿Pero qué está ocurriendo en nuestro caso?
  - Queremos que el estimador $\hat{\beta}_{1,MCO}$ capture el efecto sobre y que *únicamente* puede atribuirse a x
  - Pero en el error compuesto $\varepsilon = \gamma a + u$ existe *algo*, a, que está correlacionado con x ($\theta > 0$) y además $\gamma > 0$
- Entonces, el efecto de x sobre y está contaminado por el efecto que la habilidad a tiene sobre y

El instrumento

- Supongamos que existe una variable z correlacionada con x de forma **peculiar**

$$
\begin{array}{ccccccc}
z & \rightarrow & x & \rightarrow & \rightarrow & \rightarrow & y \\
 & & & \nwarrow & & \nearrow & \\
 & & & & \varepsilon & &
\end{array}
$$

- Nótese que z afecta a y «**sólo a través de**» x
  - Cambios en z inducen cambios en x, lo que hace que y cambie
  - Entonces y cambia solo porque x cambia: los únicos cambios que z genera son cambios en x, no directamente en y
- Imaginemos que x es un decisión que la persona toma
  - Esta decisión afecta a y, pues $x \rightarrow y$
  - Pero a veces estas decisiones se deben a cambios en ε pues $\varepsilon \rightarrow x \rightarrow y$
- Si ocurre un cambio exógeno, inesperado e independiente de x, dado por z, ¿qué efecto tiene?
  - z induce a algunos a tomar una decisión x, que genera un cambio en y, pues $z \rightarrow x \rightarrow y$
  - El último tramo, $x \rightarrow y$, es el efecto causal, inducido por z vía $z \rightarrow x$

La exclusion restriction

- Supongamos que existe una variable z correlacionada con x de forma **peculiar**

$$\begin{array}{ccccccccc} z & \rightarrow & x & \rightarrow & & \rightarrow & & \rightarrow & y \\ & & & \nwarrow & & & & \nearrow & \\ & & & & & \varepsilon = \gamma a + u & & & \end{array}$$

- Nótese que z ***no* tiene conexión con** ε
  - x está determinado por un factor en ε que no podemos controlar (a)
  - z, el instrumento, no debe estar relacionado con este factor no observado
- Para que z sea tomado en serio como instrumento debe cumplir con la «**exclusion restriction**» (ER). Esto implica 2 condiciones
  1. C1: z no está correlacionado con y directamente sino sólo a través de su efecto en x
  2. C2: z debe ser independiente de cualquier variable que determine y salvo x

La primera condición de la ER

- z no está correlacionado con y directamente sino sólo a través de su efecto en x

$$z \quad \rightarrow \quad x \quad \rightarrow \quad \rightarrow \quad \rightarrow \quad y$$

$$\varepsilon$$

- Nótese que
  - *La única forma* como z induce cambios sobre y es porque z induce cambios en x que a su vez generan cambios en y

$$z \rightarrow x \rightarrow y$$

  - I.e. z *no* es un determinante directo de y, su influencia es sólo indirecta. Por tanto z puede *omitirse* de la ecuación que determina y
  - Cualquier conexión entre z e y debe pasar a través del tratamiento x

La segunda condición de la ER

- z debe ser independiente de cualquier variable que determine y salvo x
- Digamos que z afecta indirectamente y por 2 caminos

$$
\begin{array}{ccccccc}
 & & \boldsymbol{w} & \to & \to & \to & \\
 & \nearrow & & & & & \downarrow \\
z & \to & x & \to & \to & \to & y \\
 & & & \nwarrow & & \nearrow & \\
 & & & & \varepsilon & &
\end{array}
$$

- En este caso
  - $z \to x \to y$: z genera un cambio en x, y este cambio en x genera cambio en y. Este es el efecto que puede atribuirse a x y que queremos capturar en nuestro estimado
  - $z \to w \to y$: z *también* genera un cambio en w, y este cambio en w genera un cambio en y. Este es un efecto *adicional* que *no queremos capturar* en nuestro estimado

  Cambios en z inducen cambios en x y w, y entonces mezcla el efecto $x \to y$ con el efecto $w \to y$. Nuestro estimado no sólo mide $x \to y$
- Si esto ocurre ¡debemos incluir w como variable de control!

¿Cómo sabemos que tenemos un buen instrumento?

- Solo se puede identificar un efecto causal utilizando IV si puede defender teórica y lógicamente la ER. Esta no puede probarse empíricamente sino *sólo argumentarse*
- Usualmente, un instrumento z satisface la ER cuando, al explicarlo a un interlocutor, éste *no entiende a la primera* el mecanismo que relaciona z con la dependiente y
- ¿Por qué?
    - Porque si z cumple la restricción de exclusión, entonces z *no* es un determinante directo de y. La ER *asegura que z e y parezcan* a primera vista (y para un interlocutor cualquiera) *completamente independientes.*
    - Para ver su ligazón es necesaria una explicación de cómo z afecta a y a través de x. Esta es la justificación teórica y lógica que da validez al instrumento.

Por tanto, en virtud de la ER, *un buen instrumento para x no puede ser un determinante obvio de y*

Outline

La ecuación estructural de interés

- Imaginemos que queremos estimar la relación entre el tamaño del hogar con el número de horas que una mujer trabaja

$$\text{horas} = \beta_0 + \beta_1 \text{numhijos} + \varepsilon$$

Este modelo (corto) deja de lado algunos elementos que

1. tienen efecto parcial sobre el número de horas que la mujer trabaja
2. están relacionados con el número de hijos que una mujer tiene

- ¿Qué variable no incluida en el modelo cumple con estas condiciones? E.g. *actitud al trabajo*

$$\text{horas} = \beta_0 + \beta_1 \text{numhijos} + \gamma \text{actitud} + u$$

En tanto $\gamma \neq 0$ y $\frac{Cov(\text{numhijos,actitud})}{Var(\text{numhijos})} \neq 0$, actitud es una variable omitida que crea problema de inconsistencia

- Podemos expresar esto como

$$\begin{array}{ccccc} \text{numhijos} & \rightarrow & \rightarrow & \rightarrow & \text{horas} \\ & \nwarrow & & \nearrow & \\ & & \varepsilon = \gamma \text{actitud} + u & & \end{array}$$

- Por tanto necesitamos un instrumento para estimar la ecuación de interés

Introduciendo el instrumento

- Para que un instrumento z cumpla con la ER debe ocurrir que

$$\begin{array}{ccccccc} z & \rightarrow & \text{numhijos} & \rightarrow & \rightarrow & \rightarrow & \text{horas} \\ & & & \nwarrow & & \nearrow & \\ & & & & \varepsilon & & \end{array}$$

  A&E proponen el gender mix: secuencia sexual de los 2 primeros hijos
- ¿Qué pensaría su amigo si les dicen que «el numhijos afecta el número de horas que la mujer trabaja»?
  - Probablemente dirá «obvio, eso ya lo sabía»
- ¿Pero qué pensaría si le dicen que «las madres cuyos 2 primeros hijos son del mismo sexo (z) trabajan menos que aquellas cuyos hijos tienen una proporción de sexos equilibrada»?
  - Probablemente dirá confundido «¡¿qué tiene que ver el gender mix con el número de horas que la mujer trabaja?!»
  - Pero si luego de decirles que «el género de los 2 primeros niños induce a las mujeres a tener familias más grandes, lo que lleva a que trabajen menos» la persona entiende, entonces (en principio) tenemos un muy buen instrumento

La primera condición de la ER

1. z está correlacionado con y directamente sino sólo a través de su efecto en x:
   - El gender mix (z)
     1.1 *No* influye directamente en el número de horas (y), i.e. z *no* es un determinante directo de la dependiente y. Por tanto puede omitirse de la ecuación para y
     1.2 Afecta el número de hijos (x): usualmente las personas cuyos 2 primeros hijos son del mismo sexo deciden tener un tercer hijo
   - Es este último camino que justifica que $z \rightarrow y$ *sólo por la influencia que tiene sobre* x, i.e. cumple la parte de "**solo a través de**" de la definición de la ER

La segunda condición de la ER

2. z debe ser independiente de cualquier variable que determine y salvo x.
   - ¿Es posible que el gender mix mueva otro factor w que a su vez mueva el número de horas?

$$
\begin{array}{ccccccc}
 & & \boldsymbol{w} & \rightarrow & \rightarrow & \rightarrow & \\
 & \nearrow & & & & & \downarrow \\
z & \rightarrow & \text{numhijos} & \rightarrow & \rightarrow & \rightarrow & \text{horas} \\
 & & & \nwarrow & & \nearrow & \\
 & & & & \varepsilon & &
\end{array}
$$

   - No parece ser el caso: no hay ningún otro elemento que z mueve salvo el número de hijos
   - Si fuera el caso que también $z \rightarrow w \rightarrow y$, entonces w debe incluirse en la regresión

- Esta última consideración la veremos en la siguiente subsección

Outline

1. Motivación
2. La endogeneidad y el instrumento
   - La lógica del modelo de regresión
   - La falla del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$
   - Entendiendo a la variable instrumental
   - Ilustración: Angrist y Evans (1996)
   - Cómo los instrumentos suelen fallar
3. Efectos de tratamiento homogéneos
   - Estimadores de variables instrumentales
   - Ilustración: Angrist y Krueger (1991)
4. Efectos de tratamiento heterogéneos
   - El estimador de Wald
   - Ilustración: Angrist (1990)
5. Aspectos técnicos y problemas
   - El problema de instrumentos débiles
   - Dos (o tres) tests importantes
6. Conclusión

El estado de la situación

«(...) these days, go to any seminar where an IV paper is presented and you'll hear no end of worries and arguments about whether the instrument is valid. And as time goes on, it seems like people have gotten more and more difficult to convince when it comes to validity»
Huntington-Klein, N. (2021) The effect an introduction to research design and causality

- ¡Hay una buena razón para estar preocupado! Es difícil de justificar que existe una variable
  1. fuertemente relacionada con el tratamiento
  2. que de alguna manera no está *en absoluto* relacionado con otros factores que también pueden afectar a y

  ver `http://econometricsense.blogspot.com/2015/12/do-friends-let-friends-do-ivor-is-all.html`
- De hecho, muchos instrumentos que antes eran creíbles ahora son motivo de memes. Veamos algunos ejemplos

Acemoglu y Johnson (2007)

- En Disease and development, JPE, 115(6), A&J estiman el efecto de la mortalidad (x) en el crecimiento económico (y)
  - Instrumentan la mortalidad por el momento en que la nueva tecnología médica (e.g. vacunas) se introduce en el país (z)

$$
\begin{array}{ccccccc}
\text{intro tech médica} & \rightarrow & \text{mortalidad} & \rightarrow & \rightarrow & \rightarrow & \text{crecimiento} \\
 & & & \nwarrow & & \nearrow & \\
 & & & & \varepsilon & &
\end{array}
$$

  - Encuentran un efecto negativo
- Bloom et al. (2014) Disease and development revisited. JPE, 122(6) señalan que en realidad la tecnología médica está relacionada con factores de salud y económicos preexistentes que están relacionados con el crecimiento económico

$$
\begin{array}{cccccccc}
 & & \text{factores pre-existentes} & \longrightarrow & \longrightarrow & \longrightarrow & \\
 & \nearrow & & & & & \downarrow \\
\text{intro tech médica} & \rightarrow & \text{mortalidad} & \rightarrow & \rightarrow & \rightarrow & \text{crecimiento} \\
 & & & \nwarrow & & \nearrow & \\
 & & & & \varepsilon & &
\end{array}
$$

y al controlar por estos factores preexistentes encuentran en vez un efecto positivo

Miguel, Satyanath, and Sergenti (2004)

- En Economic shocks and civil conflict, KPE, 112(4), Miguel et al. estiman el efecto de crecimiento económico (x) en el conflicto civil (y)
  - Utilizan la lluvia (z) como instrumento para crecimiento económico en África, y analizan cómo el crecimiento económico afecta el conflicto civil
  - Encuentran un efecto bastante grande: caída del 5% aumenta las posibilidades de guerra en 50%

$$
\begin{tikzcd}
\text{lluvia} \arrow[r] & \text{crecimiento} \arrow[r] & {} \arrow[r] & {} \arrow[r] & \text{conflicto} \\
 & & \varepsilon \arrow[ul] \arrow[ur] & &
\end{tikzcd}
$$

- Sarsons (2015) argumenta que la lluvia no es un instrumento válido, pues la lluvia afecta el conflicto a través de otros mecanismos tales como migración

$$
\begin{tikzcd}
 & \text{migración} \arrow[r] & {} \arrow[r] & {} \arrow[r] & {} \arrow[d] \\
\text{lluvia} \arrow[ur] \arrow[r] & \text{crecimiento} \arrow[r] & {} \arrow[r] & {} \arrow[r] & \text{conflicto} \\
 & & \varepsilon \arrow[ul] \arrow[ur] & &
\end{tikzcd}
$$

Un estudio creíble debería incluir migración en el modelo

El problema de los instrumentos manidos

- La cantidad de lluvia ha sido uno de los z más usados. Ha instrumentalizado *casi todo*, desde ingresos personales a actividad económica
  - En general, una variable z que ha sido usada por diferentes papers para instrumentalizar diferentes variables debe ser motivo de sospecha
- Imaginemos que queremos analizar $x \rightarrow y$ pero x es endógena. Usamos z como instrumento para x. Consultamos 3 papers que han analizado y:
  - Paper 1 analiza $w \rightarrow y$ y w es endógena. Éste instrumentaliza w usando z

$$\begin{array}{ccccccc} z & \rightarrow & w & \rightarrow & \rightarrow & \rightarrow & y \\ & & & \nwarrow & & \nearrow & \\ & & & & \varepsilon & & \end{array}$$

  - Paper 2 analiza $m \rightarrow y$ y m es endógena. Éste instrumentaliza m usando z

$$\begin{array}{ccccccc} z & \rightarrow & m & \rightarrow & \rightarrow & \rightarrow & y \\ & & & \nwarrow & & \nearrow & \\ & & & & e & & \end{array}$$

  - Paper 3 analiza $k \rightarrow y$ y k es endógena. Éste instrumentaliza k usando z

$$\begin{array}{ccccccc} z & \rightarrow & k & \rightarrow & \rightarrow & \rightarrow & y \\ & & & \nwarrow & & \nearrow & \\ & & & & u & & \end{array}$$

El problema de los instrumentos manidos

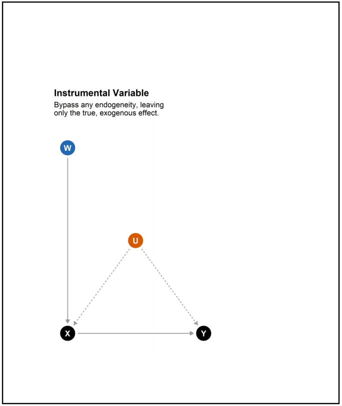
- Si estos 3 papers son válidos, entonces en realidad tenemos que

$$
\begin{array}{ccccccc}
k & \longrightarrow & \longrightarrow & \longrightarrow & \longrightarrow & & \\
\uparrow & & m & \longrightarrow & \longrightarrow & \searrow & \\
\uparrow & \nearrow & w & \longrightarrow & \longrightarrow & \searrow & \downarrow \\
\uparrow & \nearrow & & & & \searrow & \downarrow \\
z & \rightarrow & x & {\color{green}\rightarrow} & {\color{green}\rightarrow} & {\color{green}\rightarrow} & y \\
 & & & {\color{red}\nwarrow} & & {\color{red}\nearrow} & \\
 & & & & \varepsilon & &
\end{array}
$$

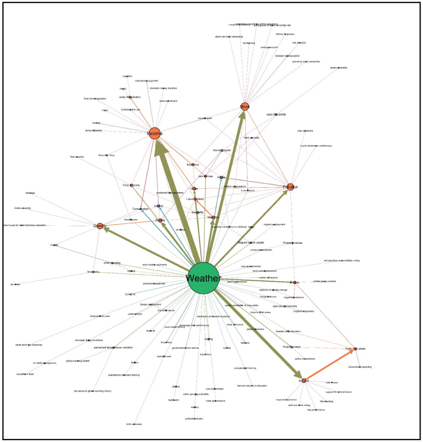
  Esto implica que z no sólo afecta el x de interés, sino también otros factores como w, m, k etc., cada uno de los cuales afectan a y
- Por tanto, para obtener el efecto $x \rightarrow y$ de interés, debemos *controlar por todos estos otros tratamientos* que también son afectados por el instrumento y que afectan a y
  - Pero controlar por todos estos factores w, m, k es a menudo *imposible*
  - Aun así, muchos papers usan $z =$ lluvia para (e.g.) $y =$ ingresos agricolas pero olvidan que hay varios otros mecanismos que afectan y más allá de su influencia a través de x
  - Si bien la primera condición para la ER se cumple, la segunda no. Y entonces el instrumento es inválido

Using weather as an IV

Expectations:



Reality:



Outline

Retomando: 1 regresor endógeno y 1 instrumento

- Vimos que si el modelo verdadero es $y = \beta_0 + \beta_1 x + \gamma a + u$ pero estimamos en vez

$$y = \beta_0 + \beta_1 x + \underbrace{\varepsilon}_{\equiv \gamma a + u}$$

  la habilidad a pasa a ser no observada (y por tanto pasa a estar en el error compuesto $\varepsilon = \gamma a + u$). Debido a que la habilidad

  1. tiene un efecto *parcial* (positivo) sobre y $(\gamma = 0)$
  2. está correlacionada con x $(\theta_1 \neq 0$ en $a = \theta_0 + \theta_1 x + v)$

  el estimador MCO, $\hat{\beta}_{1,MCO}$, es inconsistente

- Ahora que ya sabemos lo que es un instrumento, asumamos que tenemos un instrumento
- ¿Cómo podemos derivar un nuevo estimador con este instrumento?

Retomando: 1 regresor endógeno y 1 instrumento

- Una manera formal de decir que nuestro instrumento cumple con la ER

$$\begin{array}{ccccccc} z_1 & \rightarrow & x_1 & \rightarrow & \rightarrow & \rightarrow & y \\ & & & \nwarrow & & \nearrow & \\ & & & & \varepsilon & & \end{array}$$

  está dada por

$$Cov(z_1, x_1) \neq 0 \text{ y } Cov(z_1, \varepsilon) = 0$$

  Estas 2 condiciones son ***cruciales***

  1. $Cov(z_1, x_1) \neq 0$ es el **supuesto de relevancia**: z_1 afecta x_1: $z_1 \rightarrow x_1$. Esto *sí* se puede testear directamente
  2. $Cov(z_1, \varepsilon) = 0$ es el **supuesto de ortogonalidad/ER**: z_1 no está correlacionado con ningún otro factor que afecte a y. Esto *no* se puede testear

- Volvamos al ejemplo: nuestra ecuación estimable (corta) es $y = \beta_0 + \beta_1 x_1 + \varepsilon$. Tomemos la covarianza entre y y z_1

$$Cov(y, z_1) = Cov(\beta_0 + \beta_1 x_1 + \varepsilon, z_1) = \underbrace{Cov(\beta_0, z_1)}_{=0} + \beta_1 Cov(x_1, z_1) + \underbrace{Cov(\varepsilon, z_1)}_{=E\left(z'\varepsilon\right)=0} = \beta_1 Cov(x_1, z_1) \Leftrightarrow Cov(y, z_1) = \beta_1 Cov(x_1, z_1)$$

Retomando: 1 regresor endógeno y 1 instrumento

- A partir de $Cov(y, z_1) = \beta_1 Cov(x_1, z_1)$ podemos hallar la **expresión poblacional del estimador IV** para este modelo simple

$$\beta_{1,IV} = \frac{Cov(y, z_1)}{Cov(x_1, z_1)}$$

la cual se asemeja a la del estimador MCO para el mismo modelo sin controles $y = \beta_0 + \beta_1 x_1 + \varepsilon$: $\beta_{1,MCO} = \frac{Cov(y,x_1)}{Var(x_1)} = \frac{Cov(y,x_1)}{Cov(x_1,x_1)}$

- A diferencia de antes, la muestra está ahora compuesta por 3 cantidades

$$\left\{\left(y_i, x_i, \boxed{z_i}\right) : i = 1, \ldots, N\right\}$$

y usando **el principio de analogía** de Manski, hallamos la contraparte muestral de $\beta_{1,IV}$

$$\beta_{1,IV} = \frac{Cov(y, z_1)}{Cov(x_1, z_1)} \underbrace{\rightarrow}_{\text{princip. analog.}} \boxed{\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{N}(z_{1i} - \bar{z}_1)(y_i - \bar{y})}{\sum_{i=1}^{N}(z_{1i} - \bar{z}_1)(x_{1i} - \bar{x}_1)}}$$

Esta expresión nos provee el **estimador de variables instrumentales** $\hat{\beta}_{IV}$ para este modelo simple sin controles, que proviene de reemplazar

- promedios poblacionales $Cov(y, z_1)$ y $Cov(x_1, z_1)$ que no observamos, por
- promedios muestrales, $\sum_{i=1}^{N}(z_{1i} - \bar{z}_1)(y_i - \bar{y})$ y $\sum_{i=1}^{N}(z_{1i} - \bar{z}_1)(x_{1i} - \bar{x}_1)$ que sí observamos

Generalizando: 1 regresor endógeno y 1 instrumento

- ¿Cómo cambia el estimador IV si agregamos controles? Asumamos ahora que $x_1,...,x_{K-1}$ son las **variables de control** y x_K es el regresor endógeno

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \rightarrow y = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \varepsilon \equiv \mathbf{x}' \beta + \varepsilon$$

entonces el vector de variables de la ecuación estructural es

$$\mathbf{x} = \left(1, x_1, x_2, ..., x_{K-1}; \boxed{x_K}\right)_{1\times(1+K)}$$

  - $x_1,...,x_{K-1}$ son $K-1$ **variables de control**, *no* correlacionadas con ε
  - x_K es el regresor endógeno, correlacionada con algún factor en ε
- Asumamos también que
  - Seguimos teniendo 1 instrumento z_1 para el regresor endógeno x_K. En el presente contexto, z es un **instrumento excluido**
  - ¿Qué son los **instrumentos incluidos**? Son las variables $x_1,...,x_{K-1}$ que son exógenas. No necesitan ser instrumentalizadas; son instrumentos de sí mismas

  Entonces el **vector de instrumentos z** está compuesto por *todas* las variables exógenas

$$\mathbf{z}' = \left(1, x_1, x_2, ..., x_{K-1}; \boxed{z_1}\right)_{1\times(1+K)}$$

Estas satisfacen la **condición de ortogonalidad del vector de instrumentos**: $Cov(\mathbf{z}, \varepsilon) = E(\mathbf{z}\varepsilon) = 0$

Generalizando: 1 regresor endógeno y 1 instrumento

- Hagamos igual que para derivar el estimador MCO: pre-multipliquemos por $\mathbf{z}$ la ecuación estructural en error-form y tomemos expectativas

$$y = \mathbf{x}^{'}\beta + \varepsilon \Leftrightarrow E(\mathbf{z}y) = E\left(\mathbf{z}\mathbf{x}^{'}\beta\right) + E(\mathbf{z}\varepsilon) \Leftrightarrow E(\mathbf{z}y) = E\left(\mathbf{z}\mathbf{x}^{'}\right)\beta + E(\mathbf{z}\varepsilon)$$

  y por la exogeneidad de $\mathbf{z}$

$$E(\mathbf{z}y) = E\left(\mathbf{z}\mathbf{x}^{'}\right)\beta + \underbrace{E(\mathbf{z}\varepsilon)}_{=0} \Leftrightarrow \beta_{IV} = \left[E\left(\mathbf{z}\mathbf{x}^{'}\right)\right]^{-1} E(\mathbf{z}y)$$

- Esta es la **expresión poblacional del estimador IV**
  - Define β como una función de valores en la **población**: $E\left(\mathbf{z}\mathbf{x}^{'}\right)$, $E(\mathbf{z}y)$.
  - β_{IV} está identificado si la matriz $\left[E\left(\mathbf{z}\mathbf{x}^{'}\right)\right]^{-1}$ existe y en ese caso β_{IV} tiene un valor único. Para que $\left[E\left(\mathbf{z}\mathbf{x}^{'}\right)\right]^{-1}$ exista $E\left(\mathbf{z}\mathbf{x}^{'}\right)$ debe ser diferente de 0

$$E\left(\mathbf{z}\mathbf{x}^{'}\right) \neq 0$$

    esta es la representación matemática de la **condición de relevancia del vector de instrumentos** en este caso general. Esto generaliza la condición $Cov(z_1, x_1) \neq 0$ del modelo simple

Generalizando: 1 regresor endógeno y 1 instrumento

- Tenemos ahora una muestra compuesta ahora por 3 cantidades

$$\{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, ..., N\}$$

y usando **el principio de analogía** de Manski, hallamos la contraparte muestral de β_{IV}

$$\beta_{IV} = \left[E\left(\mathbf{z}\mathbf{x}^{'}\right)\right]^{-1} E\left(\mathbf{z}y\right) \underbrace{\rightarrow}_{\text{princip. analog.}} \boxed{\hat{\beta}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\mathbf{x}_i^{'}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i y_i = \left(\sum_{i=1}^{N} \mathbf{z}_i\mathbf{x}_i^{'}\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i}$$

Esta expresión nos provee el **estimador de variables instrumentales**, $\hat{\beta}_{IV}$, para este caso *más general*, que proviene de reemplazar

  - promedios poblacionales $E\left(\mathbf{z}\mathbf{x}^{'}\right)$ y $E\left(\mathbf{z}y\right)$ que no observamos, por
  - promedios muestrales, $\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i\mathbf{x}_i^{'}$ y $\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i y_i$ que sí observamos

- Nótese cómo el estimador IV con o sin controles requiere el cumplimiento de la condición de exclusión: ortogonalidad y relevancia

Generalizando: 1 regresor endógeno y M instrumentos

- Digamos que en nuestro modelo con $x_1, ..., x_{K-1}$ controles

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \varepsilon \equiv \mathbf{x}^{'} \beta + \varepsilon$$

  ya no sólo tenemos 1 instrumento para x_K sino M instrumentos *excluidos* $z_1, z_2, ..., z_M$.

- Ahora el vector de instrumentos $\mathbf{z}$ está compuesto ahora por
  - $x_1, x_2, ..., x_{K-1}$ **instrumentos incluidos**
  - $z_1, z_2, ..., z_M$ **instrumentos excluidos**

  Es decir, el vector de instrumentos pasa a ser

$$\mathbf{z}^{'} = \left(1, x_1, x_2, ..., x_{K-1}; \boxed{z_1}\right)_{1\times(1+K)} \rightarrow \mathbf{z}^{'} = \left(1, x_1, x_2, ..., x_{K-1}; \boxed{z_1, z_2, ..., z_M}\right)_{1\times(1+(K-1)+M)}$$

- Si cada uno de estos $z_1, ..., z_M$ tiene alguna correlación parcial con x_K, podríamos tener M estimadores IV $\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{'}\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$

Generalizando: 1 regresor endógeno y M instrumentos

- E.g. si $K = 3$, con x_3 regresor endógeno,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

  entonces

$$\mathbf{x} = (1, x_1, x_2; x_3)_{1\times(1+3)}$$

  y si $M = 3$, entonces

$$\mathbf{z} = (1, x_1, x_2; z_1, z_2, z_3)_{1\times(1+2+3)}$$

- Esto significa que tenemos 3 posible estimadores IV
  - si usamos $\mathbf{z}_1^{'} = (1, x_1, x_2, z_1)_{1\times(1+2+1)}$ obtenemos $\hat{\beta}_{z_1,IV}$
  - si usamos $\mathbf{z}_2^{'} = (1, x_1, x_2; z_2)_{1\times(1+2+1)}$ obtenemos $\hat{\beta}_{z_2,IV}$
  - si usamos $\mathbf{z}_3^{'} = (1, x_1, x_2; z_3)_{1\times(1+2+1)}$ obtenemos $\hat{\beta}_{z_3,IV}$
- Pero al hacer esto dejamos de lado información que los otros instrumentos proveen, lo que genera ineficiencia ¿Qué hacer entonces?
- Podemos combinar la información que todos los instrumentos producen en 1 sola variable que usamos luego como instrumento

Generalizando: 1 regresor endógeno y M instrumentos

- El nuevo instrumento corresponde a la combinación lineal de los instrumentos excluidos que está más fuertemente correlacionada con x_K. ¿Cómo? Estimando la **forma reducida** para x_K, que la regresiona contra *todas* las exógenas en $\mathbf{z}$

$$x_K = \theta_0 + \underbrace{\theta_1 x_1 + \ldots + \theta_{K-1} x_{K-1}}_{\text{instrumentos incluidos}} + \underbrace{\theta_1 z_1 + \ldots + \theta_M z_M}_{\text{M instrumentos excluidos}} + r = \boldsymbol{\theta}\mathbf{z} + r$$

  Nótese que x_K puede ser continua, binaria o discreta.

- En este caso nuestro modelo a estimar está compuesto por **2 ecuaciones**

$$\text{ec. estructural:} y = \mathbf{x}'\beta + \varepsilon = \beta_0 + \beta_1 x_1 + \ldots + \beta_{K-1} x_{K-1} + \beta_K x_K + \varepsilon$$

$$\text{primera etapa:} x_K = \mathbf{z}'\boldsymbol{\theta} + r = \theta_0 + \theta_1 x_1 + \ldots + \theta_{K-1} x_{K-1} + (\theta_{z_1} z_1 + \ldots + \theta_{z_M} z_M) + r$$

  Entonces

  1. Obtenemos los valores predichos $\hat{x}_K$ estimando la primera etapa por MCO

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

     nótese cómo se incluyen los instrumentos excluidos y también los incluidos

  2. Estimamos por MCO el modelo estructural reemplazando x_k por $\hat{x}_K$

  $$y \text{ sobre } 1, x_1, \ldots, x_{K-1}, \boxed{\hat{x}_K}$$

Generalizando: 1 regresor endógeno y M instrumentos

- Este proceso de 2 etapas provee el estimador de **two-stages least squares, 2SLS**

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

- Comparemos $\hat{\beta}_{2SLS} = \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \mathbf{x}_i \right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i y_i$ con $\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{'} \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$
  - El estimador IV se usa cuando el número de instrumentos excluidos es igual al número de variables endógenas. En este caso, el modelo está **exactamente identificado**
    - Hemos asumido 1 endógena y 1 instrumento, pero podríamos haber asumido, e.g., 2 endógenas y 2 instrumentos
  - El estimador 2SLS *sólo es necesario* cuando tenemos más instrumentos excluidos que variables endógenas. En este caso, el modelo está **sobreidentificado**
    - Si usamos $\hat{\beta}_{2SLS}$ en un modelo exactamente identificado éste colapsa en $\hat{\beta}_{IV}$
- Sea que usemos el estimador IV o 2SLS, debemos cumplir la **condición de orden**: número de *instrumentos* excluidos debe ser *no menor* al número de endógenas, L. Hemos asumido $L = 1$, pero puede que e.g. $L = 3$ y entonces necesitamos *al menos* 3 instrumentos

Desentrañando el estimador 2SLS

- Pero el estimador 2SLS nos dice algo más sobre la mecánica sobre la cual los instrumentos (válidos) funcionan

1. El estimador 2SLS $\hat{\beta}_{2SLS} = \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \mathbf{x}_i\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i \mathbf{y}_i$ usa solo los valores ajustados de los regresores endógenos para la estimación.
   - Estos valores predichos se basa en el vector $\mathbf{z}$ que incluye *todas* las variables **exógenas** en el modelo estructural

     $$\hat{x}_K = \mathbf{z}_i^{'} \boldsymbol{\theta} = \hat{\theta}_0 + \left(\hat{\theta}_1 x_1 + ... + \hat{\theta}_{K-1} x_{K-1}\right) + \left(\hat{\theta}_{z_1} z_1 + ... + \hat{\theta}_{z_M} z_M\right)$$

     y entonces los valores ajustados $\hat{x}_K$ también son exógenos
   - Por tanto, identificamos el efecto de educación usando aquella parte de la variación en la educación que es *exógena*
   - Esto permite volver al mundo ideal en el que identificamos efectos causales basado en información exógena
2. Pero no todo es felicidad: la variación exógena en x impulsada por $\mathbf{z}$ es solo un subconjunto de la variación total en la variable misma. O dicho de otra manera, IV reduce la variación en los datos, por lo que hay menos información disponible para la identificación

Outline

1. Motivación
2. La endogeneidad y el instrumento
   - La lógica del modelo de regresión
   - La falla del supuesto $Cov(\varepsilon, \mathbf{x}) = 0$
   - Entendiendo a la variable instrumental
   - Ilustración: Angrist y Evans (1996)
   - Cómo los instrumentos suelen fallar
3. Efectos de tratamiento homogéneos
   - Estimadores de variables instrumentales
   - Ilustración: Angrist y Krueger (1991)
4. Efectos de tratamiento heterogéneos
   - El estimador de Wald
   - Ilustración: Angrist (1990)
5. Aspectos técnicos y problemas
   - El problema de instrumentos débiles
   - Dos (o tres) tests importantes
6. Conclusión

La variación exógena en la educación

- Angrist y Krueger (1991) estiman el retorno a la educación usando como instrumento el trimestre (quarter of birth) en que nació la persona
- En USA el año escolar empieza a fines de agosto y termina en mayo del año siguiente. La ley americana estipula que los niños deben entrar a 1er grado el año calendario que cumplen 6. Si el niño nació
  - entre 1 de Julio y 31 de Dic., entra al 1er grado
  - entre 1 de Enero y 30 de Junio, entra al kindegarden

  Una persona nacida el 31 de Diciembre y aquella nacida el 1er de Enero se le asignó exógenemente diferentes grados
- ¿Cómo esto afecta el hecho de terminar el colegio?
  - USA tiene leyes de escolarización obligatoria: obligan a permanecer en la secundaria *hasta cumplir los 16 años*
  - Después de cumplir los 16 años, se puede abandonar la escuela legalmente.

  Entonces aquellos nacidos en el 1er o 2do trimestre pueden abandonar la escuela antes de completar el año lectivo, teniendo menos escolaridad que aquellos nacidos en el 3er y 4to trimestre.

La variación exógena en la educación

$$\text{trimestre} \quad \rightarrow \quad \text{educación} \quad \rightarrow \quad \rightarrow \quad \rightarrow \quad \text{salario}$$

$$\nwarrow \qquad \nearrow$$

$$\varepsilon$$

- Recuérdese que los buenos instrumentos requieren siempre una explicación porque a primera vista no están relacionados con y. Eso es precisamente lo que implica la ER.
- ¿Por qué el trimestre de nacimiento afectaría los salarios?
  - No tiene ningún sentido lógico y obvio por qué debería hacerlo.
  - Pero, si uno explica al interlocutor que quienes nacieron más tarde en el año (Q3, 4) obtuvieron *más educación (x)* debido a la escolarización obligatoria, entonces la relación (indirecta) entre el instrumento y los salarios tiene sentido
- La segunda condición de la ER implica argumentar que el instrumento z afecta y sólo por su efecto en inducir cambios en educación x. Este parece ser el caso.

La primera etapa

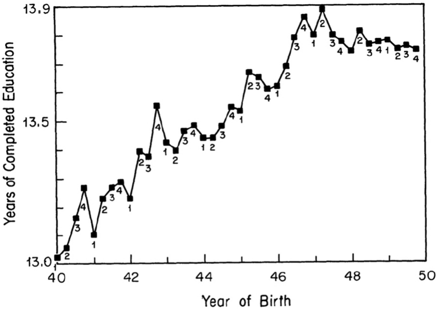


Figure 49. First-stage relationship between quarter of birth and schooling. Reprinted from Cunningham, S. and Finlay, K. (2012). "Parental Substance Abuse and Foster Care: Evidence from Two Methamphetamine Supply Shocks?" *Economic Inquiry*, 51(1):764–782. Copyright © 2012 Wiley. Used with permission from John Wiley and Sons.

La primera etapa

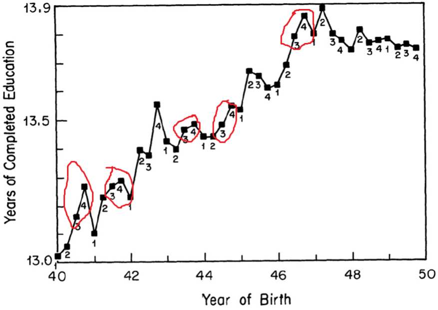


Figure 49. First-stage relationship between quarter of birth and schooling. Reprinted from Cunningham, S. and Finlay, K. (2012). "Parental Substance Abuse and Foster Care: Evidence from Two Methamphetamine Supply Shocks?" *Economic Inquiry*, 51(1):764–782. Copyright © 2012 Wiley. Used with permission from John Wiley and Sons.

- En promedio, los nacidos en el 3er y 4to trimestre tienen más escolaridad que los nacidos en 1er y 2do
- Esa relación se debilita en las cohortes posteriores, porque el retorno de niveles más altos de educación aumentó en el tiempo, desincentivando el abandono de cohortes posteriores

La forma reducida

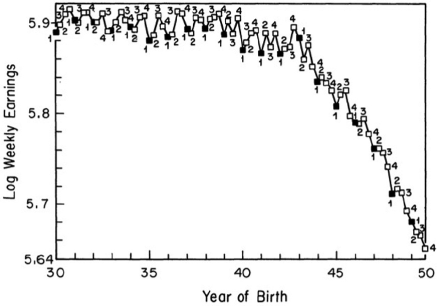


Figure 50. Reduced-form visualization of the relationship between quarter of birth and schooling. Reprinted from Angrist, J. D. and Krueger, A. B. (1991). “Does Compulsory School Attendance Affect Schooling and Earnings?” *Quarterly Journal of Economics*, 106(4):979–1014. Permission from Oxford University Press.

La forma reducida

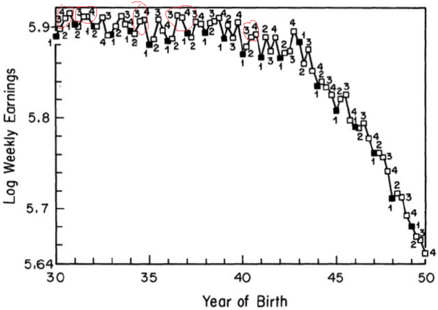


Figure 50. Reduced-form visualization of the relationship between quarter of birth and schooling. Reprinted from Angrist, J. D. and Krueger, A. B. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4):979–1014. Permission from Oxford University Press.

- En general los nacidos en el 3er y 4to trimestre tienen mayor ingresos y los nacidos en 1er y 2do trimestre tienen menos ingresos

Los resultados

Independent variable	OLS	2SLS
Years of schooling	0.0711	0.0891
	(0.0003)	(0.0161)
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No

Note: Standard errors in parentheses. First stage is quarter of birth dummies.

- Su modelo es

$$\ln y = \beta_0 + \beta_1 \text{esc} + \gamma \textbf{year} + \varepsilon$$
$$\text{esc} = \theta_0 + z_1\theta_1 + z_2\theta_2 + z_3\theta_3 + \gamma \textbf{year} + r$$

  donde y es log del salario por hora, **year** son dummies por año de nacimiento, los instrumentos son 3 dummies: para el 1er (z_1), 2do (z_2) y 3er (z_3) trimestre (categoría omitida es 4to trimestre)
- Encuentran un retorno del 7,1 % por cada año adicional de escolaridad, pero con 2SLS es más alto (8,9 %).

Outline

Efectos tratamiento heterogéneos

- Ahora pasamos a la pedagogía más contemporánea donde relajamos la suposición de que los efectos del tratamiento *son los mismos para todos*. Ahora permitiremos que cada individuo tenga una respuesta *diferente* al efecto tratamiento, δ_i

$$y_1 - y_0 = \delta \quad \underbrace{\rightarrow}_{\text{generalizando}} \quad y_{i1} - y_{i0} = \delta_i$$

- Una vez que introducimos efectos de tratamiento **heterogéneos**, introducimos una distinción entre
  - la validez interna: ¿qué parámetro poblacional identifica $\hat{\beta}_{1,IV}$? ¿eso que estimamos, es lo que queríamos estimar?
  - la validez externa: ¿son los hallazgos extrapolables a otras poblaciones aparte de las que estudiamos?
- Como veremos, bajo efectos tratamiento
  - **Homogéneos**, todas las personas reaccionan *igual* ante x y ante z. El efecto que recuperamos, $\hat{\beta}_{1,IV}$, es el efecto *para la población*
  - **Heterogéneos**, las personas reaccionan *diferente* ante x y ante z. El efecto que recuperamos es el efecto para un subconjunto de la población de interés

De vuelta al framework básico

- El **resultado que observamos**, y, es uno de los 2 resultados potenciales, y_0, y_1. Qué observamos depende del estatus de tratamiento, d:

$$y = \begin{cases} y_0 & \text{si } d = 0 \\ y_1 & \text{si } d = 1 \end{cases} \Leftrightarrow y = dy_1 + (1-d)\,y_0 = y_0 + (y_1 - y_0)\,d$$

- La diferencia de medias entre los tratados y no tratados es igual a

$$\underbrace{E(y|d=1) - E(y|d=0)}_{\text{diferencia observada}} = E(y_1|d=1) - E(y_0|d=0) = \underbrace{E(y_0|d=1) - E(y_0|d=0)}_{\text{sesgo}} + \underbrace{E(y_1 - y_0|d=1)}_{\text{ATET}}$$

  i.e. está contaminada por la diferencia en el promedio de resultados previos entre quienes *elijen* tratarse y quienes *elijen* no tratarse $E(y_0|d=1) - E(y_0|d=0)$

- Un RCT, al hacer la asignación al tratamiento **aleatoria**, hace que no exista relación entre el estado inicial y_0 y la decisión de tratarse. Entonces $\text{sesgo} \equiv E(y_0|d=1) - E(y_0|d=0) = 0$ y por tanto

$$\underbrace{E(y|d=1) - E(y|d=0)}_{\text{diferencia observada}} = 0 + \underbrace{E(y_1 - y_0|d=1)}_{\text{ATET}} = \text{ATET}$$

  i.e. obtenemos el efecto causal de interés con una diferencia de medias

Un supuesto crucial

- Esto nos dice que un RCT soluciona nuestros problemas para hallar causalidad: basta estimar una diferencia de medias o estimar β_d por MCO

$$y = \beta_0 + \beta_d d + \varepsilon$$

  esto nos dará estimadores consistentes pues la aleatorización hace que $Cov(d, \varepsilon) = 0$
- Nótese que esto asume que quienes fueron seleccionados para tomar el tratamiento *efectivamente* toman el tratamiento. No hay distinción entre *asignación* y toma de tratamiento
- Pero ¿qué pasaría si, a pesar de haber asignado aleatoriamente el tratamiento las personas *elijen* si tomarlo o no? En ese caso,
  - El **tratamiento efectivo**, d, está relacionado con factores no observados. Es decir, a pesar de la aleatorización, $Cov(d, \varepsilon) \neq 0$
  - Entonces, incluso bajo RCT, el *noncompliance* hace que el estimador MCO de TE, $\hat{\beta}_{d,MCO}$, sea inconsistente

El tratamiento potencial

- Ahora tenemos una nueva variable potencial: el **tratamiento potencial**

$$d = \begin{cases} d_0 & \text{si } z = 0 \\ d_1 & \text{si } z = 1 \end{cases} \Leftrightarrow d = zd_{1i} + (1-z)\, d_{0i} = d_{0i} + (d_{1i} - d_{0i})\, z$$

  El **tratamiento que observamos**, d, depende de la **asignación al tratamiento**, z, que **asumimos es aleatoria**
- Entonces ahora y depende de *2* variables: d y z

$$y_i\,(d_i = a) \rightarrow y_i\,(d_i = a, z_i = b) \equiv y_i\,(a,b)\,;a,b \in \{0,1\}$$

- Podemos representar esto como un sistema de 2 ecuaciones

$$\begin{array}{lcl} y_i = \mathbf{x}_i^{'}\beta + \varepsilon_i = \beta_0 + \beta_d d_i + \varepsilon_i & \underbrace{\rightarrow} & y_i = \beta_{0\boxed{i}} + \beta_{d\boxed{i}} d_i + \varepsilon_i \\ x_{Ki} = \mathbf{z}_i^{'}\theta + r_i = \theta_0 + \theta_1 z_i + r_i & \text{para el caso heterog.} & d_i = \pi_{0\boxed{i}} + \pi_{1\boxed{i}} z_i + r_i \end{array}$$

  Nótese la heterogeneidad en
  - el efecto tratamiento en el *resultado* y, $\beta_{di} = (y_{1i} - y_{0i})$
  - el efecto del instrumento en la *recepción del tratamiento* d, $\pi_{1i} = (d_{1i} - d_{0i})$
- Estamos interesados en el efecto promedio del tratamiento $E\,(\beta_{di})$ bajo este caso de **heterogeneidad**

El estimador de Wald

- El sistema anterior puede estimarse usando $\beta_{1,IV} = \frac{Cov(y,z_1)}{Cov(d,z_1)}$. Pero si x y z son binarios, la PRF se reduce a

$$\beta_{1,IV} = \frac{Cov(y,z_1)}{Cov(d,z_1)} \rightarrow \beta_{d,\text{Wald}} = \frac{E(y|z=1) - E(y|z=0)}{E(d|z=1) - E(d|z=0)} = \frac{\frac{dy}{dz}}{\frac{dd}{dz}}$$

  Esta es la **expresión poblacional del estimador de Wald**

- Pero $\beta_{d,\text{Wald}}$ no es directamente observable. Nuevamente, usemos nuestra muestra $\{(y_i, d_i, z_i) : i = 1, ..., N\}$ junto con el **principio de analogía**

$$\beta_{d,\text{Wald}} = \frac{E(y|z=1) - E(y|z=0)}{E(d|z=1) - E(d|z=0)} \underbrace{\rightarrow}_{\text{princip. analog.}} \boxed{\hat{\beta}_{d,\text{Wald}} = \frac{\widehat{E}(y|z=1) - \widehat{E}(y|z=0)}{\widehat{E}(d|z=1) - \widehat{E}(d|z=0)}}$$

  Esta expresión nos provee el **estimador muestral de Wald**, $\hat{\beta}_{d,\text{Wald}}$, que proviene de reemplazar
  - promedios condicionales poblacionales $E(.|z=1)$ que no observamos, por
  - promedios condicionales muestrales $\widehat{E}(.|z=1)$ que sí observamos

Los supuestos para el LATE

- En este caso de heterogeneidad, para que $\hat{\beta}_{d,\text{Wald}}$ estime algo relevante debemos combinar los supuestos en *la población* de
  1. potential-outcomes framework
  2. IV estimation

1. **SUTVA**: los resultados *potenciales* para *i no están relacionados* con el estado del tratamiento de otras personas

$$\text{si } z_i = z_i^{'} \text{ entonces } d_i(z) = d_i\left(z^{'}\right)$$

   Una violación ocurre si el d con un z dado es afectado por el d de otros con el mismo z

2. **Independencia** ("asignación tan buena como aleatoria"): el IV es independiente de resultados potenciales y tratamiento potencial

$$(y_i(d_{1i}, z=1), y_i(d_{0i}, z=0), d_{1i}, d_{0i}) \perp z_i$$

Los supuestos para el LATE

3. **ER**: cualquier efecto de z sobre y debe ocurrir vía el efecto de z sobre d. I.e. $y_i(d_i, z_i)$ es una función de d_i

$$y_i(d_i = \mathrm{d}, z = 1) = y_i(d_i = \mathrm{d}, z = 0) \text{ para d} = 0, 1$$

4. **Relevancia**: z está correlacionada con la variable endógena d

$$E(d_{1i} - d_{0i}) \neq 0$$

   a diferencia de la independencia y ER, esta primera etapa es comprobable pues se basa en d y z, los cuales se observan

5. **Monotonicidad**. la variable instrumental opera en la misma dirección para todos los individuos

$$\pi_{1i} \geq 0 \forall i \text{ o } \pi_{1i} \leq 0 \forall i$$

   En otras palabras,

   - el instrumento puede no tener efecto en algunas personas
   - pero para aquellos que sí afecta, son afectados en la *misma dirección*

Los 4 tipos de individuos

- En este caso de d y z binario, existen 4 tipos de individuos

1. Un **complier** tomaría el tratamiento $(d = 1)$ si $z = 1$ y no lo tomaría $(d = 0)$ si $z = 0$
2. Un **defier** *no* tomaría el tratamiento $(d = 0)$ si $z = 1$ y sí lo tomaría $(d = 1)$ si $z = 0$
3. Un **always-taker** siempre tomaría el tratamiento $(d = 1)$ sea que $z = 0$ o $z = 1$
4. Un **never-taker** nunca tomaría el tratamiento $(d = 0)$ sea que $z = 0$ o $z = 1$

	$z = 0$	$z = 1$
$d = 0$	compliers	defiers
	never takers	never takers
$d = 1$	defiers	complier
	always taker	always taker

- Nunca sabemos quién es quién realmente, pues observamos z 1 vez. Para saber con certeza si el individuo es C, D, A o N deberíamos: volver en el tiempo, modificar z (imposible!) y ver qué haría.
- Veamos qué implica esto en $\beta_{d,W} = \frac{E(y|z=1)-E(y|z=0)}{E(d|z=1)-E(d|z=0)}$

El numerador del estimador de Wald

- Dados los 4 tipos de individuos y por la ley de expec. iteradas

$$E(y|z=1)-E(y|z=0)=Pr(C)[E(y|z=1,C)-E(y|z=0,C)]+Pr(D)[E(y|z=1,D)-E(y|z=0,D)]$$
$$+Pr(A)[E(y|z=1,A)-E(y|z=0,A)]+Pr(N)[E(y|z=1,N)-E(y|z=0,N)]$$

pero por la ER, los resultados potenciales son indep. de z. Entonces

- Para los always takers

$$E(y|A,z=1)-E(y|A,z=0)=E(y_1|d=1,z=1)-E(y_1|d=1,z=0)\Leftrightarrow$$
$$E(y_1|d=1)-E(y_1|d=1)=0$$

- Para los never takers

$$E(y|N,z=1)-E(y|N,z=0)=E(y_0|d=0,z=1)=E(y_0|d=0,z=0)\Leftrightarrow$$
$$E(y_0|d=0)-E(y_0|d=0)=0$$

Entonces

$$E(y|z=1)-E(y|z=0)=Pr(C)[E(y|z=1,C)-E(y|z=0,C)]$$
$$+Pr(D)[E(y|z=1,D)-E(y|z=0,D)]$$

Pero debemos asumir que *no* hay defiers, por lo que $Pr(D)=0$,

$$E(y|z=1)-E(y|z=0)=Pr(C)[E(y|z=1,C)-E(y|z=0,C)]$$

Por tanto el numerador mide el efecto sólo para compliers

El denominador del estimador de Wald

- El denominador $E(d|z=1) - E(d|z=0)$ es el cambio en el estado de tratamiento d cuando z pasa de 0 a 1
- Dado que no hay defiers

	$z=0$	$z=1$
$d=0$	compliers	**defiers**
	never takers	never takers
$d=1$	**defiers**	complier
	always taker	always taker

$\rightarrow$

	$z=0$	$z=1$
$d=0$	*compliers*	
	never takers	never takers
$d=1$		*complier*
	always taker	always taker

- Al pasar z de $z=0$ a $z=1$
  - Los never takers no cambian su d
  - Los always takers no cambian su d
  - Sólo compliers cambian su d

  Para este d binario, el denominador es igual a la probabilidad de ser un complier

$$E(d|z=1) - E(d|z=0) = Pr(C)$$

El estimador LATE

- Introduciendo estas expresiones en la fórmula original

$$\beta_{d,\text{Wald}} = \frac{E(y|z=1) - E(y|z=0)}{E(d|z=1) - E(d|z=0)} = \frac{Pr(C)\left[E(y|z=1,C) - E(y|z=0,C)\right]}{Pr(C)} = E(y|z=1,C) - E(y|z=0,C)$$

- El estimador de $\beta_{1,\text{W}}$, $\hat{\beta}_{1,\text{W}}$, estima el valor poblacional *promedio* del efecto causal de d sobre y individual, $E(\beta_{1i})$
- ¿Pero de qué población? $\hat{\beta}_{1,\text{W}}$ es el efecto causal sólo para los compliers, i.e. aquellos quienes son inducidos al tratamiento por el instrumento
  - $\hat{\beta}_{d,\text{W}}$ se conoce como el estimado del efecto tratamiento promedio **local** (LATE)
  - La palabra **'local**' enfatiza que $\hat{\beta}_{1,\text{W}}$ estima un promedio *ponderado* sólo para aquellos individuos cuya probabilidad de tratamiento es influenciada por z

El LATE vs ATE bajo selecc. en obs.

- Esto contrasta *inmensamente* con el caso donde d es endógeno pero se puede justificar la CMI

$$E(y_0|\mathbf{x}, d = 1) = E(y_0|\mathbf{x}) \text{ y } E(y_1|\mathbf{x}, d = 1) = E(y_1|\mathbf{x})$$

  en cuyo caso **la selección en observables** ocurre
  - Si el efecto tratamiento es homogéneo

$$y_i = \beta_0 + \beta_d d_i + \varepsilon_i$$

    existe un único efecto causal. $\hat{\beta}_{d,MCO}$ estima el ATE: el valor esperado del efecto causal de d sobre y, $E(\beta_d)$, para *toda* la población
  - Si el efecto tratamiento es *heterogéneo*

$$y_i = \beta_{0\boxed{i}} + \beta_{d\boxed{i}} d_i + \varepsilon_i$$

    no existe un único efecto causal. Pero igual $\hat{\beta}_{d,MCO}$ estima el ATE: *promedio* de los efectos causales individuales de d sobre y, $E(\beta_{di})$, para *toda* la población
- El comando `teffects` de Stata estima $\hat{\beta}_{d,MCO}$ bajo efectos heterogéneos, lo cual es más general y por tanto más creíble (y colapsa en $E(\beta_d)$ si existen efectos homog.)

El LATE vs ATE bajo selecc. en no obs.

- Esto también contrasta con el caso donde d es endógeno y la CMI *no* ocurre porque **la selección depende de no observables**
- Si los efectos *son homogéneos*

$$y_i = \beta_0 + \beta_d d_i + \varepsilon_i$$
$$d = \pi_0 + \pi_1 z_i + r_i$$

existe un único efecto causal. $\hat{\beta}_{d,W}$ estima el ATE: el valor esperado de d sobre y, $E(\beta_d)$, para *toda* la población. Este es el el *típico* estimador de IV, visto en la sección 3 del módulo.

- $\hat{\beta}_{d,W}$ estima el ATE sólo si asumimos *algún* tipo de homogeneidad

	$z \to d$ homog. $(\pi_{1i} = \pi_1)$	$z \to d$ heterog. (π_{1i})
$d \to y$ homog. $(\beta_{1i} = \beta_1)$	**ATE**	
$d \to y$ heterog. (β_{1i})		LATE

¿Sirve de algo el LATE?

- Entonces, en presencia de efectos heterogéneos, el estimador IV de Wald
  - *no* es un estimador consistente del ATE
  - sí es un estimador consistente del ATE local: *no* nos dice nada sobre los efectos en los never-takers ni always-takers; para estos el instrumento z *no* afecta su d
- La implicancia de esto es muy fuerte:
  - En la mayoría de aplicaciones estaremos más interesados en estimar el ATE en toda la población
  - Pero bajo heterogeneidad, no podemos estimar el ATE sino sólo un ATE limitado
- LATE no da el ATET, porque el grupo tratado incluye compliers y *always takers*

Ejemplo del LATE

- Como ejemplo, suponga que tenemos un programa de capacitación laboral
  - Se les asigna *aleatoriamente* un número de prioridad z que influye en la probabilidad de que sean *admitidos* en el programa.
  - **No existe compliance obligatoria del tratamiento**: los trabajadores son admitidos pero deciden tomarlo $(d = 1)$ o no $(d = 0)$
- Aquellos seleccionados y que *toman* el tratamiento $z = 1, d = 1$ pueden diferir en términos de hab. no observadas (ε) de aquellos seleccionados y que no lo toman $z = 1, d = 0$
  - Esto genera endogeneidad, pues d está relacionado con ε y por tnato depende de factores no observados
  - A pesar de que z es aleatorio, la decisión de *tratamiento efectivo* es una decisión individual
- Existen 2 tipos de trabajadores
  1. Tipo 1: la 1era mitad, que sabe que se beneficiarán del programa
  $$\pi_{1i} = \pi_1^+ > 0; \beta_{1i} = \beta_1^+ > 0$$
  2. Tipo 2: la 2da mitad, que sabe que el programa *no* es efectivo para ellos, por lo que no tomarán el tratamiento aunque fueran admitidos, es decir
  $$\pi_{1i} = 0; \beta_{1i} = 0$$

Ejemplo del LATE

- Los individuos deciden enrolarse basado en su conocimiento de cuán efectivo será el programa
- Entonces
  1. el ATE es

     $$E(\beta_{1i}) = Pr(\texttt{tipo I})\,E(\beta_{1i}|\texttt{tipo I}) + Pr(\texttt{tipo II})\,E(\beta_{1i}|\texttt{tipo II}) = \frac{1}{2}\times\beta_1^{+} + \frac{1}{2}\times 0 = \frac{1}{2}\beta_1^{+}$$

     dado que en este ejemplo $Pr(\text{Tipo I}) = Pr(\text{Tipo II})$, el ATE da igual pesos a todos los individuos sin importar si se enrolan o no
  2. el LATE is

     $$E(\beta_{1i}|\texttt{complier}) = E(\beta_{1i}|\texttt{tipo I}) = \beta_1^{+}$$

     i.e. da el efecto causal para los trabajadores que tomen el programa. No incluye aquellos que *no* son compliers, i.e. aquellos que no toman el tratamiento a pesar de ser seleccionados
- En este ejemplo el LATE es el ATE para la 1era mitad que son los compliers: aquellos para quienes el instrumento influencia el hecho de tomar el tratamiento

Extensiones

- Hemos visto las propiedades de IV con ET heterog. usando
  1. d y z dummies
  2. no controles para y

  Recuérdese que LATE no ocurre porque z y d son dummies sino por heterogeneidad
- Podemos generalizar esto al caso con
  1. d y z continua
  2. controles para y
- ¿Cuándo debemos incluir controles? Si z se asigna aleatoriamente
  - no necesitamos controles para obtener un estimador consistente
  - pero agregar controles puede mejorar la eficiencia del estimador
- Si 2 investigadores tiene cada uno un z diferente
  - Los correspondientes $\hat{\beta}_{1,W}$ proveen diferentes estimadores del efecto causal
  - Cada z define un grupo diferente de compliers, y por tanto identifica un ATE diferente

Outline

La lotería de Vietnam

- Angrist (1990) examina si servir en la guerra de Vietnam tiene un efecto positivo sobre los ingresos laborales (asumiendo heterog.)

$$\text{ingresos}_i = \beta_{0\boxed{i}} + \beta_{1\boxed{i}}\text{vietnam}_i + \varepsilon_i$$

  evidentemente $Cov(\text{vietnam}_i, \varepsilon_i) \neq 0$ (¿Por qué?)
  - Si el hecho de servir en Vietnam está relacionado sólo con factores observados, entonces podemos usar métodos basados en selección en observables
  - Si existen factores no observados que determinan el tratamiento, necesitamos un instrumento para vietnam
- Durante la Guerra hubo 5 loterías para reclutar jóvenes. En cada lotería
  - A cada persona elegible se le dieron números (sin reemplazo) del 1 al 365 basado en su día de nacimiento (z)
  - Comenzaron llamando desde el número 1 hasta cubrir una cuota requerida por el Departamento de Defensa (e.g. 125)
  - A estos que fueron seleccionados se les hacían exámenes médicos. Se quedaba con aquellos quienes irán a la guerra ($\text{vietnam} = 1$)
- Debido a que aquellos con bajo z eran más probables de que $\text{vietnam} = 1$, Angrist (1990) decide construir una z_i dummy: 1 si tiene un número de lotería bajo (y es más probable que sea reclutado) y 0 si es alto

Analizando el instrumento

2. **Independencia** $\left(y_i\left(d_{1i}, z=1\right), y_i\left(d_{0i}, z=0\right), d_{1i}, d_{0i}\right) \perp z_i$
   - la asignación del número es una lotería, y por tanto es "tan bueno como aleatorio". Entonces z es independiente de
     - ganancias potenciales, $(\text{ingresos}_0, \text{ingresos}_1)$
     - servicio militar potencial, $(\text{vietnam}_0, \text{vietnam}_1)$
3. **Exclusion restriction**: $y_i\left(d_i = \mathrm{d}, z=1\right) = y_i\left(d_i = \mathrm{d}, z=0\right)$ para $\mathrm{d} = 0, 1$
   - z determina si vietnam=1 (relevancia)
   - Pero también z sólo debe afectar a ingresos mediante vietnam y no a través de w. E.g.

$$\begin{array}{ccccccc} & & \boldsymbol{w} & \rightarrow & \rightarrow & \rightarrow & \\ & \nearrow & & & & & \downarrow \\ z & \rightarrow & \text{vietnam} & \rightarrow & \rightarrow & \rightarrow & \text{ingresos} \\ & & & \nwarrow & & \nearrow & \\ & & & & \boldsymbol{\varepsilon} & & \end{array}$$

   Esto se violaría z afecta la educación de quienes evitan el draft. Si este fuera el caso, entonces z afecta ingresos indirectamente a través de 2 vías
     - a través del efecto de z en vietnam
     - a través del efecto de z en educación $= \boldsymbol{w}$
   - En general, un número de lotería aleatorio no implica que se cumpla la restricción de exclusión. Estas son suposiciones diferentes.

Analizando el instrumento

4. **Relevancia**: $E(d_{1i} - d_{0i}) \neq 0$
   - ¿Un número lotería $(z = 0)$ aumenta la probabilidad promedio de servicio militar? Si sí, entonces z cumple con el requisito de la primera etapa.
   - A diferencia de independencia y la ER, esta condición es comprobable mediante

$$\text{vietnam}_i = \pi_{0i} + \pi_{1i} z + r_i$$

5. **Monotonicidad**: $\pi_{1i} \geq 0, \forall i$ o $\pi_{1i} \leq 0, \forall i$
   - La elegibilidad para el reclutamiento
     - puede no tener efecto sobre la probabilidad de servicio militar para algunos
     - cuando tiene un efecto, los desplaza a todos en servicio

- Por lo tanto, ser elegible para el draft sirve como una variable instrumental válida que cumple además con los supuestos del LATE
- Nuestro sistema de ecuaciones que identifica el LATE está dado por

$$\text{ingresos}_i = \beta_{0i} + \beta_{1i}\text{vietnam} + \varepsilon_i$$
$$\text{vietnam}_i = \pi_{0i} + \pi_{1i}\text{lotería} + r_i$$

Los resultados

- El resultado más importante es que para veteranos blancos, el servir en Vietnam *reduce* el salario en alrededor de US$2,050 - US$2,741

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean (1)	Eligibility Effect (2)	Mean (3)	Eligibility Effect (4)	(5)
1981	16,461	-435.8 (210.5)	0.267	0.159 (0.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (293)
1969	2,299	-2.0 (34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

Los resultados

- En base a estos números podemos hallar el LATE

Earnings year	Earnings: Mean (1)	Earnings: Eligibility Effect (2)	Veteran Status: Mean (3)	Veteran Status: Eligibility Effect (4)	Wald Estimate of Veteran Effect (5)
1981	16,461	-435.8 (210.5)	0.267	0.159 (0.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (293)
1969	2,299	-2.0 (34.5)			

Sabemos que $\hat{\beta}_{d,\mathrm{W}} = \frac{\widehat{E}(y|z=1)-\widehat{E}(y|z=0)}{\widehat{E}(d|z=1)-\widehat{E}(d|z=0)}$

- $\widehat{E}(y|z=1)-\widehat{E}(y|z=0)$ es la diferencia de salarios promedio entre aquellos con número de lotería bajo y alto
- $\widehat{E}(x|z=1)-\widehat{E}(x|z=0)$ es la diferencia de participación entre aquellos con número de lotería bajo y alto

Los resultados

- En base a estos números podemos hallar el LATE

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Para la generación

- de 1981 tenemos que

$$\hat{\beta}^{1981}_{d,\mathrm{W}} = \frac{\widehat{E}(y|z=1) - \widehat{E}(y|z=0)}{\widehat{E}(d|z=1) - \widehat{E}(d|z=0)} = \frac{-435{,}8}{0{,}159} = -2741$$

- de 1971 tenemos que

$$\hat{\beta}^{1971}_{d,\mathrm{W}} = \frac{\widehat{E}(y|z=1) - \widehat{E}(y|z=0)}{\widehat{E}(d|z=1) - \widehat{E}(d|z=0)} = \frac{-325{,}9}{0{,}159} = -2050$$

La interpretación de *este* LATE

- El efecto encontrado es para la *subpoblación* inducida por la lotería a tomar el tratamiento, los *complier*s, no para toda la población
- IV estima el efecto promedio del servicio militar en los ingresos de
  - las subpoblaciones que fueron *induc*idos en por la lotería a tomar o no el tratamiento
  - No identifica el efecto causal del servicio militar en
    - patriotas (always-takers): estos siempre servirían en Vietnam, $d = 1$, pues no les importa un número de lotería alto ($z = 0$) o bajo ($z = 1$). Siempre van a servir

$$d_{0i} = d_{1i} = 1$$

    - quienes se casaron o entraron a estudiar para no ir a Vietnam (never-takers): estos *no* servirían en Vietnam, $d = 0$, independientemente de si su número de lotería es alto ($z = 0$) o bajo ($z = 1$) . Nunca van a servir

$$d_{0i} = d_{1i} = 0$$

- No identificamos el ATE sino el LATE. El efecto para la subpoblación que obtenemos depende *completamente* del instrumento usado

Outline

Complicaciones adicionales del estimador IV

- Cuando $Cov(x_1, \varepsilon) \neq 0$, i.e. cuando hay endogeneidad, el estimador MCO es inconsistente. En este caso es necesario usar un instrumento z
- Pero resulta que el estimador de IV *nunca es insesgado*, a diferencia del estimador MCO que sí puede ser insesgado si $E(\varepsilon|x_1) = 0$
  - De hecho, en el caso de un modelo exactamente identificado, $\hat{\beta}_{1,IV}$ *no tiene valor esperado* (Kinal 1980)
  - Esto significa que, *en pequeñas muestras, el estimador IV puede tener un sesgo sustancial*. Por tanto *es necesario usar muestras grandes* para justificar el uso del estimador IV
- Entonces, uno podría pensar que si $Cov(x_1, \varepsilon) = 0$ y $Cov(x_1, z_1) \neq 0$, con una muestra grande, $N \to \infty$, el estimador IV es consistente para el efecto causal
- *No necesariamente*: depende de si los instrumentos son **débiles** o **no**

> Con instrumentos débiles los estimadores de variables instrumentales pueden funcionar mal ¡incluso con $N \to \infty$!

El problema de instrumentos débiles: inconsistencia

- Sigamos con el caso simple: $y = \beta_0 + \beta_1 x_1 + \varepsilon$ con 1 instrumento z_1. Sabemos que en este modelo simple $\hat{\beta}_{IV} = \frac{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)(y_i-\bar{y})}{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)(x_{1i}-\bar{x}_1)}$
- Si tomamos el plim de esta expresión y la trabajamos tenemos

$$\text{Inconsistencia} = \text{plim}\left(\hat{\beta}_{1,IV}\right) - \beta_1 = \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right)\frac{Corr\left(z_1,\varepsilon\right)}{Corr\left(z_1,x_1\right)}$$

  Esta expresión nos dice que
  - El estimador $\hat{\beta}_{IV}$ es consistente si $Corr\left(z_1,\varepsilon\right) = 0$
  - Pero en la vida real *rara vez la condición de exclusión se cumple totalmente*, i.*e. que definitivamente* $Corr\left(z_1,\varepsilon\right) = 0$ sino más bien que $Corr\left(z_1,\varepsilon\right) \approx 0$
- No obstante, la magnitud de la inconsistencia, $\left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right)\frac{Corr(z_1,\varepsilon)}{Corr(z_1,x_1)}$
  - *se vuelve arbitrari*amente grande a medida que $Corr\left(z_1,x_1\right) \to 0$
  - *se reduce* a medida que $Corr\left(z_1,x_1\right) \to 1$

  Por tanto, la condición de relevancia $Corr\left(z_1,x_1\right) \neq 0$, no es suficiente: con $Cov\left(z_1,\varepsilon\right) \approx 0$ e **instrumentos débiles** la inconsistencia del estimador $\hat{\beta}_{IV}$ puede ser muy grande

▸ Derivación

El estimador MCO vs el estimador IV, round 2

- Nótese que
  - Ya sabemos que si $Corr(x, \varepsilon) \neq 0$ el estimador MCO es inconsistente
  - Acabamos de ver que si $Corr(z_1, \varepsilon) \approx 0$ el estimador IV *también* es inconsistente.

  Pero ¿cuál de los 2 es más inconsistente?
- Escribamos el plim del estimador MCO para el mismo modelo de regresión simple $y = \beta_0 + \beta_1 x_1 + \varepsilon$

$$\text{plim}\left(\hat{\beta}_{1,MCO}\right) = \beta_1 + \frac{Cov(x_1, \varepsilon)}{Var(x_1)} = \beta_1 + \frac{\frac{Cov(x_1,\varepsilon)}{\sigma_\varepsilon \sigma_{x_1}}}{\frac{\sigma_{x_1}\sigma_{x_1}}{\sigma_\varepsilon \sigma_{x_1}}} = \beta_1 + \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right) Corr(x_1, \varepsilon)$$

y entonces

$$\text{Inconsistencia} = \text{plim}\left(\hat{\beta}_{1,MCO}\right) - \beta_1 = \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right) Corr(x_1, \varepsilon)$$

El estimador MCO vs el estimador IV, round 2

- Comparando ambas expresiones,
  - $\text{Inconsistencia}_{IV} = \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right)\frac{Corr(z_1,\varepsilon)}{Corr(z_1,x_1)}$
  - $\text{Inconsistencia}_{MCO} = \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right)Corr\left(x_1,\varepsilon\right)$

  la inconsistencia en IV es mayor que la del estimador OLS si

$$\text{Inconsistencia}_{IV} > \text{Inconsistencia}_{OLS} \Leftrightarrow \underbrace{\frac{\left|Corr\left(z_1,\varepsilon\right)\right|}{\left|Corr\left(z_1,x_1\right)\right|}}_{\text{IV}} > \underbrace{\left|Corr\left(x_1,\varepsilon\right)\right|}_{\text{OLS}}$$

- Entonces si el instrumento débil, con $\left|Corr\left(z_1,x_1\right)\right|$ bajo, ¡el estimador IV puede ser más inconsistente que el estimador MCO!

¿Cómo saber si un instrumento débil?

- En el modelo sobreidentificado con 1 regresor endógeno, podemos calcular el estadístico F de que los coeficientes de los instrumentos $(z_1, ..., z_M)$ son igual a 0 en la regresión de primera etapa de 2SLS
- El **estadístico** F nos dice cuánta información (condicional a $\mathbf{x}$) contienen los instrumentos
  - Cuanto más información, mayor es el valor esperado del estadístico F
  - Como regla: *no* tenemos instrumentos débiles si el F de la primera etapa es ≥ 10 ▸ Derivación
- Stock y Yogo (2005) proporcionan un test para instrumentos débiles que no asume que sobreidentificación de nuestro modelo, donde «no débil» significa que el sesgo del 2SLS es como máximo 10 % del sesgo OLS

$$H_0 : \text{los instrumentos son débiles y } H_a : \text{los instrumentos son no débiles}$$

  - La prueba compara el F de la primera etapa con un **valor crítico** que depende del número de instrumentos. Para una prueba con significancia del 5 %, los valores críticos oscilan entre 9.08 y 11.52.
  - El problema es que el test de S&Y asume *homocedasticidad*

Ilustración: BJB (1995)

- Bound, Jaeger, y Baker (1995) analizan el setting de A&K (1991)

Independent variable	OLS	2SLS	OLS	2SLS
Years of schooling	0.063	0.142	0.063	0.081
	(0.000)	(0.033)	(0.000)	(0.016)
First stage *F*		13.5		4.8
Excluded instruments				
Quarter of birth		Yes		Yes
Quarter of birth × year of birth		No		Yes
Number of excluded instruments		3		30

Note: Standard errors in parentheses. First stage is quarter of birth dummies.

- Una vez que A&K incluyen las 30 dummies de trimestres de nacimiento × año, el estadístico F cae a 4.8
- Esto ocurre porque, como se vio en el gráfico en la sección anterior, la relación entre el trimestre de nacimiento y la escolaridad se hizo más pequeña para las últimas cohortes Gráfico

Ilustración: BJB (1995)

- Luego incluyen los instrumentos débiles: agregan los 150 instrumentos, dados por trimestres de nacimiento $\times$ estado

Independent variable	**OLS**	**2SLS**	**OLS**	**2SLS**
Years of schooling	0.063	0.083	0.063	0.081
	(0.000)	(0.009)	(0.000)	(0.011)
First stage *F*		2.4		1.9
Excluded instruments				
Quarter of birth		Yes		Yes
Quarter of birth $\times$ year of birth		Yes		Yes
Quarter of birth $\times$ state of birth		Yes		Yes
Number of excluded instruments		180		178

Note: Standard errors in parentheses.

- Y aquí vemos que el problema persiste: los instrumentos son débiles y, por lo tanto, el sesgo del estimador IV es cercano al sesgo de OLS
- Incluso peor, ¡este problema ocurre a pesar de que A&K usan 300,000 a 500,000 observaciones! Es decir, un número grande observaciones no soluciona el problema si los instrumentos son débiles
- Esto también sugiere que *no deberíamos usar muchos instrumentos*

La importancia de la primera etapa

- Un punto que vale enfatizar es que el uso de la regresión de forma la primera etapa

$$x_K = \mathbf{z}^{'}\boldsymbol{\theta} + r = \theta_0 + \underbrace{\theta_1 x_1 + \ldots + \theta_{K-1} x_{K-1}}_{\text{instrumentos incluidos}} + \underbrace{(\theta_1 z_1 + \ldots + \theta_M z_M)}_{\text{instrumentos excluidos}} + r$$

  No sólo requiere evaluar si el instrumento es relevante $\left(H_0 : \theta_1 = \theta_2 = \ldots = \theta_M = 0\right)$ y no débil $(F > 10)$. Es también importante tomar nota del signo y magnitud de los $\hat{\theta}_1, \ldots, \hat{\theta}_M$

- Esto es porque los argumentos de por qué un instrumento es válido debe incluir una discusión sobre la naturaleza de la relación entre x_K y z
  - E.g. en una ecuación que instrumenta el nivel educativo (x_K) por la cercanía a la escuela de la persona cuando era niño (z) se espera que $\theta_1 > 0$
  - Si en la muestra se encuentra que el efecto parcial es negativo, es decir, $\theta_1 < 0$, entonces es probable que z no sea un buen instrumento para x_K
- Esto *no tiene nada que ver* con que se cumpla la ortogonalidad $Cov(z, \varepsilon) = 0$

Ilustración: Angrist y Krueger (1991)

Outcome variable	Birth cohort	Quarter-of-birth effect I	II	III
Total schooling	1930–1939	−0.124	−0.86	−0.015
		(0.017)	(0.017)	(0.016)
	1940–1949	−0.085	−0.035	−0.017
		(0.012)	(0.012)	(0.011)
High school grad	1930–1939	−0.019	−0.020	−0.004
		(0.002)	(0.002)	(0.002)
	1940–1949	−0.015	−0.012	−0.002
		(0.001)	(0.001)	(0.001)
College grad	1930–1939	−0.005	0.003	0.002
		(0.002)	(0.002)	(0.002)
	1940–1949	−0.003	0.004	0.000
		(0.002)	(0.002)	(0.002)

Note: Standard errors in parentheses.

- Nótese cómo los coeficientes para el trimestre *I* y *II* son estadísticamente significativos, pero negativos, tal como se espera. Los coeficientes para el trimestre *III* no son estadísticamente significativos
- El instrumento de trimestre de nacimiento *solo debería afectar la finalización de la escuela secundaria*
  - Dado que no vincula a nadie más allá de la escuela secundaria, no debería afectar la cantidad de años más allá de las probabilidades de terminar la escuela secundaria o la universidad. Esta la 2da condición de la ER
  - Vemos que el efecto de estos trimestres es 0 para college graduates

¿Qué hacer si tenemos instrumentos débiles?

- Si tenemos un modelo sobre identificado con instrumentos fuertes y otros débiles, *debemos descartar los instrumentos más débiles* y usar el subconjunto más relevante
- Si tenemos un modelo sobre identificado pero *sin* suficientes instrumentos fuertes, hay 2 opciones
  1. Encontrar instrumentos adicionales más fuertes. ¡Más fácil decirlo que hacerlo!
  2. Emplear métodos distintos como **Limited information maximum likelihood estimator** (**LIML**). Este produce estimadores insesgados para la **mediana** con modelos sobreidentificados asumiendo homogeneidad y genera una reducción del sesgo de muestra finita.
- Si tenemos un modelo exactamente identificado, no se puede descartar instrumentos (por definición)

> The real solution for a weak instrument problem is get **better instruments** (...) So you should just continue searching for stronger instruments that simultaneously satisfy the exclusion restriction. Good luck with that. *Seriously, good.*
> *Cunningham, S. (2021) Causal Inference: The Mixtape*

El problema de instrumentos débiles: SEs incluso más altos

- Comparemos las varianzas de IV y OLS

$$\frac{AVar\left(\hat{\beta}_{1,IV}\right)}{AVar\left(\hat{\beta}_{1,OLS}\right)} = \frac{\frac{1}{N} \times \frac{\sigma_{\varepsilon}^2}{\sigma_x^2 \times \rho_{x,z}^2}}{\frac{1}{N} \times \frac{\sigma_{\varepsilon}^2}{\sigma_x^2}} = \frac{1}{\rho_{x,z}^2}$$

- Asumamos que z cumple con la condición de relevancia

$$Cov\left(x_1, z_1\right) \neq 0 \Leftrightarrow \rho_{x,z}^2 \neq 0$$

  Entonces la diferencia entre $AVar$s depende de la fuerza de la asociación entre x_1 y z_1, $\rho_{x,z}^2$
  - si la fuerza es alta $\rho_{x,z}^2 >> 0$, la diferencia entre las varianzas asintóticas será muy pequeña.
  - si la fuerza es baja $\rho_{x,z}^2 \approx 0$, la diferencia entre las varianzas asintóticas será muy grande
- Con **instrumentos débiles** $\rho_{x,z}^2 \approx 0$ la varianza del estimador IV puede ser mucho mayor que la del estimador OLS

Outline

¿Se puede testear la exogeneidad de los instrumentos?

- **En el caso exactamente identificado** no es posible probar H_0: instrumentos son exógenos
  - La única forma de evaluar si los instrumentos son exógenos es argumentando
- **En el caso sobreidentificado** *sí* es posible probar las **restricciones de sobreidentificación**, i.e. que los instrumentos *adicionales* son exógenos bajo el supuesto de *que hay suficientes instrumentos válidos*
  - E.g. si tenemos $K = 1$ con x_1 regresor endógeno, no controles, entonces $\mathbf{x} = (1; x_1)_{1\times(1+1)}$ y si hay 2 instrumentos excluidos ($M = 2$) entonces $\mathbf{z} = (1; z_1, z_2)_{1\times(1+0+2)}$.
  - Podríamos calcular 2 estimadores 2SLS: si $\mathbf{z}_1^{'} = (1, z_1)_{1\times(1+1)}$ obtenemos $\hat{\beta}_{z_1,2SLS}$ usando el 1er instrumento y si $\mathbf{z}_2^{'} = (1, z_2)_{1\times(1+1)}$ obtenemos $\hat{\beta}_{z_2,2SLS}$ usando el 2do instrumento
  - Dado esto, si z_1 y z_2 son exógenos, $\hat{\beta}_{z_1,2SLS}$ y $\hat{\beta}_{z_2,2SLS}$ serán similares y si z_1 o z_2 (o ambos) son *no* exógenos, $\hat{\beta}_{z_1,2SLS}$ y $\hat{\beta}_{z_2,2SLS}$ serán muy diferentes

¿Se puede testear la exogeneidad de los instrumentos?

- Asumiendo homoscedasticidad, podemos usar el **J test**. Este se basa en que la exogeneidad de instrumentos significa que no están correlacionados con $\hat{\varepsilon}_i$
  1. Estimar por 2SLS el modelo de interés
     $y = \mathbf{x}'\beta + \varepsilon = \beta_0 + \beta_1 x_1 + \ldots + \beta_{K-1} x_{K-1} + \beta_K x_K + \varepsilon$, donde x_K es endógena
  2. Generar residuos usando la x_K endógena *observada*, *no* predicha:
     $\hat{\varepsilon}_{i,2SLS} = y_i - \left(\hat{\beta}_{0,2SLS} + \hat{\beta}_{1,2SLS} x_{1i} + \hat{\beta}_{2,2SLS} x_{2i} + \ldots + \hat{\beta}_{K,2SLS} x_{Ki}\right)$
  3. Regresionar $\hat{\varepsilon}_{i,2SLS}$ contra el vector $\mathbf{z}$: $\hat{\varepsilon}_{i,2SLS}$ sobre $1, x_{1i}, \ldots, x_{K-1i}, z_{1i}, \ldots, z_{Mi}$
  4. Calcular el estadístico F de los coeficientes de los M instrumentos excluidos. El test se distribuye como χ^2 con $M-1$ DoF

     $$J = M \times F \sim \chi^2_{M-1}$$

     donde $M-1$ es el **grado de sobreidentificación**, el número de instrumentos menos el número de regresores endógenos
- El supuesto básico de este test es que al menos 1 de los instrumentos sea exógeno y no débil. Si tenemos 2 instrumentos mediocres, el test no ayuda
- Bajo heteroscedasticidad, el test se conoce como el **test de Sargan**

El test de Hausman-Durbin-Wu

- Ante endogeneidad, el estimador 2SLS produce estimadores consistentes. Pero es menos eficiente que el MCO, pues usa sólo una parte de la información en x: sólo aquella inducida por z, $z \to x$. Entonces $\hat{\beta}_{IV}$ será estimado de forma muy imprecisa
  - ¿Pero qué pasa si las diferencias entre $\hat{\beta}_{OLS}$ y $\hat{\beta}_{IV}$ son casi 0? Entonces la inconsistencia que produce x_K es poca, y podemos usar el estimador MCO
  - Pero si las diferencias entre $\hat{\beta}_{OLS}$ y $\hat{\beta}_{IV}$ son significativas, entonces el regresor endógeno generará problemas para el estimador MCO
- Si nuestro modelo es $y = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \varepsilon$, la forma reducida de x_K es

$$x_K = \theta_0 + \theta_1 x_1 + ... + \theta_{K-1} x_{K-1} + \theta_{z_1} z_1 + r$$

Entonces la endogeneidad de x_K equivale a

$$Cov(x_K, \varepsilon) = \theta_1 Cov(x_1, \varepsilon) + ... + \theta_{K-1} Cov(x_{K-1}, \varepsilon) + \theta_{z_1} Cov(z_1, \varepsilon) + Cov(\varepsilon, r)$$

Dado que $Cov(x_1, \varepsilon) = ... = Cov(x_{K-1}, \varepsilon) = Cov(z_1, \varepsilon) = 0$, entonces $Cov(x_K, \varepsilon) = Cov(\varepsilon, r)$ y por tanto no hay endogeneidad si

$$Cov(x_K, \varepsilon) = Cov(\varepsilon, r) = 0$$

El test de Hausman-Durbin-Wu

- Escribamos el error en la ecuación estructural como

$$\varepsilon = \eta r + e; Cov(r, e) = 0 \text{ y } E(e) = 0$$

  Entonces **no habrá** endogeneidad si

$$Cov(x_K, \varepsilon) = Cov(\varepsilon, r) = \eta = 0$$

- Dado lo anterior, podemos incluir $\eta r + e$ en la ecuación poblacional para ver si existe endogeneidad

$$y = \theta_0 + \theta_1 x_1 + ... + \theta_{K-1} x_{K-1} + \theta_{z_1} z_1 + (\eta r + e)$$

  No observamos r pero podemos estimarlo usando la forma reducida de x_K
- El test de Hausman-Durbin-Wu evalúa la endogeneidad testeando η
  1. Estimamos la forma reducida x_K regresionándolo contra todas las exógenas
  2. Obtenemos los residuos $\hat{r}$ de la forma reducida
  3. Estimamos por MCO la ecuación estructural introduciendo $\hat{r}$

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \eta \hat{r} + error$$

  4. Testeamos $H_0 : \eta = 0$, usando un t-test robusto a heterosk. *Si rechazamos concluimos que x_K es endógena*
- Esto muestra que la endogeneidad se origina por la relación entre no observables que explican y y no observables que explican x_K

Los métodos de IV en la práctica

- Las variables instrumentales son un diseño poderoso para identificar efectos causales cuando los datos sufren de selección en *no* observables.
  - Si uno encuentra un instrumento válido, ha logrado identificar un efecto causal puro
- Pero tremenda recompensa tiene muchas limitaciones serias
  1. La defensa del IV requiere teoría y no es fácil establecer argumentos que permitan convencer a la gente de que las condiciones de ER se cumplen
  2. Bajo heterogeneidad, $\hat{\beta}$ solo identifica el LATE. Esto puede que *no* sea un resultado relevante para policy si el ATE para los compliers es diferente al ATE para las otras subpoblaciones.
  3. IV requiere muchas condiciones de identificación y por lo tanto, muchos encuentran estimación IV menos creíble, *no* porque no logra identificar un efecto causal sino porque es difícil imaginar un instrumento que satisfaga todas las condiciones.

¿Vale la pena usar IV?

- Pero, nuevamente, recuérdese que la 1era condición de la ER no es suficiente. No sólo se requiere que $E(x|\mathbf{z}) \neq 0$ sino que la relación entre el instrumento z y x sea fuerte
  - Si nuestro F de la 1era etapa es menor 10 entonces el instrumento es débil
  - En este caso nuestro estimador es inconsistente y ni siquiera podemos hacer inferencia
- Por eso, a menudo debemos elegir entre
  1. Un estimador posiblemente inconsistente que tiene SEs relativamente pequeños (OLS)
  2. Un estimador consistente pero con un SE muy alto que no permite concluir nada interesante (IV, 2SLS).

  Un enfoque es usar OLS a menos que podamos rechazar la exogeneidad de las variables explicativas.

El PEOR error que podemos hacer es usar un instrumento que no es válido y decir que «tenemos un efecto causal»

¿Y qué hay de casos más generales?

- Hemos visto qué ocurre si tenemos *sólo* 1 regresor endógeno. Pero puede que tengamos 2 o más, con diferentes números de instrumentos
- E.g. tomemos el caso donde x_1 y x_2 son reg. endógenos (sin controles)
  - Si los instrumentos excluidos son (z_1, z_2), tenemos un modelo exact. identificado. Podemos usar el estimador $\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{'}\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$ reemplazando $\mathbf{x} = (x_1, x_2)$ por $\mathbf{z} = (z_1, z_2)$
  - Si los instrumentos excluidos son (z_1, z_2, z_3) tenemos un modelo sobreidentificado. Tenemos un modelo de 3 ecuaciones

$$
\begin{aligned}
&y = \beta_0 + \beta_1 x_1 + \varepsilon & \underbrace{\rightarrow}_{\text{con 2 regresores endógenos}} \quad & \text{outcome:} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\
&x_1 = \theta_0 + \theta_1 z_1 + r & & \text{primera etapa:} x_1 = \theta_{10} + \theta_{11} z_1 + \theta_{12} z_2 + \theta_{13} z_3 + r_1 \\
& & & \text{primera etapa:} x_2 = \theta_{20} + \theta_{21} z_1 + \theta_{22} z_2 + \theta_{23} z_3 + r_2
\end{aligned}
$$

podemos usar el estimador $\hat{\beta}_{2SLS} = \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \mathbf{x}_i\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i y_i$. Este estimador es igual al MCO de

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \varepsilon$$

donde $\hat{x}_1$ y $\hat{x}_2$ son las predicciones de sus 1eras etapas estimadas por MCO

Si las mecánicas son las mismas, ¿por qué no las hemos estudiado?

- La respuesta es obvia: si encontrar 1 instrumento es difícil, encontrar 2 es MÁS DIFÍCIL
  - Es por eso que la mayoría de papers que usan IV sólo consideran 1 endógena, que es la variable de interés
  - Las otras variables se *asumen controles*, y por tanto exógenas
- Siempre tenemos que ser escépticos de papers que usan modelos sobreidentificados
  - De hecho, mientras mayor sea el número de instrumentos, aun cuando estos pueden ser válidos, ¡mayor es el sesgo que obtenemos aun cuando nuestra muestra tenga miles de observaciones!
- Pero asumamos que en los ejemplos de la diapositiva anterior los instrumentos en $\mathbf{z}$ son válidos
  - Podemos usar el estimador IV o el de 2SLS
  - Podemos usar el estimador de GMM que es más eficiente que estos estimadores. Esto lo veremos en la práctica.

¡Gracias!


THIS IS WHERE I'D PUT
MY INSTRUMENTAL VARIABLE
IF I HAD ONE
imgflip.com

Derivación de la consistencia del estimador MCO

- La expresión para el estimador $\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})y_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$ puede trabajarse más

$$\hat{\beta}_{1,MCO} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(\beta_0 + \beta_1 x_i + \gamma a_i + \varepsilon_i)}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \beta_0 \frac{\sum_{i=1}^{N}(x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^{N}(x_i - \overline{x})(x_i)}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \gamma \frac{\sum_{i=1}^{N}(x_i - \overline{x})(a_i)}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \frac{\sum_{i=1}^{N}(x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

- Podemos reducir esta expresión usando algunas propiedades mecánicas
  - $\frac{\sum_{i=1}^{N}(x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \overline{x}\right)}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{(n\overline{x} - n\overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = 0$
  - $\frac{\sum_{i=1}^{N}(x_i - \overline{x})(x_i)}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = 1$
- Entonces

$$\begin{aligned}
\hat{\beta}_{1,MCO} &= \beta_0 \underbrace{\frac{\sum_{i=1}^{N}(x_i - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}}_{=0} + \beta_1 \underbrace{\frac{\sum_{i=1}^{N}(x_i - \overline{x})(x_i)}{\sum_{i=1}^{N}(x_i - \overline{x})^2}}_{=1} + \gamma \frac{\sum_{i=1}^{N}(x_i - \overline{x})a_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \frac{\sum_{i=1}^{N}(x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \\
&= \beta_1 + \gamma \frac{\sum_{i=1}^{N}(x_i - \overline{x})a_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2} + \frac{\sum_{i=1}^{N}(x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i - \overline{x})^2}
\end{aligned}$$

Derivación de la consistencia del estimador MCO

- Queremos ver si $\hat{\delta}_{MCO}$ es consistente, y por tanto usemos el operador que nos permite analizar esto

$$\text{plim}\left(\hat{\beta}_{1,MCO}\right) = \text{plim}\left(\beta_1 + \gamma\frac{\sum_{i=1}^{N}(x_i-\overline{x})\,a_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2} + \frac{\sum_{i=1}^{N}(x_i-\overline{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right) = \beta_1 + \gamma\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})\,a_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right) + \text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right)$$

  y para seguir operando debemos dividir por N el numerador y el denominador

$$\text{plim}\left(\hat{\beta}_{1,MCO}\right) = \beta_1 + \gamma\text{plim}\left(\frac{\frac{\sum_{i=1}^{N}(x_i-\overline{x})a_i}{N}}{\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}}\right) + \text{plim}\left(\frac{\frac{\sum_{i=1}^{N}(x_i-\overline{x})\varepsilon_i}{N}}{\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}}\right) = \beta_1 + \gamma\frac{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})a_i}{N}\right)}{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}\right)} + \frac{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})\varepsilon_i}{N}\right)}{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}\right)}$$

  expresado así, el numerador y el denominador son las covarianzas muestrales. Estas convergen a las covarianzas poblacionales, i.e.$\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})y_i}{N}\right) = Cov(x,y)$ por lo que

$$\begin{aligned}\text{plim}\left(\hat{\beta}_{1,MCO}\right) &= \beta_1 + \gamma\frac{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})a_i}{N}\right)}{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}\right)} + \frac{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})\varepsilon_i}{N}\right)}{\text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}\right)} \\ &= \beta_1 + \gamma\frac{Cov(x,a)}{Var(x)} + \frac{Cov(x,\varepsilon)}{Var(x)}\end{aligned}$$

Derivación de la consistencia del estimador MCO

- Hemos asumido que $E(x\varepsilon) = 0$ lo que implica que $Cov(x,\varepsilon) = 0$, por tanto

$$\text{plim}\left(\hat{\beta}_{1,MCO}\right) = \beta_1 + \gamma\frac{Cov(x,a)}{Var(x)} + \frac{\overbrace{=0}^{Cov(x,\varepsilon)}}{Var(x)} = \beta_1 + \gamma\frac{Cov(x,a)}{Var(x)}$$

- Por tanto

$$\text{plim}\left(\hat{\beta}_{1,MCO}\right) - \beta_1 = \gamma\frac{Cov(x,a)}{Var(x)}$$

y recordando que un estimador $\hat{\beta}_{1,MCO}$ es consistente si $\text{plim}\left(\hat{\beta}_{1,MCO}\right) - \beta_1 = 0$, el término de la derecha, $\gamma\frac{Cov(x,a)}{Var(x)}$, es la magnitud de la inconsistencia. Si $\hat{\beta}_{1,MCO}$ es consistente entonces

$$\gamma\frac{Cov(x,a)}{Var(x)} = 0 \Leftrightarrow \gamma = 0 \text{ y/o } \frac{Cov(x,a)}{Var(x)} = 0$$

◂ Back

Derivación de la inconsistencia del estimador IV

- Partimos del modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Leftrightarrow y_i - \overline{y} = \beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})$

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{N} (z_{1i} - \overline{z}_1)(y_i - \overline{y})}{\sum_{i=1}^{N} (z_{1i} - \overline{z}_1)(x_{1i} - \overline{x}_1)} = \frac{\sum_{i=1}^{N} (z_{1i} - \overline{z}_1)[\beta_1 (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})]}{\sum_{i=1}^{N} (z_i - \overline{z})(x_{1i} - \overline{x}_1)}$$

y operando

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{N} [\beta_1 (z_{1i} - \overline{z}_1)(x_i - \overline{x}) + (z_{1i} - \overline{z}_1)(\varepsilon_i - \overline{\varepsilon})]}{\sum_{i=1}^{N} (z_i - \overline{z})(x_{1i} - \overline{x}_1)} = \frac{\beta_1 \sum_{i=1}^{N} (z_{1i} - \overline{z}_1)(x_i - \overline{x}) + \sum_{i=1}^{N} (z_{1i} - \overline{z}_1)(\varepsilon_i}{\sum_{i=1}^{N} (z_i - \overline{z})(x_{1i} - \overline{x}_1)}$$

y dividiendo por N

$$\hat{\beta}_{IV} = \beta_1 + \frac{\frac{\sum_{i=1}^{N} (z_{1i} - \overline{z}_1)\varepsilon_i}{N}}{\frac{\sum_{i=1}^{N} (z_{1i} - \overline{z}_1)(x_{1i} - \overline{x}_1)}{N}}$$

Derivación de la inconsistencia del estimador IV

- Nuevamente tomando el plim de esta expresión tenemos que

$$\text{plim}\left(\hat{\beta}_{IV}\right) = \text{plim}\left(\beta_1 + \frac{\frac{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)\varepsilon_i}{N}}{\frac{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)(x_{1i}-\bar{x}_1)}{N}}\right) = \beta_1 + \frac{\text{plim}\left(\frac{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)\varepsilon_i}{N}\right)}{\text{plim}\left(\frac{\sum_{i=1}^{N}(z_{1i}-\bar{z}_1)(x_{1i}-\bar{x}_1)}{N}\right)} = \beta_1 + \frac{Cov\,(z_1,\varepsilon)}{Cov\,(z_1,x_1)}$$

  y por tanto

$$\text{plim}\left(\hat{\beta}_{IV}\right) - \beta_1 = \frac{Cov\,(z_1,\varepsilon)}{Cov\,(z_1,x_1)}$$

- Esta es la expresión de la inconsistencia. Podemos obtener una expresión alternativa: multiplicamos por $\left(\frac{1}{\sigma_\varepsilon \times \sigma_{x_1} \times \sigma_{z_1}}\right)$

$$\text{plim}\left(\hat{\beta}_{IV}\right) = \beta_1 + \frac{\left(\frac{1}{\sigma_\varepsilon \times \sigma_{x_1} \times \sigma_{z_1}}\right) Cov\,(z_1,\varepsilon)}{\left(\frac{1}{\sigma_\varepsilon \times \sigma_{x_1} \times \sigma_{z_1}}\right) Cov\,(z_1,x_1)} = \beta_1 + \frac{\frac{1}{\sigma_{x_1}} \times \frac{Cov(z_1,\varepsilon)}{\sigma_\varepsilon \times \sigma_{z_1}}}{\frac{1}{\sigma_\varepsilon} \times \frac{Cov(z_1,x_1)}{\sigma_{x_1} \times \sigma_{z_1}}} = \beta_1 + \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right) \frac{\boxed{Corr\,(z_1,\varepsilon)}}{\boxed{Corr\,(z_1,x_1)}}$$

  entonces la expresión alternativa para la consistencia es

$$\text{plim}\left(\hat{\beta}_{IV}\right) - \beta_1 = \frac{Cov\,(z,\varepsilon)}{Cov\,(z,x)} = \left(\frac{\sigma_\varepsilon}{\sigma_{x_1}}\right) \frac{Corr\,(z_1,\varepsilon)}{Corr\,(z_1,x_1)}$$

  y esperamos que sea igual a 0

Back

Derivación de la distribución asintótica del estimador IV

- Tomemos como punto de partida $\hat{\beta}_{IV} - \beta_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)\varepsilon_i}{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)(x_{1i}-\overline{x}_1)}$ y veamos cada parte el término de la derecha
- El numerador: si $N \to \infty$, $\overline{z}_1 - \mu_Z \approx 0$. Entonces en muestras grandes

$$\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)\varepsilon_i \approx \frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\mu_Z)\varepsilon_i$$

Para probar que asintóticamente está distribuida normalmente calculemos

1. Expectativa $E[(z_{1i}-\mu_Z)(,\varepsilon_i)] = Cov(z_{1i},\varepsilon_i)$ y porque el instrumento es exógeno $Cov(z_{1i},\varepsilon_i) = 0$

$$E\left(\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\mu_Z)\varepsilon_i\right) = \frac{1}{N}\sum_{i=1}^{N}E((z_{1i}-\mu_Z)\varepsilon_i) = 0$$

2. Varianza $Var\left(\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\mu_Z)\varepsilon_i\right)$ y porque las observaciones son i.i.d

$$Var\left(\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\mu_Z)\varepsilon_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}Var((z_{1i}-\mu_Z)\varepsilon_i) = \frac{1}{N^2}N\sigma_A^2 = \frac{1}{N}\sigma_A^2$$

Entonces $(z_{1i}-\mu_Z)\varepsilon_i$ satisface los requerimientos de TLC (si una secuencia $x_1,\ldots,x_N$ es i.i.d. y $0 < \sigma_x^2 < \infty$, entonces $\frac{\overline{x}-\mu_x}{\sigma_x} \overset{a}{\sim} N(0,1)$) por lo que

$$\frac{1}{N}\sum_{i=1}^{N}(Z_i-\mu_z)u_i \overset{a}{\sim} N\left(0,\frac{\sigma_A^2}{N}\right)$$

Derivación de la distribución asintótica del estimador IV

- El denominador $\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)(x_{1i}-\overline{x}_1)$ es la covarianza muestral que converje a la covarianza poblacional

$$\text{plim}\left[\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)(x_{1i}-\overline{x}_1)\right]=Cov(z_{1i},x_{1i})$$

  y porque el instrumento es relevante $Cov(z_{1i},x_{1i})\neq 0$
- Entonces

$$\hat{\beta}_{1,IV}-\beta_1=\frac{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)\varepsilon_i}{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)(x_{1i}-\overline{x}_1)}\approx\frac{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\mu_z)\varepsilon_i}{Cov(z_{1i},x_{1i})}$$

  y las condiciones para el teorema de Slutsky (a_n es constante y $S_n\overset{a}{\sim}S$ entonces $\frac{S_n}{a_n}\overset{a}{\sim}\frac{S}{a}$) ocurren y por tanto

$$\hat{\beta}_{1,IV}-\beta_1\approx\frac{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)\varepsilon_i}{\frac{1}{N}\sum_{i=1}^{N}(z_{1i}-\overline{z}_1)(x_{1i}-\overline{x}_1)}\overset{a}{\sim}\frac{1}{Cov(z_{1i},x_{1i})}N\left(0,\frac{\sigma_A^2}{N}\right)=N\left(0,\frac{1}{[Cov(z_{1i},x_{1i})]^2}\frac{\sigma_A^2}{N}\right)$$

  pero $\sigma_A^2=Var((z_{1i}-\overline{z}_1)\varepsilon_i)=E\left((z_{1i}-\overline{z}_1)^2\varepsilon_i^2\right)-[E((z_{1i}-\overline{z}_1)\varepsilon_i)]^2=$ $E(z_{1i}-\overline{z}_1)^2E\left(\varepsilon_i^2\right)=\sigma_z^2\sigma_\varepsilon^2$ y $\frac{Cov(z_{1i},x_{1i})}{\sigma_z\sigma_x}=\rho_{x,z}\Leftrightarrow[Cov(z_{1i},x_{1i})]^2=\rho_{x,z}^2\times\sigma_z^2\times\sigma_x^2$ y

$$\hat{\beta}_{1,IV}\overset{a}{\sim}N\left(\beta_1,\frac{1}{N}\frac{\sigma_\varepsilon^2}{\rho_{x,z}^2\times\sigma_x^2}\right)$$

Derivación de la regla de $F > 10$

- Denota el sesgo asintótico del estimador MCO como $\hat{\beta}_1^{OLS} - \beta_1$ y el sesgo del 2SLS es $E\left(\hat{\beta}_1^{2SLS}\right) - \beta_1$
- Es posible demostrar que cuando hay muchos instrumentos el sesgo del estimador 2SLS es aproximadamente

$$E\left(\hat{\beta}_1^{2SLS}\right) - \beta_1 = \frac{\hat{\beta}_1^{OLS} - \beta_1}{E(F) - 1}$$

donde $E(F)$ es la expectativa del estadístico F de primera etapa. Entonces, si $E(F) = 10$

$$E\left(\hat{\beta}_1^{2SLS}\right) - \beta_1 = \frac{\hat{\beta}_1^{OLS} - \beta_1}{9} \Rightarrow \frac{E\left(\hat{\beta}_1^{2SLS}\right) - \beta_1}{\hat{\beta}_1^{OLS} - \beta_1} = \frac{1}{9} \approx 10\,\%$$

- Por tanto, el sesgo de 2SLS, en relación con el sesgo de OLS, es aproximadamente 10 %. Esto es lo suficientemente pequeño como para ser aceptable en muchas aplicaciones.

◂ Back