

Tópicos de Economía Aplicada

*Êstimaciones*

## Modelo causal de Rubin y Educación

El economista empírico tiene un modelo que utiliza para explicitar los efectos de una variable de tratamiento sobre una variable resultado y para entender qué método permite recuperar (inferir) este efecto. Más en general:

- ▶ Una población dada tiene unos **parámetros** que queremos estimar (por ejemplo, un efecto de una variable sobre otra)
- ▶ El **modelo** define el proceso que genera los datos observados de esa población
- ▶ El método construye un **estimador** para aplicar en los datos
- ▶ El resultado en unos datos es una **estimación**

El objetivo es que esta estimación sea una adecuada medición del parámetro (estimación puntual)

En esta clase:

- ▶ vamos a ver un ejemplo de modelos: el modelo de educación, que vincula el nivel educativo a los ingresos laborales
- ▶ luego vamos a ver métodos o diseños de investigación para estimar el efecto: empezamos por asignación aleatoria

# Tabla de contenidos

## Modelo de educación

- El problema de la elección de la educación

- Modelo empírico

- Simulación

- Una regresión sobre los datos simulados

## Modelo causal de Rubin

## Asignación aleatoria

- El ideal del experimento

- Ejemplo

- Potenciales problemas

## Modelo de educación

- El problema de la elección de la educación

- Modelo empírico

- Simulación

- Una regresión sobre los datos simulados

## Modelo causal de Rubin

## Asignación aleatoria

- El ideal del experimento

- Ejemplo

- Potenciales problemas

# Un modelo de educación

Queremos estudiar diferentes métodos para identificar un efecto  
Para eso, vamos a considerar un problema sencillo en donde los individuos pueden tener decisión en su tratamiento

El ejemplo es un modelo de elección de educación

Especificaremos de manera general un modelo de capital humano lo más sencillo posible para luego utilizarlo como ejemplo en las estimaciones.

# Un modelo de educación

Supongamos un individuo cuyo ingreso en la vida (valor presente neto del ingreso laboral, descontado al momento en que decide su educación) depende de la educación y también de otras características

$$y = f(d, \theta, X, \tau)$$

donde  $y$  es el ingreso laboral (en valor presente),  $d = \{0, 1\}$  es la variable binaria que identifica el nivel educativo (diploma universitario),  $\theta \geq 0$  es una variable de habilidades (innatas),  $X$  es un vector de otras variables predeterminadas conocidas por el individuo al momento de elegir la educación, como el nivel socioeconómico familiar. Por otro lado,  $\tau$  indica un período de tiempo (por ejemplo el año de la elección de educación). Se asume  $\frac{\partial f}{\partial d} \geq 0$ ,  $\frac{\partial f}{\partial \theta} \geq 0$ .

## Un modelo de educación

Los costos de estudiar para cada persona dependen de condiciones reflejadas en un vector de variables  $Z$  y potencialmente de  $\theta$ ,  $X$  y de un indicador  $s$

$$c = g(\theta, s, Z, X, \tau)$$

Una variable  $z$  puede ser el costo de estudiar o la distancia a la universidad; una variable  $x$  puede ser la educación de los padres;  $s$  puede ser un indicador (un puntaje en un test) que cambie las chances de ser elegible para una beca.

Asumimos que el individuo decide su educación para maximizar (el valor presente de) su ingreso,

$$\max_{d \in \{0,1\}} f(d, \theta, X, \tau) - d \times g(\theta, s, Z, X, \tau)$$

y entonces elige educarse siempre que

$$f(1, \theta, X, \tau) - g(\theta, s, Z, X, \tau) \geq f(0, \theta, X, \tau)$$

# Un modelo empírico:

Para pensar en una contraparte empírica de este modelo reescribimos las ecuaciones.

- ▶ Asumimos funciones lineales
- ▶ Asumimos errores (con media cero) inobservables para el econometrista
- ▶ Asumimos que los efectos de las variables  $X$  y  $Z$  son comunes a todos los individuos pero que los efectos de la educación dependen de  $\theta$ .

Los costos y el ingreso para cada individuo  $i$ :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 \theta_i + \alpha_0 d_i + \alpha_1 \theta_i d_i + \gamma_0 \tau + u_i$$

$$c_i = \delta_0 + \delta_1 z_{1i} + \delta_x \mathbf{1}\{x_{3i} = \bar{x}\} + \delta_s \mathbf{1}\{s_i \geq \bar{s}\} + \delta_\tau \tau + \dots + v_i$$

donde las variables  $X$ ,  $s$ ,  $\tau$  y  $Z$  son información observable (para el individuo y en los datos) antes de tomar la decisión de educación, las variables  $\theta$  y  $v$  son conocidas sólo para el individuo (no es observable en los datos) antes de tomar la decisión, y la variable  $u$  sólo es conocida por el individuo al final de su vida laboral.



# Elección de educación

Luego,

$$d_i = \begin{cases} 1 & \text{si } \mathbb{E}_u [y_i | X_i, \theta_i, \tau, d = 1] - \mathbb{E}_u [y_i | X_i, \theta_i, \tau, d = 0] \geq c_i \\ 0 & \text{en otro caso} \end{cases}$$

Vale la pena notar que

$$\begin{aligned} \mathbb{E}_u [y_i | X_i, \theta_i, \tau, d = 1] - \mathbb{E}_u [y_i | X_i, \theta_i, \tau, d = 0] &= \alpha_i \\ \alpha_i &\equiv \alpha_0 + \alpha_1 \theta_i \end{aligned}$$

# Datos (simulados) de educación:

Ahora simulamos unos datos

Nuestro ejemplo de educación básico se basa en las siguientes ecuaciones

$$y_i = 1 + ,2x_{1i} + ,09x_{2i} + ,07\theta_i + \alpha_i d_i + ,06\tau + 0,7u_{1i}$$

$$c_i = 0,4 - z_{0i} + 1,1z_{1i} - 0,01x_{2i} + \dots + 0,1u_{2i}$$

$$d_i = \mathbf{1} \{ \alpha_i \geq c_i \}$$

donde  $u_{1i}$ ,  $u_{2i}$  son los errores (inobservables) tienen distribución normal estándar independiente,  $\alpha_i = 0,05 + 0,05 * \theta_i$ , por lo que  $cov(\theta_i, \alpha_i) > 0$ , donde  $\theta$  es habilidad y  $\alpha$  es retorno a la educación,  $x_1 = 1$  indica hombre,  $x_2$  es la cantidad de años de estudio del padre,  $\tau = \{0, 1\}$  hace referencia a dos períodos,  $z_0$  y  $z_1$  son variables de costos y donde  $d_i$  es la decisión de educación, que será 1 si  $\alpha_i \geq c_i$ . La variable  $x_3$  es una variable de región que usaremos a lo largo de los ejemplos para definir diferentes formas de selección al tratamiento. Además, hay dos variables relevantes:  $a_1$  que es una variable aleatoria binaria, y  $s = 20 + 2\theta_i + err$  que simula el resultado en un test.

# Estimación del efecto de la educación:

El desafío es estimar el efecto de la educación cuando no observamos la habilidad.

El efecto de la educación depende de  $\theta_i$  en este ejemplo:

$$\alpha_i = \alpha_0 + \alpha_1 \theta_i$$

Por la simulación que utilizamos,

$$\mathbb{E}[\alpha] = 0,5$$

Por otro lado, un mayor  $\theta_i$  también implica mayor probabilidad de decidir educarse (porque hay una decisión de educación que depende de  $\theta_i$ )

## Regresión: sesgo de variable omitida

El parámetro de interés es  $\mathbb{E} [\alpha_i]$

Una regresión de  $y$  en  $X$  y  $d$  (sin observar  $\theta$ ) tiene un sesgo y no recupera  $\mathbb{E} [\alpha_i]$ : no puede separar el efecto de la educación del efecto de tener más habilidades ( $\theta$  más alto). Definiendo el error  $e = \beta_3\theta + u_1$ :

$$\mathbb{E} [y|X, d] = \beta_0 + X'\beta + (\alpha_0 + \alpha_1\mathbb{E} [\theta|X, d]) d + \mathbb{E} [e|X, d]$$

$$\mathbb{E} [y|X, d = 1] = \beta_0 + X'\beta + (\alpha_0 + \alpha_1\mathbb{E} [\theta|X, d = 1]) + \mathbb{E} [e|X, d = 1]$$

$$\mathbb{E} [y|X, d = 0] = \beta_0 + X'\beta + \mathbb{E} [e|X, d = 0]$$

luego

$$\begin{aligned}\mathbb{E} [\alpha^{OLS}] &= \mathbb{E} [y|X, d = 1] - \mathbb{E} [y|X, d = 0] \\ &= \alpha_0 + \alpha_1\mathbb{E} [\theta|d = 1] \\ &\quad + \beta_3 (\mathbb{E} [\theta|d = 1] - \mathbb{E} [\theta|d = 0]) \\ &\quad + \mathbb{E} [u_1|d = 1] - \mathbb{E} [u_1|d = 0]\end{aligned}$$

# Sesgo de variable omitida

El sesgo que incurrimos con esta regresión es la diferencia entre  $\mathbb{E} [\alpha_i]$  y  $\mathbb{E} [\alpha^{OLS}]$  donde  $\mathbb{E} [\alpha_i] = \alpha_0 + \alpha_1 \mathbb{E} [\theta]$  y donde

$$\mathbb{E} [\alpha^{OLS}] = \alpha_0 + \alpha_1 \mathbb{E} [\theta | d = 1] \quad (1)$$

$$+ \beta_3 (\mathbb{E} [\theta | d = 1] - \mathbb{E} [\theta | d = 0]) \quad (2)$$

$$+ \mathbb{E} [u_1 | d = 1] - \mathbb{E} [u_1 | d = 0] \quad (3)$$

(1) Si estudiar depende de  $\theta$  entonces  $\mathbb{E} [\theta | d = 1] > \mathbb{E} [\theta]$ : cuando  $\alpha_1 > 0$  no se mide el efecto promedio sino el efecto dentro del grupo de los que estudiaron (Esta parte del sesgo se debe a la heterogeneidad del efecto de la educación).

