

# Microeconometría II

## Lecture 1

*“...cada vez que un hombre se enfrenta con diversas alternativas, opta por una y elimina las otras.”*

*El jardín de los senderos que se bifurcan. J.L. Borges*

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Introducción

- El gobierno está pensando en un nuevo programa de capacitación dirigido a jóvenes sin educación formal. El programa capacita en algún oficio al trabajador con el objetivo de disminuir el tiempo que pasa hasta encontrar su primer empleo.
- En este caso la “vacuna” es el nuevo programa de capacitación y su efectividad debiera verse en una reducción del tiempo de desempleo.
- La variable sobre la que se mide la efectividad de la política (duración del desempleo) se denomina **variable de resultado**.
- El desafío principal de llevar a cabo una evaluación de impacto es la **identificación de una relación de causalidad entre el programa y los resultados de interés**.
- El modelo de resultados potenciales es un procedimiento que nos permite evaluar el impacto de las intervenciones sobre diferentes resultados de manera precisa.

# Introducción

- Para identificar una relación de causalidad entre la implementación del programa y el resultado del mismo, el modelo de resultados potenciales se pregunta: **cuál hubiera sido el resultado de no haberse implementado el programa?**
- Supongamos un trabajador  $u$ .
- Denotemos con  $Y_T(u)$  al tiempo que pasa (medido en meses) hasta que encuentra empleo después de haber recibido la capacitación del programa y con  $Y_C(u)$  al tiempo que pasa desempleado hasta encontrar empleo de no haber recibido la capacitación del programa.
- Los valores  $Y_T(u)$  y  $Y_C(u)$  se conocen en la literatura como **resultados potenciales**.
- Ex-ante, el trabajador  $u$  potencialmente pasará  $Y_T(u)$  ó  $Y_C(u)$  meses en desempleo.
- Ex-post, nosotros solo observamos uno de esos resultados potenciales:  $Y_T(u)$  si recibió la capacitación ó  $Y_C(u)$  si no la recibió.

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- **Modelo de Resultados Potenciales (Rubin, 1974)**
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de inferencia causal

- Concepto fundamental: **exposición potencial**. Que cada unidad de la población pueda estar expuesta potencialmente a cualquiera de las causas.
- Para empezar a formalizar nuestro análisis vamos a considerar una **población** objetivo  $U$  de unidades denotadas por  $u$ .
- Las unidades  $u$  son los objetos básicos de estudio.
- Ejemplos de unidades son: individuos, empresas, familias, parcelas de tierra, equipos de laboratorio etc.
- Una **variable** es una función que se define sobre cada unidad  $u$  de  $U$ .
- El valor de una variable para una unidad  $u$  determinada es un número asignado por alguna medición aplicada sobre  $u$ .
- Por ejemplo, para cada unidad  $u$  en  $U$  hay asociado un valor  $Y(u)$  de la variable  $Y$ .

# Modelo de inferencia causal

- Todas las distribuciones, probabilidades y valores esperados de las variables involucradas en el programa se calculan sobre  $U$ .
- Por simplicidad se asumirá que hay sólo dos niveles de tratamiento, expresados por  $T$  (el tratamiento) y  $C$  (el control) respectivamente.
- Sea  $s_T$  una variable que indica la causa a la cual fue expuesta cada  $u \in U$ , o sea  $s_T = 1$  indica exposición al tratamiento y  $s_T = 0$  indica exposición al control.
- En un estudio controlado,  $s_T$  está diseñado por el investigador. En un estudio no controlado  $s_T$  está determinado en alguna medida por factores que se encuentran más allá del control del investigador.
- En cualquiera de los dos casos un aspecto clave de la noción de causa en este modelo es que **el valor  $s_T(u)$  para cada unidad 'pudo haber sido diferente'**.

# Modelo de inferencia causal

- La interpretación de  $Y_T(u)$  e  $Y_C(u)$  para un elemento dado  $u$  es que  $Y_T(u)$  es el valor de la variable de resultado que se hubiera observado si el individuo fue expuesto a la causa  $T$  mientras que  $Y_C(u)$  es el valor de la variable de resultado que se hubiera observado **para el mismo individuo** si fue expuesto a la causa  $C$ .
- Entre los dos **resultados potenciales** correspondientes a los dos potenciales tratamientos solo **UN** resultado es observado. El otro, denominado **contrafáctica** no se observa.



# Modelo de inferencia causal

- El efecto de la causa  $T$  sobre  $u$  medido por  $Y$  y en relación a la causa  $C$  es la diferencia entre  $Y_T(u)$  e  $Y_C(u)$ . En el modelo estará representado por la siguiente diferencia:

$$Y_T(u) - Y_C(u) \quad (1)$$

- La diferencia anterior representa el efecto causal de  $T$  (con respecto a  $C$ ) sobre  $u$  (medido a través de  $Y$ ).
- La expresión (1) es la forma en que el modelo de inferencia causal expresa el concepto de causalidad más básico. Establece que  $T$  causa un efecto  $[Y_T(u) - Y_C(u)]$  sobre  $u$ .
- Esta forma de definir el **efecto causal** usando dos resultados potenciales se denomina **causalidad contrafactual**.

# Modelo de inferencia causal

## Problema Fundamental de la Inferencia Causal

**Es imposible observar los valores de  $Y_T(u)$  y  $Y_C(u)$  sobre una misma unidad y por lo tanto es imposible observar el efecto de  $T$  sobre  $u$ .**

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- **Solución Estadística del PFIC**
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de inferencia causal

- La amenaza implícita en el PFIC es que la inferencia causal es imposible. Sin embargo, hay una “solución estadística”
- La solución estadística hace uso de la población  $U$  en un sentido típicamente estadístico.
- Se define el **efecto causal promedio** ( $ATE$  - *average treatment effect*) como la diferencia entre dos valores esperados,  $E(Y_T)$  y  $E(Y_C)$ .

$$ATE = E[Y_T(u) - Y_C(u)] = E(Y_T) - E(Y_C) \quad (2)$$

# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	$T$	$Y_T(1)$	$Y_C(1)$	$Y_T(1)$
2	$T$	$Y_T(2)$	$Y_C(2)$	$Y_T(2)$
3	$C$	$Y_T(3)$	$Y_C(3)$	$Y_C(3)$
4	$T$	$Y_T(4)$	$Y_C(4)$	$Y_T(4)$
5	$C$	$Y_T(5)$	$Y_C(5)$	$Y_C(5)$
6	$C$	$Y_T(6)$	$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N$	$T$	$Y_T(N)$	$Y_C(N)$	$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuando  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T | s_T = 1) - E(Y_C | s_T = 0)$

# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	1	$Y_T(1)$		$Y_T(1)$
2	1	$Y_T(2)$		$Y_T(2)$
3	0		$Y_C(3)$	$Y_C(3)$
4	1	$Y_T(4)$		$Y_T(4)$
5	0		$Y_C(5)$	$Y_C(5)$
6	0		$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	$Y_T(N)$		$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuando  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T | s_T = 1) - E(Y_C | s_T = 0)$

# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	1	$Y_T(1)$	$Y_C(1)$	$Y_T(1)$
2	1	$Y_T(2)$	$Y_C(2)$	$Y_T(2)$
3	0	$Y_T(3)$	$Y_C(3)$	$Y_C(3)$
4	1	$Y_T(4)$	$Y_C(4)$	$Y_T(4)$
5	0	$Y_T(5)$	$Y_C(5)$	$Y_C(5)$
6	0	$Y_T(6)$	$Y_C(6)$	$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	$Y_T(N)$	$Y_C(N)$	$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuándo  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T | s_T = 1) - E(Y_C | s_T = 0)$

# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	

- Cuando  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T | s_T = 1) - E(Y_C | s_T = 0)$



# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	$E(Y_T   s_T = 1)$

- Cuando  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T | s_T = 1) - E(Y_C | s_T = 0)$

# Tabla de resultados potenciales

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$	$y_u$
1	1			$Y_T(1)$
2	1			$Y_T(2)$
3	0			$Y_C(3)$
4	1			$Y_T(4)$
5	0			$Y_C(5)$
6	0			$Y_C(6)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1			$Y_T(N)$
		$E(Y_T)$	$E(Y_C)$	$E(Y_C s_T = 0)$

- Cuándo  $E(Y_T) - E(Y_C) \stackrel{?}{=} E(Y_T|s_T = 1) - E(Y_C|s_T = 0)$

# Modelo de inferencia causal

- Se considera que la población  $U$  de individuos es "grande" y los datos observados para cada individuo  $u$  son los valores de  $s_T(u)$  e  $y_u$ .
- La información que obtenemos se refiere a:

$$E(y_u | s_T = 1) = E(Y_T | s_T = 1), \quad E(y_u | s_T = 0) = E(Y_C | s_T = 0)$$

- Los valores observados de la variable de resultado se pueden describir a través de:

$$\begin{aligned} y_u &= s_T \times Y_T(u) + (1 - s_T) \times Y_C(u), \quad u = 1, \dots, N \\ &= Y_C(u) + s_T \times [Y_T(u) - Y_C(u)], \quad u = 1, \dots, N \end{aligned} \quad (3)$$

- Esta ecuación se conoce como "switching equation".
- Es esencial reconocer que  $E(Y_T)$  y  $E(Y_T | s_T = 1)$  no son equivalentes (idem para la causa  $C$ ).

# Independencia

- Cuando los elementos de la población son asignados aleatoriamente a la causa  $C$  o  $T$  es posible verificar que la causa a la que está expuesta cada unidad  $u$  de la población será estadísticamente independiente de cualquier otra variable, incluyendo a  $Y_T$  e  $Y_C$ .
- Si el procedimiento de aleatorización se realiza correctamente entonces  $s_T$  es independiente de  $Y_T$  e  $Y_C$  y otras variables en  $U$ .

# Independencia

- Si el supuesto de independencia es válido entonces

$$E(Y_T) = E(Y_T | s_T = 1)$$

y

$$E(Y_C) = E(Y_C | s_T = 0)$$

- Por lo tanto el efecto causal promedio se obtiene como

$$ATE = E[y_u | s_T(u) = 1] - E[y_u | s_T(u) = 0]$$

Si la aleatorización es posible (supuesto de independencia es válido), siempre se puede estimar el efecto causal promedio como una diferencia de medias.

- La expresión anterior revela que se puede utilizar la información de distintos individuos para conocer  $ATE$ .

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- **El Director de Capacitación perfecto**
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

## El Director de Capacitación perfecto: ejemplo

- Los datos hipotéticos de la tabla muestran los resultados potenciales de dos programas de entrenamiento para el empleo:
  - $Y_T$ : meses hasta conseguir el primer empleo después de recibir el nuevo programa de entrenamiento.
  - $Y_C$ : meses hasta conseguir el primer empleo después de recibir el viejo programa de entrenamiento.

$u$	$Y_T(u)$	$Y_C(u)$
1	10	14
2	11	9
3	14	10
4	12	9
5	6	7
6	9	10
$E(Y)$	10.33	9.83

- ATE** =  $E(Y_T) - E(Y_C) = 10.33 - 9.83 = 0.5$ , en promedio, el nuevo programa de capacitación no reduce el tiempo de desempleo.

# El director de capacitación perfecto: ejemplo

- El **director de capacitación perfecto** elige para cada trabajador el mejor tratamiento (el que lo deja desempleado el menor tiempo posible).
- Qué observaríamos?

$u$	$s_T(u)$	$Y_T(u)$	$Y_C(u)$
1	1	10	?
2	0	?	9
3	0	?	10
4	0	?	9
5	1	6	?
6	1	9	?
$E(Y S)$		8.33	9.33

- ATE** =  $E(Y_T|s_T = 1) - E(Y_C|s_T = 0) = 8.33 - 9.33 = -1$ , en promedio, el nuevo programa de capacitación reduce el tiempo de desempleo en un mes!.



# El Director de Capacitación perfecto: ejemplo

- En el ejemplo del Director de Capacitación perfecto, el entrenamiento que cada unidad recibe depende del resultado potencial de esa unidad.
- El mecanismo de asignación,  $s_T(u)$ , no es independiente de los resultados potenciales.
- Esto provoca que el estimador del ATE sea sesgado.
- Sobre la base de los resultados observados, concluiríamos erróneamente que el nuevo programa de entrenamiento funciona.
- Como conseguir que el mecanismo de asignación sea independiente de los resultados potenciales? **Asigne aleatoriamente.**

# El Director de Capacitación perfecto: ejemplo

- El verdadero ATE de la primera tabla es un parámetro (está calculado con los resultados potenciales).
- El ATE de la segunda tabla es una estimación (está calculado con los datos observados).
- Necesitamos un estimador que sea INSESGADO: si la asignación se repitiera una y otra vez, el promedio de los estimadores debiera ser igual al parámetro.

# El Director de Capacitación perfecto: ejemplo

- En el ejemplo del Director de Capacitación perfecto hay 20 formas diferentes de asignar 3 unidades a la nueva capacitación y 3 a la vieja.

	#1 $\Rightarrow$	111000	ATE estimado = 3
	#2 $\Rightarrow$	110100	ATE estimado = 2
	...		
Director perfecto:	#7 $\Rightarrow$	100011	ATE estimado = -1
	...		
	#20 $\Rightarrow$	000111	ATE estimado = -2

- El promedio de los 20 ATE observados es 0.5 y coincide con el ATE verdadero.
- El Director de Capacitación perfecto** siempre elige el #7.
- El mecanismo de asignación independiente** selecciona el # aleatoriamente.

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- **Modelo de Regresión Lineal Simple**
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de Regresión Simple

- Considere nuevamente la *switching equation* (3)

$$y_u = Y_C(u) + s_T \times [Y_T(u) - Y_C(u)]$$

- Escribamos los resultados potenciales como

$$Y_C(u) = E[Y_C(u)] + \epsilon_C(u) \quad (4)$$

$$Y_T(u) = E[Y_T(u)] + \epsilon_T(u) \quad (5)$$

piense que siempre podemos escribir a una variable aleatoria como su esperanza matemática más una variable de media cero.

- Reemplazando estos resultados potenciales en la ecuación (3) tenemos:

$$\begin{aligned} y_u &= E[Y_C(u)] + s_T(u) \times \{E[Y_T(u)] - E[Y_C(u)]\} + \epsilon_u \\ &= \underbrace{E[Y_C(u)]}_{\beta_0} + s_T(u) \times \underbrace{\{E[Y_T(u)] - E[Y_C(u)]\}}_{\beta_1} + \epsilon_u \end{aligned} \quad (6)$$

donde  $\epsilon_u = \epsilon_C(u) + s_T(u) \times \{\epsilon_T(u) - \epsilon_C(u)\}$ .

# Modelo de Regresión Simple

- La ecuación (6) recibe el nombre de **Modelo de Regresión Lineal Simple (MRLS)**.  $y_u$  se conoce como variable dependiente y  $s_T(u)$  como variable explicativa.
- $\beta_0$  y  $\beta_1$  son los parámetros del modelo.
- Note que  $\beta_1 = E[Y_T(u)] - E[Y_C(u)]$  es el ATE.
- Recuerde que, si observáramos todos los datos de la población, lo único que podríamos calcular es:  $E[y_u | s_T(u) = 1] - E[y_u | s_T(u) = 0]$ , entonces

$$E[y_u | s_T = 1] = \beta_0 + \beta_1 + E[\epsilon_u | s_T = 1]$$

y

$$E[y_u | s_T = 0] = \beta_0 + E[\epsilon_u | s_T = 0].$$

# Modelo de Regresión Simple

- Por lo tanto, de las dos ecuaciones anteriores

$$\begin{aligned} E[y_u|s_T = 1] - E[y_u|s_T = 0] &= \beta_1 + E[\epsilon_u|s_T = 1] - E[\epsilon_u|s_T = 0] \\ &= \underbrace{\phantom{\beta_1}}_{\text{ATE}} + \underbrace{\phantom{E[\epsilon_u|s_T = 1] - E[\epsilon_u|s_T = 0]}}_{\text{sesgo}} \end{aligned}$$

- **Si se cumple el supuesto de independencia entre  $s_T$  y los resultados potenciales el sesgo desaparece y  $E[y_u|s_T = 1] - E[y_u|s_T = 0] = \beta_1 = ATE$ .**
- Para ver esto matemáticamente note que:

$$\begin{aligned} \text{sesgo} &= E[\epsilon_u|s_T = 1] - E[\epsilon_u|s_T = 0] = E[\epsilon_T(u)|s_T = 1] - E[\epsilon_C(u)|s_T = 0] \\ &= E\{[Y_T(u) - E(Y_T(u))]|s_T = 1\} - E\{[Y_C(u) - E(Y_C(u))]|s_T = 0\} \\ &= \{E[Y_T(u)|s_T = 1] - E[Y_T(u)]\} - \{E[Y_C(u)|s_T = 0] - E[Y_C(u)]\} \\ &= 0 \end{aligned}$$

# Modelo de Regresión Simple

- Entonces, el parámetro  $\beta_1$  es igual a la diferencia de esperanzas condicionales entre grupos y se puede escribir como:

$$\begin{aligned}\beta_1 &= E[y_u | s_T = 1] - E[y_u | s_T = 0] \\ &= E(Y_T - Y_C) = E(y_1 - y_0) \\ &= \frac{COV(y, s)}{VAR(s)}.\end{aligned}$$

- Para que esta igualdad se cumpla es fundamental el **supuesto de ignorabilidad del tratamiento (asignación aleatoria)**.



# Modelo de Regresión Simple

$$\begin{aligned}\frac{COV(y, s)}{VAR(s)} &= \frac{E[(sy_1 + (1 - s)y_0)s] - E[(sy_1 + (1 - s)y_0)]E[s]}{p(1 - p)} \\&= \frac{E[s^2y_1] + E[s(1 - s)y_0] - E[sy_1]p - E[(1 - s)y_0]p}{p(1 - p)} \\&= \frac{E[s^2y_1](1 - p) - E[(1 - s)y_0]p}{p(1 - p)} \\&= \frac{E[E[sy_1|s]](1 - p) - E[E[(1 - s)y_0|s]]p}{p(1 - p)} \\&= \frac{E[y_1|s = 1]P(s = 1)(1 - p) - E[y_0|s = 0]P(s = 0)p}{p(1 - p)} \\&= \frac{\{E[y_1|s = 1] - E[y_0|s = 0]\}p(1 - p)}{p(1 - p)} \\&= \{E[y_1|s = 1] - E[y_0|s = 0]\} = E(y_1 - y_0)\end{aligned}$$

bajo asignación aleatoria del tratamiento.

# Modelo de Regresión Simple

- En la práctica, el coeficiente poblacional  $\beta_1$  se estima utilizando el método de mínimos cuadrados clásicos (MCC).
- $\hat{\beta}_1 = \frac{\widehat{COV}(y,s)}{\widehat{VAR}(s)} = \bar{y}_1 - \bar{y}_0$
- **Contraste de hipótesis:**  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 > 0$  ( $< 0 \neq 0$ )
- **Estadístico de contraste:**  $t = \frac{\hat{\beta}_1}{std(\hat{\beta}_1)}$
- **Regla de decisión:** valor-p del estadístico de contraste menor a nivel de significación, implica Rechazo de  $H_0$ .

# Modelo de Regresión Simple

- En Stata el comando para estimar por MCC es: `reg`

`reg` *variable dependiente* *variable explicativa*

# Modelo de Regresión Lineal

- En la práctica es muy difícil realizar una aleatorización de las unidades poblacionales.
- Piense en el ejemplo del programa de capacitación. Es muy poco probable que el gobierno designe aleatoriamente a aquellos a quienes les brindará la capacitación.
- En general siempre habrá cierta autoselección en el tratamiento.
- Si esto ocurre, entonces habrá alguna correlación entre  $s_T$  e  $y_u$  y el supuesto de independencia no se cumplirá.
- Aún en estos casos es posible recuperar el impacto del programa si se controla por las variables que inducen la correlación entre  $s_T$  e  $y_u$ .
- Para que el modelo de regresión tenga una interpretación causal necesitamos cambiar el supuesto de independencia.
- **Supuesto de Independencia Condicional o Selección sobre Observables:** llamemos  $w_u$  al vector de variables observables. Entonces

$$\{Y_T(u), Y_C(u)\} \perp\!\!\!\perp s_T(u) | w_u \quad (7)$$

# Modelo de Regresión Lineal

- Piense en el ejemplo del programa de capacitación laboral donde  $s_T(u)$  indica si el trabajador  $u$  participa del programa e  $y_u$  es el tiempo hasta encontrar empleo.
- Los elementos en  $w_u$  podrían ser: educación, edad y experiencia en el mercado de trabajo.
- Suponga que los trabajadores que están más interesados en participar del programa son aquellos más jóvenes, con menos educación y menos experiencia.
- Entonces como la educación, edad y experiencia ayudan a explicar  $Y_T(u)$  e  $Y_C(u)$  la asignación aleatoria no se mantiene.
- Sin embargo, si agrupamos a los trabajadores por edad, educación y experiencia es posible que la asignación se vuelva independiente.
- Por ejemplo, considere el grupo de trabajadores que tiene 25 años de edad, con 12 años de escolarización y 4 años de experiencia. **Lo que requiere la independencia condicional es que dentro de este grupo la asignación funcione como aleatoria.** Esto funciona si los grupos son “grandes”.

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- **Modelo de Regresión Lineal Múltiple**

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de Regresión Lineal

- Lo que dice el supuesto de independencia condicional es que condicionando sobre las características  $w_u$  el sesgo de selección desaparece.
- En otras palabras, la asignación del tratamiento es tan buena como si se hubiera hecho un experimento aleatorizado entre unidades tratadas y no tratadas que resultan comparables en términos de sus características observables.
- Descomponiendo la parte aleatoria de los resultados potenciales  $\epsilon_C(u)$  y  $\epsilon_T(u)$  en las ecuaciones (4) y (5), en una parte lineal que depende de las características observables  $w_u$  y un término de error con media cero tenemos:

$$Y_C(u) = E[Y_C(u)] + [w_u - E(w_u)]\gamma + \eta_C(u) \quad (8)$$

$$Y_T(u) = E[Y_T(u)] + [w_u - E(w_u)]\gamma + \eta_T(u) \quad (9)$$

donde las características observables están escritas en desvíos con respecto a sus medias simplemente por convención y  $\gamma$  es un vector de parámetros.

- Reemplazando estos resultados potenciales en la ecuación (3) tenemos:

$$y_u = \beta_0 + \beta_1 s_T(u) + [w_u - E(w_u)]\gamma + v_u \quad (10)$$

donde  $v_u = \eta_C(u) + s_T(u)[\eta_T(u) - \eta_C(u)]$

# Modelo de Regresión Lineal

- Entonces, por el supuesto de independencia condicional:

$$\begin{aligned} E[y_u | w_u, s_T] = E[y_u | w_u] &= \beta_0 + \beta_1 s_T(u) + [w_u - E(w_u)]\gamma + E[v_u | w_u, s_T] \\ &= \beta_0 + \beta_1 s_T(u) + [w_u - E(w_u)]\gamma \end{aligned} \quad (11)$$

- Note que podemos relajar el supuesto de independencia condicional a **independencia condicional en media:**

$$E[y_u | w_u, s_T] = E[y_u | w_u] \quad (12)$$

- El coeficiente  $\beta_1$  sigue siendo el efecto promedio del tratamiento,

$$\begin{aligned} E[y_u | w_u, s_T = 1] - E[y_u | w_u, s_T = 0] &= (\beta_0 + \beta_1 + [w_u - E(w_u)]\gamma) - (\beta_0 + [w_u - E(w_u)]\gamma) \\ &= \beta_1 \end{aligned} \quad (13)$$



# Modelo de Regresión Lineal

- **Remark 1:** es claro de la ecuación (10) que si controlamos por  $w_u$  en lugar de por  $[w_u - E(w_u)]$  lo único que se modifica es la ordenada al origen.
- **Remark 2:** el primer supuesto implícito en este desarrollo es que el efecto de las características observables es el mismo,  $\gamma$ , en ambos grupos, tratamiento y control.
- **Remark 3:** el segundo supuesto implícito es que controlando por las características observables también se controla por las no observables.

# Modelo de Regresión Lineal

- **Remark 2.**

- Podemos permitir efectos diferentes de las características observables escribiendo:

$$Y_C(u) = E[Y_C(u)] + [w_u - E(w_u)]\gamma_0 + \eta_C(u) \quad (14)$$

$$Y_T(u) = E[Y_T(u)] + [w_u - E(w_u)]\gamma_1 + \eta_T(u) \quad (15)$$

- Reemplazando estos resultados potenciales en la ecuación (3) tenemos:

$$y_u = \beta_0 + \beta_1 s_T(u) + [w_u - E(w_u)]\gamma_0 + s_T(u)[w_u - E(w_u)]\delta + v_u \quad (16)$$

donde  $\delta = \gamma_1 - \gamma_0$  y  $v_u = \eta_C(u) + s_T(u)[\eta_T(u) - \eta_C(u)]$

# Modelo de Regresión Lineal

- Ahora

$$E[y_u|w_u, s_T = 1] = \beta_0 + \beta_1 + [w_u - E(w_u)]\gamma_0 + [w_u - E(w_u)]\delta$$

$$E[y_u|w_u, s_T = 0] = \beta_0 + [w_u - E(w_u)]\gamma_0$$

$$E[y_u|w_u, s_T = 1] - E[y_u|w_u, s_T = 0] = \beta_1 + [w_u - E(w_u)]\delta$$

- Esta última ecuación determina el denominado **efecto tratamiento promedio condicional**,  $ATE(w)$ .
- Remark 4:** de esta última ecuación queda claro que:

$$\begin{aligned} E_w \{ E[y_u|w_u, s_T = 1] - E[y_u|w_u, s_T = 0] \} &= E[y_u|s_T = 1] - E[y_u|s_T = 0] \\ &= E[y_1] - E[y_0] = \beta_1 \end{aligned}$$

por la ley de expectativas iteradas y el supuesto de independencia condicional.

- Remark 5:** para calcular  $E_w$  necesitamos observar unidades en el grupo de tratamiento y de control con las mismas características observables.

# Modelo de Regresión Lineal

- Esta última condición se denomina en la literatura de causalidad como **soporte común** (“superposición” u “overlapping”).
- Matemáticamente el **supuesto de superposición** es,

$$0 < \Pr(s_T(u) = 1 | w_u) < 1 \quad (17)$$

La probabilidad condicional de recibir el tratamiento también se denomina “propensity score”.

- El supuesto de superposición significa que para cualquier configuración de las características observables,  $w_u$ , existe una probabilidad positiva de observar unidades tanto en el grupo de tratamiento como en el de control.

# Modelo de Regresión Lineal

- El impacto promedio del tratamiento se puede recuperar también desde las ecuaciones (14) y (15) por separado:

$$Y_C(u) = E[Y_C(u)] + [w_u - E(w_u)]\gamma_0 + \eta_C(u) = \alpha_0 + w_u\gamma_0 + \eta_C(u) \quad (18)$$

$$Y_T(u) = E[Y_T(u)] + [w_u - E(w_u)]\gamma_1 + \eta_T(u) = \alpha_1 + w_u\gamma_1 + \eta_T(u) \quad (19)$$

donde  $\alpha_0 = E[Y_C(u)] - E(w_u)\gamma_0$  y  $\alpha_1 = E[Y_T(u)] - E(w_u)\gamma_1$

- Tomando esperanzas condicionales tenemos

$$E[y_u | w_u, s_T = 0] = \alpha_0 + w_u\gamma_0 \quad (20)$$

$$E[y_u | w_u, s_T = 1] = \alpha_1 + w_u\gamma_1 \quad (21)$$

$$E[y_u | w_u, s_T = 1] - E[y_u | w_u, s_T = 0] = \alpha_1 - \alpha_0 + w_u(\gamma_1 - \gamma_0) \quad (22)$$

$$E_w \{E[y_u | w_u, s_T = 1] - E[y_u | w_u, s_T = 0]\} = \alpha_1 - \alpha_0 + E_w[w_u](\gamma_1 - \gamma_0) = \beta_1$$

# Modelo de Regresión Lineal

- En la práctica cualquiera de las ecuaciones presentadas: (10), (16), (18) y (19) se deben estimar usando el método de MCC.
- En Stata el comando para estimar por MCC es: `reg`

`reg` *variable dependiente* *variables explicativas*

- Ejemplo: programa de capacitación laboral `jobtraining.dta`

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- **Modelo de inferencia causal: supuestos**
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de inferencia causal: Supuestos

- **Supuesto 0:** Para cada unidad  $u$ , tenemos:  $(Y_T(u), Y_C(u), w_u)$ , donde  $w_u$  son características observables de  $u$ . La muestra de datos  $(Y_T(u), Y_C(u), w_u)_{u=1}^n$  es i.i.d. seleccionada desde una población  $P_o$ .
- **Supuesto 1 (Independencia Condicional en Media):**  
 $\mathbb{E}[Y_T(u) \mid w_u, s_T(u)] = \mathbb{E}[Y_T(u) \mid w_u]$  y  $\mathbb{E}[Y_C(u) \mid w_u, s_T(u)] = \mathbb{E}[Y_C(u) \mid w_u]$ .
- **Supuesto 2 (Superposición u “overlap”):**  
 $0 < \Pr(s_T(u) = 1 \mid w_u) < 1$ .



# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- **Modelo de inferencia causal: identificación**

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de inferencia causal: identificación

- Dados los supuestos 1 y 2, se puede identificar el ATE.

$$\tau_{ATE}(w) = \mu_1(w_u) - \mu_0(w_u), \forall w_u$$

donde  $\mu_s(w_u) = E[y_u | s_T = s, W = w_u]$

- Dado el **Supuesto 1**, las siguientes igualdades se cumplen ( $s = 0, 1$ ):

$$\mu_s(w_u) = E[y_u(s) | W = w_u] = E[y_u(s) | s_T(u) = s, W = w_u] = E[y_u | s_T(u) = s, W = w_u]$$

y  $\mu_s(w_u)$  está identificado.

- Entonces, se puede estimar el efecto tratamiento promedio,  $\tau$ , estimado primero el ATE para una subpoblación con covariables  $W = w_u$ ,

$$\begin{aligned}\tau(w) &\equiv E[Y_T(u) - Y_C(u) | W = w_u] = E[Y_T(u) | W = w_u] - E[Y_C(u) | W = w_u] \\ &= E[Y_T(u) | W = w_u, s_T(u) = 1] - E[Y_C(u) | W = w_u, s_T(u) = 0] \\ &= E[y_u | W = w_u, s_T(u) = 1] - E[y_u | W = w_u, s_T(u) = 0]\end{aligned}$$

# Modelo de inferencia causal: identificación

- Note que para que esta estimación sea posible, necesitamos estimar  $\mathbb{E}[y_u \mid W = w_u, s_T(u) = s]$  para todos los valores de  $w_u$  y  $s$  en el soporte de estas variables.
- Aquí es donde entra el **Supuesto 2**.
- Si el **Supuesto 2** no se cumple para  $W = w_u$ , no es posible estimar las dos esperanzas matemáticas en esos puntos,  $\mathbb{E}[y_u \mid W = w_u, s_T(u) = 1]$  y  $\mathbb{E}[y_u \mid W = w_u, s_T(u) = 0]$ , porque en esos valores de  $w_u$  habría solo observaciones del tratamiento o del control.
- En otras palabras, los supuestos 1 y 2 implican que dos unidades de diferentes grupos pero con las mismas variables observables deberían tener el mismo resultado potencial esperado:  $E[Y(s) \mid s_T(u) = 1, W] = E[Y(s) \mid s_T(u) = 0, W]$ ,  $s = 1, 0$ .

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Efecto promedio del tratamiento sobre los tratados

- Una segunda medida del impacto de un programa es el **efecto promedio del tratamiento sobre los tratados (Average Treatment Effect on the Treated, ATT)**:

$$ATT = \mathbb{E}(y_1 - y_0 | s_T = 1) \quad (23)$$

- El ATT es el efecto promedio para aquellos que participaron del programa.
- Note que de la ecuación (23) se desprende que si se cumple el supuesto de independencia (por ejemplo en un experimento aleatorizado):  $ATE = ATT$
- Si hay alguna autoselección en el tratamiento el ATT puede ser recuperado haciendo un supuesto de independencia más débil que el que necesita el ATE.

# Efecto promedio del tratamiento sobre los tratados

- Si asumimos que  $s_T$  es independiente de  $y_0$  sin imponer ninguna restricción sobre la relación entre  $s_T$  e  $y_1$  tenemos

$$\begin{aligned} E(y_u | s_T = 1) - E(y_u | s_T = 0) &= E(y_0 | s_T = 1) - E(y_0 | s_T = 0) + E(y_1 - y_0 | s_T = 1) \\ &= [E(y_0 | s_T = 1) - E(y_0 | s_T = 0)] + ATT \end{aligned} \quad (24)$$

- Si  $y_0$  es independiente de  $s_T$  el primer término del lado derecho de (24) desaparece y recuperamos el ATT.
- Igualmente, este supuesto (aleatorización entre los que no participan del programa o política) es bastante fuerte y en la práctica se recurre a las mismas estrategias que utilizamos para calcular el *ATE*.

# Efecto promedio del tratamiento sobre los tratados

- **Supuesto ATT 1 (Independencia Condicional en Media):**

$$\mathbb{E}[Y_C(u) \mid w_u, s_T(u)] = \mathbb{E}[Y_C(u) \mid w_u].$$

- **Supuesto ATT 2 (Superposición u “overlap”):**

$$Pr(s_T(u) = 1 \mid w_u) < 1.$$

# Efecto promedio del tratamiento sobre los tratados

- Recordando que  $y_u = y_0 + s_T(y_1 - y_0)$  podemos escribir

$$\begin{aligned} E(y_u \mid w_u, s_T = 1) - E(y_u \mid w_u, s_T = 0) &= E(y_0 \mid w_u, s_T = 1) - E(y_0 \mid w_u, s_T = 0) \\ &+ E(y_1 - y_0 \mid w_u, s_T = 1) \\ &= [E(y_0 \mid w_u, s_T = 1) - E(y_0 \mid w_u, s_T = 0)] \\ &+ ATT(w) \end{aligned} \tag{25}$$

- Por el **supuesto ATT 1** el primer término del lado derecho de la ecuación (25) desaparece y la diferencia en esperanzas condicionales que se pueden calcular identifica  $ATT(w)$ .



# Efecto promedio del tratamiento sobre los tratados

- El efecto promedio sobre los tratados es:

$$E_w \{E(y_u | w_u, s_T = 1) - E(y_u | w_u, s_T = 0)\} = E_w \{E(y_1 - y_0 | w_u, s_T = 1)\} \quad (26)$$

- $E(y_u | w_u, s_T = 1)$  se puede calcular desde la subpoblación de unidades tratadas usando solo los valores de  $w_u$  en ese grupo.
- Por otro lado  $E(y_u | w_u, s_T = 0)$  se tiene que calcular para la subpoblación de unidades tratadas. Aquí es donde entra el **supuesto ATT 2**.
- Si hay valores de  $w_u$  solo para unidades del tratamiento no podríamos calcular  $E(y_u | w_u, s_T = 0)$  para esos valores porque no habría observaciones en el control.
- El **supuesto ATT 2:  $Pr(s_T(u) = 1 | w_u) < 1$**  descarta esta posibilidad.

# Efecto promedio del tratamiento sobre los tratados

- Estimación
- Las dos medidas de impacto,  $ATE$  y  $ATT$  se pueden estimar por  $MCC$ .
- En la práctica estimamos  $E[y \mid \mathbf{w}, s_T = 1]$  con el plano de regresión muestral,  $\hat{m}_1(\mathbf{w}, \hat{\delta}_1)$ , con las observaciones del tratamiento y estimamos  $E[y \mid \mathbf{w}, s_T = 0]$  con el plano de regresión muestral,  $\hat{m}_0(\mathbf{w}, \hat{\delta}_0)$ , con las observaciones del control.
- Un estimador consistente del  $ATE$  es:

$$ATE = \frac{1}{N} \sum_{i=1}^N \left[ \hat{m}_1(\mathbf{w}_i, \hat{\delta}_1) - \hat{m}_0(\mathbf{w}_i, \hat{\delta}_0) \right]$$

- Un estimador consistente del  $ATT$  es:

$$ATT = \frac{1}{\sum_{i=1}^N s_T(i)} \sum_{i=1}^N s_T(i) \times \left[ \hat{m}_1(\mathbf{w}_i, \hat{\delta}_1) - \hat{m}_0(\mathbf{w}_i, \hat{\delta}_0) \right]$$

# Efecto promedio del tratamiento

- **IMPORTANTE:** note que en cualquiera de estos casos lo que estamos haciendo es estimar  $E[Y(0) \mid S = 1, W = w]$  con  $\hat{m}_0(\mathbf{w}_i, \hat{\delta}_0)$ , y  $E[Y(1) \mid S = 0, W = w]$  con  $\hat{m}_1(\mathbf{w}_i, \hat{\delta}_1)$ .
- Esto es, para estimar el resultado contrafáctico promedio usamos el resultado promedio observado en el otro grupo.
- **Esto solo funciona si las covariables  $w$  son las mismas!!**
- Es decir, la regresión sobre los “controles”,  $\hat{m}_0(\mathbf{w}_i, \hat{\delta}_0)$ , se usa para predecir los resultados potenciales no observados de los “tratados”.
- Lo que uno querría es que esta predicción de los controles,  $\hat{m}_0(\mathbf{w}_i, \hat{\delta}_0)$ , se haga en usando las covariables de los tratados,  $W_i^T$ .
- Pero si las covariables en los dos grupos son diferentes la predicción puede ser mala.

# Efecto promedio del tratamiento

- Como ejemplo considere

$$\begin{aligned}\hat{m}_0(\mathbf{w}_i, \hat{\delta}_0) &= \hat{\pi}_0 + \hat{\pi}_2 W_i^C \\ &= \hat{\pi}_0 + \hat{\pi}_2 W_i^C + \hat{\pi}_2 W_i^T - \hat{\pi}_2 W_i^T \\ &= \hat{Y}_i^C + \hat{\pi}_2 (W_i^C - W_i^T)\end{aligned}$$

donde  $\hat{Y}_i^C = \hat{\pi}_0 + \hat{\pi}_2 W_i^T$

- Si  $W_i^T$  y las covariables en el control  $W_i^C$  son parecidas, la especificación precisa de la función de regresión no importa mucho para la predicción.
- