Economía Aplicada

Cuasi experimentos: Variables Instrumentales y Fuentes de Endogeneidad

Departamento de Economía
Universidad Carlos III de Madrid

Evaluación de política con cuasi experimentos

- En un cuasi experimento o experimento natural existe alguna fuente de aleatoriedad que permite considerar a alguna variable como si fuese asignada de manera aleatoria. Consideramos dos tipos de cuasi experimentos:
- El caso de que el tratamiento (D) pueda considerarse como si fuese asignado de manera aleatoria (quizás condicionando en ciertas variables adicionales X).
- El caso de que una variable (Z) que afecta al tratamiento (D) se puede considerar como si fuese asignada de manera aleatoria (quizás condicionando en ciertas variables adicionales X). En ese caso Z puede utilizarse como variable instrumental para D.

  - El artículo de Angrist es un ejemplo de este caso.

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records, Angrist, AER(1990)

- ¿Hacer el servicio militar tiene un efecto negativo en los ingresos?
- Que los veteranos tengan menores ingresos que los no veteranos no implica que ser veterano sea la causa de los menores ingresos.
- Esta comparación simple de los ingresos entre los dos grupos no es adecuada si ser veterano no es independiente de los ingresos potenciales.
- Comparar los ingresos controlando por un conjunto de características observables solamente tiene sentido si ser veterano es independiente de los ingresos potenciales una vez que tenemos en cuenta esas características.

Efecto del estatus de veterano en los ingresos

Llamemos Y_i a los ingresos, D_i representa el estado de veterano en la era de Vietnam, y X_i un conjunto de controles:

¿Es correcto estimar la siguiente esperanza condicional para estimar el efecto de ser veterano de Vietnam en los ingresos?

$$E[Y_i|D_i, X_i] = \beta_0 + \alpha D_i + \gamma' X_i$$

- Probablemente existan diferencias no observables entre los hombres que eligen enrolarse en el servicio militar y los que no, y probablemente esas diferencias estén correlacionadas con los ingresos potenciales.
- Si D_i está correlacionada con variables no observadas que pertenecen a la ecuación, las estimaciones de MCO serán inconsistentes.
- Una posible solución es encontrar una variable instrumental válida.

Una VI para el estatus de veterano

- Preocupaciones sobre la justicia de la política de conscripción en EEUU llevó a instituir una selección por sorteo en 1970.
- Este sorteo se realizó anualmente en 1970, 1971 y 1972. Se le asignó un número aleatorio (del 1 al 365) a cada fecha de nacimiento de las cohortes de 19 años, los hombres con números debajo de un cierto valor (fijado por el Ministerio de Defensa) eran llamados a servir en el ejército.
- El estatus de veterano no está completamente determinado por el sorteo: algunos se alistaban voluntariamente, otros evitaban el enrolamiento por razones de salud o de estudios. Pero el sorteo está altamente correlacionado con el estatus de veterano.

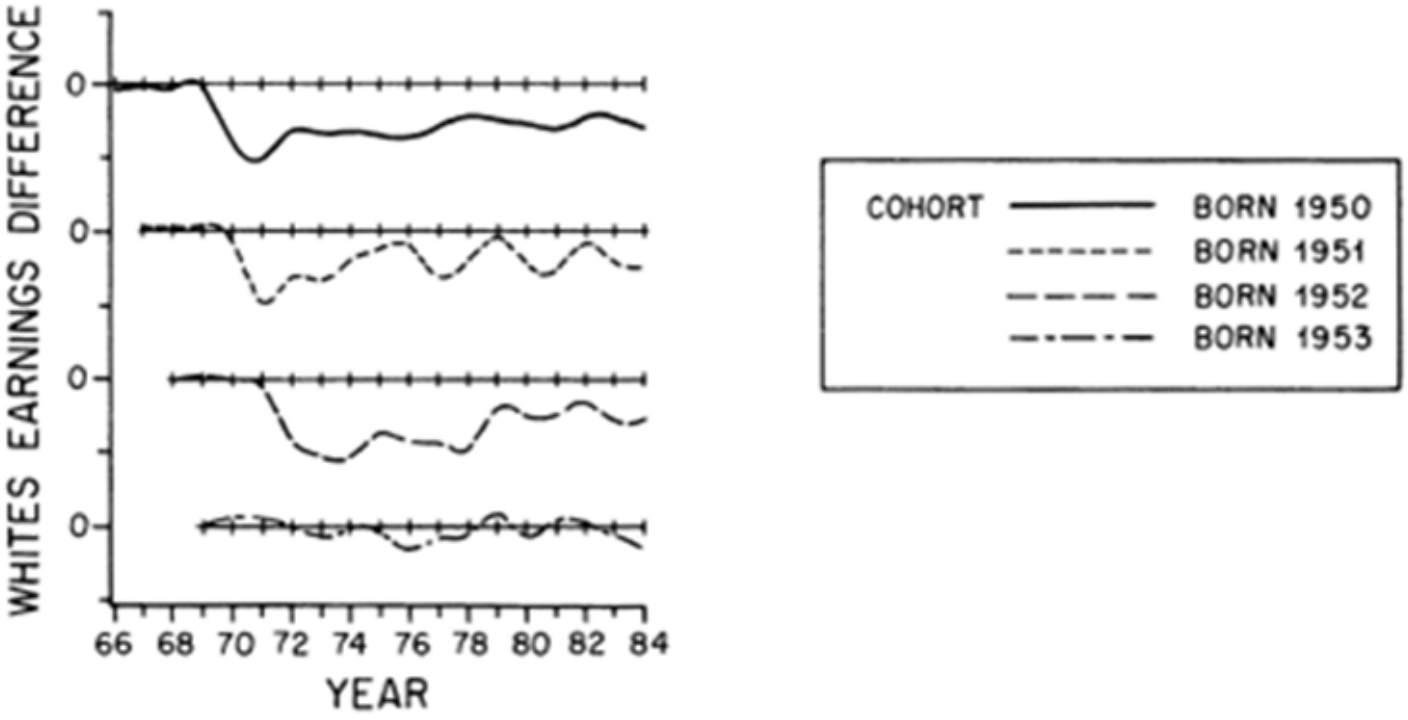
Reclutamiento Aleatorio como instrumento 1/2

- Llamemos Z_i a la variable binaria de reclutamiento aleatorio.
- De modo de poder identificar el efecto causal de D_i en los ingresos es crucial que la única razón para que $E(Y_i|Z_i)$ cambie con Z_i sea la variación en $E(D_i|Z_i)$: o sea que el reclutamiento aleatorio afecte a los ingresos sólo a través de su efecto en el estatus de veterano.
- Un chequeo simple es ver si existe asociación entre Z_i y alguna característica personal que no puede ser afectada por D_i. Otro chequeo es mirar si existe asociación entre Z_i y la variable de interés en muestras donde no hay relación entre D_i y Z_i.

Reclutamiento Aleatorio como instrumento 2/2

- Angrist analiza los ingresos de 1969 (previos al sorteo de 1970) y encuentra que no hay efecto del reclutamiento aleatorio en los ingresos.
- También analiza la cohorte de hombres nacidos en 1953. Aunque hubo sorteo para esa cohorte, nadie fue finalmente seleccionado. Por lo tanto no hay correlación entre el número obtenido en el sorteo y el estatus de veterano para esa cohorte. Angrist encuentra que no existe una relación significativa entre los ingresos y el reclutamiento aleatorio para los nacidos en 1953.
- Ambos resultados apoyan el argumento de que el reclutamiento aleatorio afecta a los ingresos sólo a través del estatus de veterano.

Diferencias en Ingresos por Reclutamiento Aleatorio - Un Gráfico



Notes: The figure plots the difference in FICA taxable earnings by draft-eligibility status for the four cohorts born 1950–53. Each tick on the vertical axis represents $500 real (1978) dollars.

Diferencias en Ingresos por Reclutamiento Aleatorio - Regresiones

TABLE 1—DRAFT-ELIGIBILITY TREATMENT EFFECTS FOR EARNINGS

	Whites							
	FICA Taxable Earnings				Total W-2 Compensation			
Year	1950	1951	1952	1953	1950	1951	1052	1953
66	−21.8							
	(14.9)							
67	−8.0	13.1						
	(18.2)	(16.4)						
68	−14.9	12.3	−8.9					
	(24.2)	(19.5)	(19.2)					
69	−2.0	18.7	11.4	−4.0				
	(34.5)	(26.4)	(22.7)	(18.3)				
70	−233.8	−44.8	−5.0	32.9				
	(39.7)	(36.7)	(29.3)	(24.2)				
71	−325.9	−298.2	−29.4	27.6				
	(46.6)	(41.7)	(40.2)	(30.3)				
72	−203.5	−197.4	−261.6	2.1				
	(55.4)	(51.1)	(46.8)	(42.9)				
73	−226.6	−228.8	−357.7	−56.5				
	(67.8)	(61.6)	(56.2)	(54.8)				
74	−243.0	−155.4	−402.7	−15.0				
	(81.4)	(75.3)	(68.3)	(68.1)				
75	−295.2	−99.2	−304.5	−28.3				
	(94.4)	(89.7)	(85.0)	(79.6)				

Estimador de Wald 1/2

- En un modelo con D endógena: $Y_i = \beta_0 + \alpha D_i + \varepsilon_i$
- Con Z una VI válida, se puede escribir $\alpha = Cov(Y_i, Z_i)/Cov(D_i, Z_i)$
- Si Z es una variable binaria, que toma el valor uno con probabilidad p, para cualquier variable W podemos escribir:

$$\begin{aligned} Cov(W_i, Z_i) &= E[W_i Z_i] - E[W_i]E[Z_i] \\ &= \{E[W_i|Z_i = 1] - E[W_i|Z_i = 0]\}\, p * (1 - p) \end{aligned}$$

- Por lo tanto:

$$\alpha = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{E[Y_i|Z_i=1] - E[Y_i|Z_i=0]}{E[D_i|Z_i=1] - E[D_i|Z_i=0]}$$

Estimador de Wald 2/2

- Si además D es binaria, como en el caso de representar al tratamiento:
  - $E[D_i|Z_i = 1]$ es la proporción de tratados $(D = 1)$ en el total de observaciones con $Z = 1$.
  - $E[D_i|Z_i = 0]$ es la proporción de tratados $(D = 1)$ en el total de observaciones con $Z = 0$.
  - El denominador captura el impacto del instrumento en la probabilidad de formar parte del grupo de tratamiento.
- El análogo muestral de α se conoce como estimador de Wald.

Estimador de Wald (incluyendo controles X):

$$\hat{\alpha}_W(X) = \frac{\overline{Y}(X,Z=1) - \overline{Y}(X,Z=0)}{\overline{P}_{D=1}(X,Z=1) - \overline{P}_{D=1}(X,Z=0)}$$

Estimador de Wald en este caso

- Numerador:

  $\overline{Y}(X, Z = 1)$: ingresos medios de los sorteados ($Z_i = 1$).

  $\overline{Y}(X, Z = 0)$: ingresos medios de los no sorteados ($Z_i = 0$).

- Denominador:

  $\overline{P}_{D=1}(X, Z = 1)$: proporción de veteranos ($D_i = 1$) en el grupo de sorteados ($Z_i = 1$): de todos los elegidos en el sorteo, qué proporción efectivamente fue al ejército.

  $\overline{P}_{D=1}(X, Z = 0)$: proporción de veteranos en el grupo de no sorteados ($Z_i = 0$): de los que no fueron elegidos en el sorteo, qué proporción igualmente fue al ejército.

Resultados

Table 2 IV estimates of the effects of military service on US white men born 1950

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean	Eligibility effect	Mean	Eligibility effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	−435.8	0.267	0.159	−2,741
		(210.5)		(.040)	(1,324)
1970	2,758	−233.8			−1,470
		(39.7)			(250)
1969	2,299	−2.0			
		(34.5)			

Notes: Figures are in nominal US dollars. There are about 13,500 observations with earnings in each cohort. Standard errors are shown in parentheses.

Tabla tomada de Angrist y Pischke, Mostly Harmless Econometrics.

- Para los hombres nacidos en 1950, hay efectos negativos significativos de haber sido seleccionado en los ingresos en 1970, cuando estos hombres comenzaban el servicio militar y en 1981, diez años después.
- En contraste, no hay evidencia de una asociación entre haber sido elegido en el sorteo y los ingresos en 1969, el año que el sorteo para los hombres nacidos en 1950 se realizó pero nadie fue realmente seleccionado.

Estimador de Wald

- Para pasar de los efectos del sorteo a los efectos de ser veterano se necesita el denominador del estimador de Wald, que es el efecto del sorteo en la probabilidad de ser veterano.
- Esta información se reporta en la columna (4), que muestra que los hombres sorteados tienen una probabilidad mayor de servir en el ejército en Vietnam que los no sorteados. La diferencia $\overline{P}_{D=1}(X, Z=1) - \overline{P}_{D=1}(X, Z=0)$ es de 0.16 puntos.
- En 1981, mucho después de que estos veteranos dejaran el ejército, el estimador de Wald es aproximadamente 17% del ingreso medio.
- Los efectos son en términos porcentuales mayores en 1970, cuando los soldados afectados estaban todavía en el ejército.

Fuentes de Endogeneidad

Fuentes de endogeneidad: Motivación

- Hasta ahora, hemos justificado la endogeneidad de los controles como un problema de variables omitidas

- Hoy vamos a estudiar dos fuentes alternativas de endogeneidad:
  - errores de medida en las variables
  - simultaneidad

Error de medida en las variables

Error de medida en variable dependiente

- Supongamos que el verdadero modelo es $Y^* = \beta_0 + \beta_1 X + u^*$
- En lugar de Y^*, observamos $Y = Y^* + e$ de modo que

$$Y = \beta_0 + \beta_1 X + (u^* + e)$$

- Si $\mathrm{E}(e) \neq 0$, la estimación de MCO de β_0 será inconsistente
- Si $\mathrm{cov}(x, e) \neq 0$, la estimación de MCO de β_1 será inconsistente

Error de medida en una varible de control

- Supongamos que el modelo verdadero es

$$Y = \beta_0 + \beta_1 X^* + u^*$$

- En lugar de X^*, observamos $X = X^* + e$ with $\mathrm{E}(e) = 0$ de modo que

$$Y = \beta_0 + \beta_1 X + (u^* - \beta_1 e)$$

Cuando e no está correlacionado con X y con u^*

Suponga que el error de medición no está correlacionado con el control observado y el término de error estructural:

$$\text{cov}(X, e) = 0 \text{ and } \text{cov}(e, u^*) = 0$$

$$\Rightarrow \text{cov}(X, u) = \text{cov}(X, u^* - \beta_1 e) = \text{cov}(X, u^*) - \beta_1 \text{cov}(X, e) = 0$$

- MCO es consistente
- Las estimaciones tendrán desviaciones estándar mayores:

  $\text{Var}(u) = \text{Var}(u^*) + \beta_1^2 \text{Var}(e)$

Cuando e no está correlacionado con X^* y con u^*

Si, de manera más realista, el error de medición no está correlacionado con el verdadero control y el término de error estructural:

$$\text{cov}(X, e) = 0 \text{ and } \text{cov}(e, u^*) = 0$$

$$\Rightarrow \text{cov}(X, u) = \text{cov}(X^* + e, u^* - \beta_1 e) = -\beta_1 \text{Var}(e) \neq 0$$

- MCO no es consistente: $\text{plim}\ \hat{\beta}_{OLS} = \beta \left(1 - \frac{\text{Var}(e)}{\text{Var}(x)}\right) < \beta$ (sesgo de atenuación)
- Con varios controles, todos sus estimadores serán inconsistentes (aunque la dirección del sesgo para los otros controles no está clara)

Ejemplo: ecuaciones de ahorro

- Suponga que desea estimar la propensión marginal a ahorrar
- Ecuación de ahorro: $sav = \beta_0 + \beta\, inc^* + u$
- Ingresos observados: $inc = inc^* + e$
- Ecuación realmente estimada: $sav = \beta_0 + \beta\, inc + (u - \beta\, e)$

Ahorro e instrumentos

- Si el error de medición no está correlacionado con los ingresos reales,

$$\text{cov}(inc, e) = \text{Var}(e) \neq 0 \Rightarrow \text{cov}(inc, u - \beta e) = -\beta \text{var}(e)$$

- MCO no es consistente: plim $\hat{\beta}_{OLS} = \beta \left(1 - \frac{\text{Var}(e)}{\text{Var}(inc)}\right) < \beta$ (sesgo de atenuación)
- Podemos utilizar VI: necesitamos una variable
  - correlacionada con el verdadero ingresos (relevancia)
  - no correlacionado con el error de medición en los ingresos observados
- Cualquier segunda medida de los ingresos (de la empresa, un cónyuge, ...) sería un buen instrumento
- Alternativamente, otra proxy de los ingresos, como el tamaño de la casa.

Simultaneidad: estimación de una función de demanda

Un sistema de ecuaciones de oferta y demanda

- Función de oferta: $q^s = \gamma_0 + \beta^s p + \gamma x^s + u^s$
- Función de demanda: $q^d = \alpha_0 + \beta^d p + \alpha x^d + u^d$

- q^s is la cantidad ofrecida, q^d es la demandada, y p es el precio
- x^s es un factor exógeno que es observado por el econometrista y afecta solo a la oferta
- x^d es un factor exógeno que afecta a la demanda
- u^s son los efectos de factores no observables que afectan la curva de oferta
- u^d son los efectos de factores no observables que afectan la curva de demanda

Precio de equlibrio

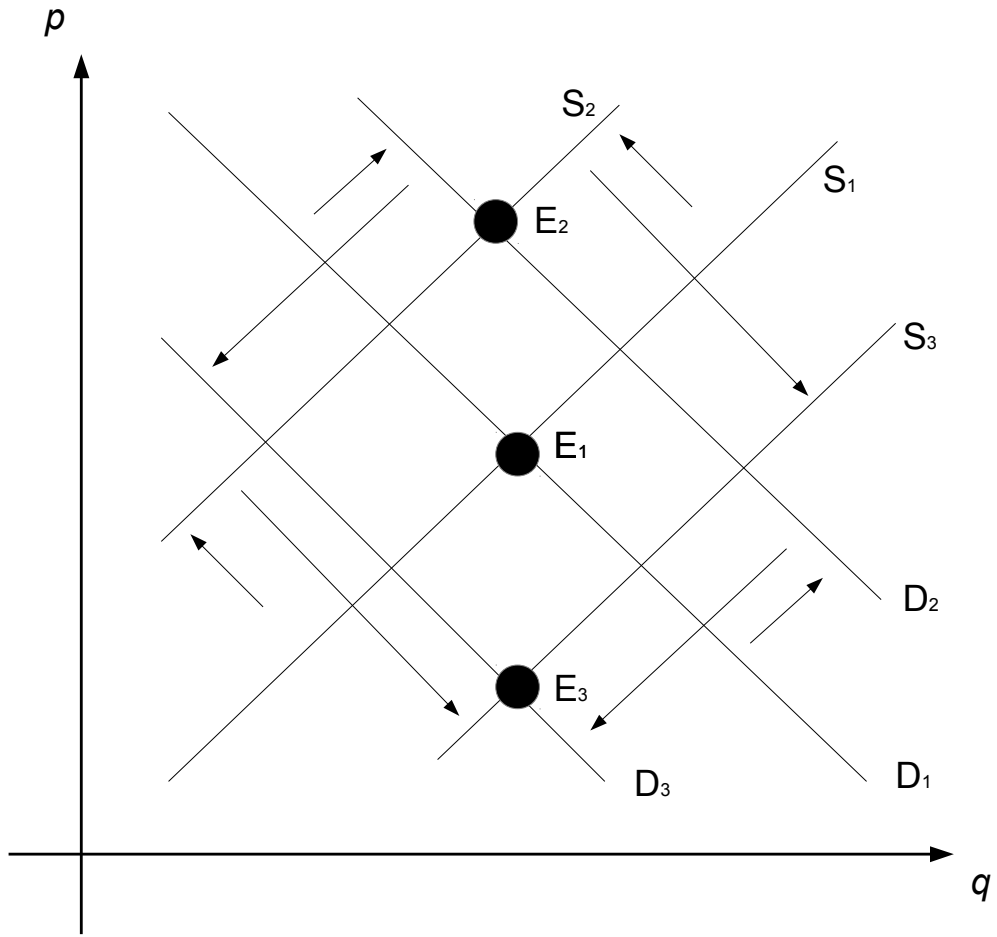
En el equilibrio, $q^s = q^d = q$

- $q^s = q^d \Rightarrow \gamma_0 + \beta^s p + \gamma x^s + u^s = \alpha_0 + \beta^d p + \alpha x^d + u^d$
- De este modo, en el equilibrio,
  $p = \left(\frac{1}{\beta^s - \beta^d}\right)\left[(\alpha_0 - \gamma_0) + (\alpha x^d - \gamma x^s) + (u^d - u^s)\right]$
- En el equilibrio, los precios dependen de factores que desplazan la demanda (x^d y u^d) y aquellos que desplazan la oferta (x^s y u^s)

Simultaneidad

- Los precios y cantidades observados se determinan simultáneamente
- Ambos dependen de los factores que determinan la demanda (x^d and u^d) y aquellos que desplazan la oferta (x^s and u^s)
- Si hacemos una regresión de q como una función de p y x^s por medio de MCO, $\hat{\beta}^s$ no es consistente porque $\text{cov}(p, u^s) \neq 0$
- Del mismo modo, si hacemos una regresión de q como función de p y x^d por medio de MCO, $\hat{\beta}^d$ no es consistente porque $\text{cov}\left(p, u^d\right) \neq 0$

Una interpretación gráfica



Las observaciones no revelan la relación negativa entre demanda y precio.

Instrumentos para la ecuación de demanda

- Queremos estimar la ecuación de demanda: $q = \alpha_0 + \beta^d p + \alpha x^d + u^d$
- En equilibrio, $p = f\left(x^d, x^s, u^d, u^s\right)$
- Un instrumento es cualquier variable que, independientemente de los otros controles, está correlacionada con p (relevancia) y no está correlacionada con u^d
- x^d ya son controles en la ecuación de demanda, por lo que no pueden ser instrumentos
- x^s son potencialmente buenos instrumentos:
  - $\text{cov}\left(x^s, p\right) \neq 0$ (relevancia) (ya que p es una función de x^s)
  - $\text{cov}\left(x^s, u^d\right) = 0$ (exogeneidad) (en caso contrario x^sno es exógeno en primer lugar)

Resumen

- Si un regresor se mide con error, entonces puede ser endógeno.
- Si tenemos variables adicionales que actúan como proxies del regresor, podríamos implementar MC2E
- Cuando se determinan dos variables simultáneamente, ambas son endógenas
- Si queremos estimar una ecuación de demanda, necesitamos MC2E y factores que solo desplacen a la oferta.