(2) Si estudiar depende de  $\theta$  entonces  $\mathbb{E} [\theta | d = 1] > \mathbb{E} [\theta | d = 0]$  y la estimación está sesgando (sobrestimando) el efecto de estudiar: los que estudian tienen alto  $\theta$  y esta variable tiene efecto directo sobre  $y$  ( $\beta_3 > 0$ ). (Esta parte del sesgo se debe a endogeneidad por variable omitida, con  $\text{cov}(d, \theta) \neq 0$ )

(3) Si  $u_1$  es inobservable para el individuo al momento de tomar la decisión, entonces  $\mathbb{E} [u_1 | d] = 0$  (aquí no hay sesgo).  
 $\text{cov}(d, \theta) \neq 0$

## Usando los datos simulados:

Veamos en los datos (simulados) cuál es el efecto.

Sabemos que por nuestra simulación el promedio de  $\alpha_i$  es 0.5.

Supongamos que hacemos la regresión

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \alpha d_i$$

Esta regresión da como resultado

$$\hat{y} = 1,6 + 0,196x_1 + 0,087x_2 + 0,697d$$

donde queda evidente el fuerte sesgo en la estimación de  $\alpha$  (no así para otros parámetros).

. reg y d x1 x2 if x3>=6 // se mantiene el sesgo por variables omitidas dado que d es endógena.

Source	SS	df	MS	Number of obs =	12500
Model	3964.19205	3	1321.39735	F( 3, 12496) =	1744.25
Residual	9466.64986	12496	.757574412	Prob > F =	0.0000
				R-squared =	0.2952
				Adj R-squared =	0.2950
Total	13430.8419	12499	1.07455332	Root MSE =	.87039

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d	.681329	.0156774	43.46	0.000	.6505988 .7120592
x1	.1990982	.0155976	12.76	0.000	.1685245 .229672
x2	.0879819	.0017207	51.13	0.000	.0846091 .0913547
_cons	1.562213	.0181678	85.99	0.000	1.526601 1.597825

. reg y d x1 x2 theta if x3>=6 // controlando por theta desaparece el sesgo (es variable omitida relevante)

Source	SS	df	MS	Number of obs =	12500
Model	7184.619	4	1796.15475	F( 4, 12495) =	3593.04
Residual	6246.2229	12495	.499897791	Prob > F =	0.0000
				R-squared =	0.5349
				Adj R-squared =	0.5348
Total	13430.8419	12499	1.07455332	Root MSE =	.70703

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d	.5083431	.0129162	39.36	0.000	.4830254 .5336608
x1	.195057	.0126704	15.39	0.000	.1702211 .2198929
x2	.0876944	.0013977	62.74	0.000	.0849546 .0904342
theta	.0998815	.0012444	80.26	0.000	.0974422 .1023207
_cons	.7624454	.017807	42.82	0.000	.7275409 .7973498

## Usando los datos simulados

La estimación por OLS se puede calcular de la siguiente manera (donde  $\bar{x}_{d=0} \equiv \frac{\sum_{i \in d=0} x_i}{\sum_i \mathbf{1}\{d=0\}}$  es el promedio de la variable para las observaciones con  $d = 0$ )

$$\alpha^{OLS} = \bar{\alpha}_{d=1} + [\beta_3 (\bar{\theta}_{d=1} - \bar{\theta}_{d=0})] + [\bar{u}_{d=1} - \bar{u}_{d=0}]$$

donde el último término no implica un sesgo sino que surge sólo por ser una muestra pequeña.

El segundo término es un sesgo por endogeneidad (variable omitida) que en este caso se solucionaría observando  $\theta$ .

El desafío es (en nuestro ejemplo) estimar el efecto de la educación sin observar  $\theta$ .



# Resumiendo

Hasta ahora:

- ▶ Pensamos un modelo
- ▶ Escribimos una versión empírica
- ▶ Simulamos el modelo
- ▶ Realizamos una estimación (mediante una regresión)

Durante el curso

- ▶ asumiremos diferentes procesos de generación de los datos dentro de este modelo.
- ▶ consideraremos diferentes estimadores
- ▶ evaluaremos los estimadores mediante derivaciones y proposiciones formales
- ▶ usaremos los datos simulados para evaluar estimadores alternativos de nuestro parámetro de interés (asumiendo  $\theta$  como inobservable).

El primer ejemplo que veremos es el de asignación aleatoria. Pero antes presentaremos el modelo causal de Rubin, una versión más sintética que el modelo de educación.

## Modelo de educación

- El problema de la elección de la educación

- Modelo empírico

- Simulación

- Una regresión sobre los datos simulados

## Modelo causal de Rubin

## Asignación aleatoria

- El ideal del experimento

- Ejemplo

- Potenciales problemas

# El modelo causal de Rubin: respuestas potenciales

- ▶ Para cada individuo hay respuestas potenciales (*potential outcomes*) según sea tratado o no tratado
- ▶ Podemos especificar

$$\begin{aligned}y_i^1 &= \beta + \alpha_i + u_i \text{ si el individuo es tratado } (d_i = 1) \\y_i^0 &= \beta + u_i \text{ si el individuo no es tratado } (d_i = 0)\end{aligned}$$

donde  $\beta$  es un nivel común a todos los individuos,  $\alpha_i$  es el efecto individual y  $u_i$  es un componente aleatorio de cada individuo.

- ▶ La variable binaria  $d$  identifica al tratado ( $d = 1$ )
- ▶ El efecto del tratamiento para el individuo  $i$  es

$$\alpha_i = y_i^1 - y_i^0$$

- ▶ Se explicita así que el efecto del tratamiento puede ser heterogéneo entre las personas (el subíndice  $i$  en el parámetro)

# Modelo causal de Rubin: respuestas potenciales

- El individuo tendrá entonces

$$y_i = d_i y_i^1 + (1 - d_i) y_i^0$$

$$y_i = \beta + \alpha_i d_i + u_i$$

- A su vez, el tratamiento depende de variables  $Z$  y un inobservable (para el econometrista)  $v$ , de forma tal que

$$d_i = \begin{cases} 1 & \text{si } g(Z_i, v_i) \geq 0 \\ 0 & \text{si } g(Z_i, v_i) < 0 \end{cases}$$

o lo que es lo mismo  $d_i = \mathbf{1}_{\{g(Z_i, v_i) \geq 0\}}$ , donde  $\mathbf{1}_A\{x\}$  es la función indicadora que toma valor uno si  $x \in A$  y cero si  $x \notin A$ . (1 si se cumple la condición y cero si no).

- Un caso especial es cuando hay una asignación al tratamiento definida por  $Z$ , por ejemplo  $z = \{0, 1\}$  o  $z \geq z^*$

# Modelo causal de Rubin: contrafactual

- ▶ *Missing data problem:*
- ▶ El problema central es que no es factible observar  $y_i^0$  e  $y_i^1$  a la vez (si observáramos ambos valores podríamos observar el  $\alpha_i$  para cada individuo, la situación ideal)
- ▶ El individuo es tratado y obtiene  $y_i^1$  o bien es no tratado y obtiene  $y_i^0$  pero no puede ser las dos cosas a la vez.
- ▶ A este problema se lo suele llamar *missing data problem*
- ▶ El desafío de la estimación del efecto de  $d$  sobre  $y$  es el de conseguir un contrafactual (una aproximación a  $y^0$  para los tratados) lo más adecuada posible.
- ▶ Los métodos que buscan atribuir un efecto causal del impacto de un programa o “tratamiento”  $d$  sobre un resultado  $y$  los denominamos métodos de *evaluación de impacto*

# Modelo causal de Rubin: parámetros y estimaciones de impacto

- ▶ La información disponible permite construir diferentes parámetros del impacto del tratamiento. (Por ahora nos concentramos en las definiciones más importantes y veremos las otras más adelante.)
- ▶ Aquí damos la versión poblacional (la expresión muestral tendrá promedios en vez de esperanzas).
- ▶ El efecto promedio en la población (Average Treatment Effect)

$$\alpha^{ATE} = \mathbb{E} [\alpha_i]$$

Este parámetro es el efecto promedio del tratamiento, tanto de los que fueron tratados como de los que no fueron tratados. Es un indicador amplio y general. Sirva para responder la pregunta cuál sería el diferencial de ingreso por educarse esperado de un individuo elegido al azar.

- El efecto promedio para los tratados (Average Treatment on the Treated)

$$\alpha^{ATT} = \mathbb{E} [\alpha_i | d_i = 1]$$

Este parámetro se concentra sólo en el efecto sobre los tratados. Si los tratados son un grupo particular, con efectos particularmente altos, entonces el efecto esperable sobre otros grupos de individuos será menor. Sirve para estimar cuál sería el diferencial de ingreso por educarse esperado de un individuo que estudió elegido al azar.

- El efecto promedio sobre los asignados al tratamiento (Intention To Treat effect)

$$\alpha^{ITT} = \mathbb{E} [y_i | Z_i = 1] - \mathbb{E} [y_i | Z_i = 0]$$

cuando

$$d_i = \mathbf{1} \{ z_i = 1 \cap v_i > v^* \} \text{ y } z_i \in \{0, 1\}$$

La utilización más frecuente de este parámetro es cuando hay una asignación  $Z$  exógena o aleatoria al tratamiento, pero en donde el tratamiento puede no tener cumplimiento perfecto. Por ejemplo, se brinda una beca de educación aleatoriamente a unos estudiantes y a otros no, pero de los elegidos para la beca, algunos deciden no educarse. En ese caso, lo que se mide no es el efecto de la educación, sino el efecto de la beca sobre el ingreso (teniendo en cuenta que algunos que reciben la beca no se educarán).



# Modelo causal de Rubin: parámetros y estimaciones de impacto vs regresiones

- ▶ Como comparación también podemos definir un estimador de regresión
- ▶ La variable  $y_i = \beta + \alpha_i d_i + u_i$ , el resultado para cada individuo  $i$ , puede reescribirse

$$y_i = \beta + \alpha^{ATE} d_i + e_i$$

con

$$e_i = u_i + (\alpha_i - \alpha^{ATE}) d_i$$

a partir de sumar y restar  $d_i \alpha^{ATE}$

- ▶ El estimador de una regresión por OLS de  $y$  en  $d$  identificará

$$\begin{aligned}\alpha^{OLS} &= \alpha^{ATE} + \mathbb{E}[e_i | d_i = 1] - \mathbb{E}[e_i | d_i = 0] \\ &= \alpha^{ATE} + \mathbb{E}[\alpha_i - \alpha^{ATE} | d_i = 1] + \mathbb{E}[u_i | d_i = 1] - \mathbb{E}[u_i | d_i = 0]\end{aligned}$$

- ▶ Es decir que el estimador  $OLS$  será un  $\alpha^{ATE}$  en la medida en que  $e_i$  no esté correlacionado con  $d_i$
- ▶ Por otro lado, el estimador  $OLS$  será un  $\alpha^{ATT}$  en la medida en que  $u_i$  no esté correlacionado con  $d_i$

# Modelo causal de Rubin: parámetros y estimaciones de impacto: ejemplos y aclaraciones

- ▶ Las definiciones no se modifican si los efectos fueran homogéneos
- ▶ Si los efectos fueran homogéneos no habría diferencias entre ellos
- ▶ A lo largo del curso analizaremos las diferentes formas de estimar  $\alpha^{ATE}$  y otros parámetros
- ▶ Estos parámetros se podrán estimar según la forma de generar la información y de asignación al tratamiento

# Modelo causal de Rubin: parámetros y estimaciones de impacto: resumen

- ▶ Average Treatment Effect:  $\alpha^{ATE} = \mathbb{E} [\alpha_i]$
- ▶ Average Treatment on the Treated:  $\alpha^{ATT} = \mathbb{E} [\alpha_i | d_i = 1]$
- ▶ Intention To Treat:  $\alpha^{ITT} = \mathbb{E} [y_i | Z_i = 1] - \mathbb{E} [y_i | Z_i = 0]$ ,  
cuando  $d_i = \mathbf{1} \{z_i = 1 \cap v_i > v^*\}$  y  $z_i \in \{0, 1\}$

# Modelo causal de Rubin: estimaciones, resumen (continuación)

- ▶ Es importante aclarar que este resumen define los parámetros, no una forma práctica de medirlos
- ▶ A lo largo del curso veremos cómo pasar de estos parámetros a métodos y estimadores concretos de medición

# Modelo causal de Rubin: parámetros en la simulación de educación

- ▶ Ahora veremos en los datos: vamos a computar estos parámetros en nuestros datos simulados: lo haremos como si observáramos  $\alpha_i$ ; sólo lo podemos hacer porque estamos usando nuestra simulación.
- ▶ Van a ser los parámetros que tendremos como referencia en cada ejercicio de estimación: querríamos estimar ATE.
- ▶ Utilizando nuestro ejemplo de educación, con datos simulados

$$y_i = 1 + ,2x_{1i} + ,09x_{2i} + ,07\theta_i + \alpha_id_i + ,06\tau + 0,7u_{1i}$$

donde para algunas regiones (identificadas con la variable  $x_3$ ) la asignación se basa en:

$$d_i = \mathbf{1}\{\alpha_i \geq c_i\}$$

$$c_i = \delta_0 - z_{0i} + 1,1z_{1i} - 0,01x_{2i} + \dots + 0,1u_{2i}$$

► Average Treatment Effect:  $\hat{\alpha}^{ATE} = \frac{\sum_i \alpha_i}{N}$

. sum alfai if x3>=6

► Average Treatment on the Treated:  $\hat{\alpha}^{ATT} = \frac{\sum_{i \in d_i=1} \alpha_i}{\sum_i d_i}$

. sum alfai if x3>=6&d==1

```
. use modeloeducacion.dta, clear
```

```
. *=====
```

```
. **Parámetros
```

```
. sum alfai // ATE
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
alfai	50000	.5001181	.2597051	.05475	2.449853

```
. sum alfai if x3>=6 // ATE de regiones 6+
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
alfai	12500	.497057	.2577744	.0633007	2.449853

```
. sum alfai if x3>=6&d==1 // ATT de regiones 6+
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
alfai	6434	.5391511	.2724067	.0675655	1.911025

```
. sum alfai if x3==2&a1==1 // ATT de región 1 (asignación aleatoria)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
alfai	1943	.4966661	.2554631	.0767537	1.834277

```
. sum alfai if x3==2&d==1 // ATT de regiones 2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
alfai	3115	.5046221	.2628834	.0726849	2.093348

## Modelo de educación

- El problema de la elección de la educación

- Modelo empírico

- Simulación

- Una regresión sobre los datos simulados

## Modelo causal de Rubin

## Asignación aleatoria

- El ideal del experimento

- Ejemplo

- Potenciales problemas



# Asignación aleatoria

- ▶ Asumimos que los datos se generan asignando el tratamiento de manera aleatoria
- ▶ Veremos qué implica este supuesto y cómo podemos testear si los datos que estamos utilizando son consistentes con este supuesto
- ▶ Veremos que un simple estimador de diferencia de medias (que se puede recuperar con una regresión) nos permite identificar el parámetro que buscamos
- ▶ Discutiremos los potenciales problemas y limitaciones de esta aproximación

# Asignación aleatoria: el ideal del experimento

- ▶ Ante el problema de estimar un efecto o un parámetro una pregunta relevante es “¿cuál sería el experimento ideal?”
- ▶ El experimento ideal se plantea sin restricciones (ni presupuestarias, ni de relevamiento de información)
- ▶ Si bien el experimento ideal raramente está disponible ni puede implementarse, es útil pensarlo
  - ▶ para definir qué tipo de datos y de variaciones “exógenas” son necesarias para identificar el efecto
  - ▶ para definir si se puede identificar el efecto (si no hay experimento ideal posible, entonces el efecto no está identificado)

# Asignación aleatoria: el ideal del experimento

- ▶ El ideal del experimento en economía sigue al ideal de los experimentos en medicina
- ▶ El objetivo es estudiar el efecto de una droga o un medicamento sobre un determinado grupo de población
  - ▶ Tomemos un grupo grande de personas (los afectados por la gripe X)
  - ▶ Le comunicamos a todo el grupo que van a participar de un experimento para una nueva droga. Van a tener que tomar una pastilla después de cada comida durante una semana.
  - ▶ Aleatoriamente se toma a la mitad de este grupo y se les da un blister con 28 pastillas redondas naranjas con la droga
  - ▶ La otra mitad se les da un blister con 28 pastillas redondas naranjas sin droga (un placebo)
  - ▶ (los participantes no saben si están tomando la droga o el placebo)
  - ▶ A la semana nos aseguramos que todo el mundo tomó sus 28 pastillas
  - ▶ Comparamos el progreso entre ambos grupos como el efecto de la pastilla

# Asignación aleatoria: el ideal del experimento

- ▶ Hace falta enfatizar algunos puntos importantes en este protocolo
1. El grupo es grande (luego la asignación aleatoria cumplirá con que las características idiosincráticas de cada persona no sea relevante en el promedio del grupo y ambos grupos quedarán balanceados)
  2. Se asigna aleatoriamente después de conocer el total del universo de las personas que participan en el experimento
  3. Las personas no saben a qué grupo están asignados (puede que ambos grupos piensen que están en tratamiento)
  4. Todo el mundo toma todas las pastillas y no toman otra cosa (compara tratamiento vs sin tratamiento)

# Asignación aleatoria

- ▶ El principal problema de la evaluación de impacto es que no podemos observar  $y_i^1$  a la vez que  $y_i^0$  para un individuo  $i$  (*missing data problem*)
- ▶ Por ejemplo, si  $d$  es una elección, un estimador basado en la diferencia  $\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0]$  puede tener un sesgo de selección
- ▶ Una asignación aleatoria al tratamiento puede generar un contrafactual correcto, evitando el sesgo de selección
- ▶ Dos supuestos esenciales de la asignación aleatoria:

$$R1 : \mathbb{E}[u_i|d_i = 1] = \mathbb{E}[u_i|d_i = 0]$$

$$R2 : \mathbb{E}[\alpha_i|d_i = 1] = \mathbb{E}[\alpha_i|d_i = 0]$$

- ▶ Los supuestos indican que los grupos tratado y no tratado son iguales a cualquier respecto salvo el tratamiento
- ▶ Estos dos supuestos (características de los datos) son esenciales para recuperar  $\alpha^{ATE}$
- ▶ El parámetro podrá ser capturado por el siguiente estimador basado en una diferencia de promedios

$$\begin{aligned}\hat{\alpha}^{ATE} &= \frac{\sum_{i \in d=1} y_i}{\sum_i d_i} - \frac{\sum_{i \in d=0} y_i}{\sum_i (1 - d_i)} \\ &= \bar{y}^1 - \bar{y}^0\end{aligned}$$

- ▶ Una vez asegurada la asignación aleatoria, el mismo coeficiente (y su intervalo de confianza) puede recuperarse de la regresión

$$y_i = \beta_0 + \beta_1 d_i + \epsilon_i$$

donde  $\hat{\alpha}^{ATE} = \hat{\beta}_1$

## Un ejemplo con datos simulados

En nuestros datos simulados del modelo de educación:

En la región 1 la asignación es aleatoria (depende sólo de una variable aleatoria  $\xi$ ):

$$\begin{aligned}y_i &= 1 + ,2x_{1i} + ,09x_{2i} + ,07\theta_i + \alpha_i d_i + ,06\tau + 0,7u_{1i} \\ a_{i1} &= \mathbf{1}\{\xi \geq 0,7\}, \text{ con } \xi \sim U(0, 1) \\ d_i &= a_{i1}\end{aligned}$$

► Parámetro ATE:

```
. sum alfai if x3==1
```

► Test de diferencias de medias:

```
. ttest y if x3==1, by(d)
```

► Regresión:

```
. reg y d if x3==1
```

```

. **Asignación aleatoria (sólo en región 1)
. use modeloeeducacion.dta, clear
. sum alfai if x3==1 // ATE

```

Variable	Obs	Mean	Std. Dev.	Min	Max
alfai	12500	.503629	.2615682	.05475	2.182912

```

. ttest y if x3==1, by(d) // test de diferencia de medias
Two-sample t test with equal variances

```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	8704	2.338549	.0096343	.8988338	2.319664 2.357435
1	3796	2.855836	.0171138	1.05441	2.822283 2.889389
combined	12500	2.495639	.0087484	.9781054	2.47849 2.512787
diff		-.5172867	.0184542		-.5534598 -.4811136

```

diff = mean(0) - mean(1) t = -28.0308
Ho: diff = 0 degrees of freedom = 12498
Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

```

```

. reg y d if x3==1 // test de diferencia de medias realizado con u
> na regresión

```

Source	SS	df	MS	Number of obs =	12500
Model	707.290366	1	707.290366	F( 1, 12498) =	785.73
Residual	11250.3807	12498	.900174482	Prob > F =	0.0000
Total	11957.671	12499	.956690219	R-squared =	0.0591
				Adj R-squared =	0.0591
				Root MSE =	.94878

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d	.5172867	.0184542	28.03	0.000	.4811136 .5534598
_cons	2.338549	.0101696	229.95	0.000	2.318615 2.358483



# Asignación aleatoria: potenciales problemas

Si bien este procedimiento es muy directo y sencillo, la aplicación en la práctica está plagada de potenciales problemas y dificultades

1. La implementación de una asignación aleatoria suele chocar con **problemas éticos** (de privar a un grupo de personas del potencial beneficio del tratamiento), problemas legales (principalmente por parte de la administración pública) y restricciones prácticas (los funcionarios pueden estar en contra de una asignación aleatoria). En todo caso, la *restricción presupuestaria* suele ser un argumento muy realista que impide realizar una asignación a todos los elegibles, lo que podría salvar los problemas éticos.
2. Un problema en la asignación aleatoria es la **muestra pequeña**. En ese caso, podría ser que el tratamiento y el control estén afectados por cuestiones idiosincráticas o características de las personas. Si la muestra es pequeña puede terminar habiendo, por ejemplo, más jóvenes en el tratamiento que en el control. Una forma de analizar este problema es mostrar el *balance entre control y tratamiento*: mostrar que no haya diferencias significativas en las variables (observables) entre tratamiento y control.

3. El grupo de tratamiento puede estar afectado por drop-out, por el **abandono**, o más en general, por **cambios de composición**. Si hay abandono (si un grupo asignado al tratamiento no se trata) la comparación no será entre tratados y no tratados sino entre asignados y no asignados al tratamiento (se identifica un ITT y no un ATE). Este puede ser un problema será tanto más grave e incluirá potencialmente un sesgo fuerte en la medida en que la tasa de abandono sea más elevada (con un 2% de tasa de abandono no hay un sesgo muy fuerte aunque haya una selección). En la práctica debe seguirse a todos los asignados al tratamiento para tener *información de todos los que abandonan*.
4. Otra forma de abandono es la de cumplir con el protocolo pero no proveer la información después del tratamiento (attrition). Si este comportamiento fuera completamente aleatorio no constituiría un problema, pero esto es difícil de asegurar. En la práctica debe tomarse información sobre todos los individuos; si el attrition estuviera presente es importante analizar en qué medida puede ser aleatorio analizando las características de los que no contestan.

5. Puede haber **contaminación por derrame (spillover)** del grupo de control. Hay varias fuentes de contaminación. La primera es que los tratados puedan afectar a los resultados de los no tratados. (Por ejemplo, una transferencia al 50 % de la población de una ciudad puede afectar el ingreso del otro 50 %.) Si el tratamiento tiene estas características es conveniente realizar la *asignación en otra escala*: asignar ciudades de tratamiento y ciudades de control y al interior de las ciudades de tratamiento asignar aleatoriamente unidades de tratamiento y unidades no tratadas.
6. Puede haber **contaminación por compensación** del grupo de control. Este efecto surge del hecho de que los que no están en tratamiento compensan esto buscando otro tratamiento (por ejemplo, otra fuente de financiamiento, otro programa de capacitación). Este tipo de contaminación es difícil de controlar, pero potencialmente puede medirse consultando al tratado y al control si participaron de otros programas a la vez.

7. Puede haber **contaminación por otras razones** del grupo de control. En economía los individuos saben si están siendo asignados a control o a tratamiento, y esto, de por sí, puede generar un efecto. El hecho de que haya un tratado puede afectar el comportamiento de un no tratado. El haber sido asignado a control puede afectar las respuestas. Este tipo de contaminación es difícil de controlar, pero pueden intentar controlarse los detalles del protocolo para evitar lo más posible estos efectos. (Por ejemplo, brindando incentivos o compensaciones al control para que participe de las encuestas)

# Potenciales problemas en la práctica

- ▶ Analizando el balance: vemos si hay diferencias significativas en observables entre tratamiento y control

```
. ttest x2 if x3==1, by(d)
```

- ▶ Corrigiendo el balance: si hubiera diferencias se pueden incluir covariables en la regresión (sólo variables que no sean afectadas por la política)

```
. reg y d x1 x2 if x3==1
```

- ▶ Si hubiera decisión / cambio de composición: generamos datos con esta característica. Se estima un ITT (intention to treat)
- ▶ Si hubiera attrition: simulamos datos con attrition. Analizamos si este attrition es aleatorio o está correlacionado con variables vinculadas a  $y$ . Cuando hay correlación se induce un sesgo.

```
. reg attrition x1 x2
```

```
. *balance
. ttest x1 if x3==1, by(d) // las diferencias de media no son significativas (balanceado)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	8704	.5272289	.0053517	.4992867	.5167383	.5377194
1	3796	.5273973	.0081042	.4993146	.5115082	.5432863
combined	12500	.52728	.0044657	.4992752	.5185266	.5360324
diff		-.0001684	.0097116		-.0192046	.0188678

```
***
. reg y d x1 x2 if x3==1 // si no estuviera balanceado se podrian incluir controles en 1
> a estimación del efecto
```

Source	SS	df	MS	Number of obs = 12500	
Model	3003.01371	3	1001.00457	F( 3, 12496) = 1396.88	
Residual	8954.65733	12496	.716601899	Prob > F = 0.0000	
Total	11957.671	12499	.956690219	R-squared = 0.2511	
				Adj R-squared = 0.2510	
				Root MSE = .84652	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
d	.5118887	.0164657	31.09	0.000	.4796135	.5441639
x1	.2010516	.0151657	13.26	0.000	.1713245	.2307786
x2	.0895107	.0016273	55.01	0.000	.086321	.0927003
_cons	1.642441	.0161588	101.64	0.000	1.610767	1.674114

# Potenciales problemas en la práctica

En la simulación introducimos la posibilidad de decisión una vez se han asignado los grupos (abandono) y la posibilidad de que no respondan la información (attrition)

$y_{x1}$  : la variable de resultado observada si hay abandono

$dx$  : la variable de tratamiento si hay abandono (los que fueron asignados a tratamiento y decidieron tratarse tienen  $dx=1$ )

$y_x$  : una variable de resultado observada cuando hay attrition (hay algunos que no contestan y el valor pasa a missing)

```
. reg yx1 dx if x3==1 // si usáramos dx como variable de tratamiento habría un sesgo muy fuerte
```

Source	SS	df	MS	Number of obs =	12500
Model	1114.01165	1	1114.01165	F( 1, 12498)	= 1269.87
Residual	10964.0617	12498	.877265295	Prob > F	= 0.0000
				R-squared	= 0.0922
				Adj R-squared	= 0.0922
Total	12078.0733	12499	.966323171	Root MSE	= .93662

  

yx1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dx	.6880397	.0193079	35.64	0.000	.6501933 .7258861
_cons	2.310301	.0096832	238.59	0.000	2.29132 2.329282

```
. reg yx1 a1 if x3==1 // si usáramos a1 como variable de tratamiento tenemos un estimador ITT
```

Source	SS	df	MS	Number of obs =	12500
Model	601.014438	1	601.014438	F( 1, 12498)	= 654.48
Residual	11477.0589	12498	.91831164	Prob > F	= 0.0000
				R-squared	= 0.0498
				Adj R-squared	= 0.0497
Total	12078.0733	12499	.966323171	Root MSE	= .95829

  

yx1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
a1	.4768425	.0186392	25.58	0.000	.4403067 .5133782
_cons	2.338549	.0102715	227.67	0.000	2.318415 2.358683

. \* problema de attrition

```
. reg yx a1 if x3==1 // si usáramos a1 como variable de tratamiento el sesgo persiste
```

Source	SS	df	MS	Number of obs =	11848
Model	1005.47151	1	1005.47151	F( 1, 11846)	= 1142.59
Residual	10424.4059	11846	.879993743	Prob > F	= 0.0000
				R-squared	= 0.0880
				Adj R-squared	= 0.0879
Total	11429.8774	11847	.964790866	Root MSE	= .93808

  

yx	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
a1	.6597915	.0195192	33.80	0.000	.6215307 .6980523
_cons	2.338549	.010055	232.58	0.000	2.31884 2.358259



# Evaluación de impacto en general: fortalezas y debilidades

- ▶ La principal fortaleza de este esquema de evaluación de impacto es que se trata de un tratamiento exógeno este modo de identificar y medir efectos eliminando los sesgos de selección, aproximándose a una relación causal
- ▶ Las limitaciones del método se vinculan con el hecho de que usualmente no puede identificar el **mecanismo económico** (no es el objetivo de este método)
- ▶ En la evaluación de impacto es siempre conveniente preguntarse por la **validez externa**. Esto es, ¿es el resultado “generalizable”? En economía se suele extrapolar el impacto de una política a otras políticas similares; a otras condiciones; en otras etapas del ciclo económico; con otras dosis. En realidad los efectos no son lineales ni invariantes del contexto.
- ▶ La evaluación de impacto suele aplicarse en microdatos y en políticas de escala piloto. Usualmente no brinda herramientas para pensar en los potenciales cambios de **equilibrio general** si las políticas en gran escala tuvieran fuerte impacto y relevancia.
- ▶ Por estas razones puede ser conveniente complementar la evaluación de impacto con teoría o análisis económico.

# Referencias

Leer:

Angrist & Pischke, "Mostly Harmless Econometrics: An empiricist's Companion", Cap 2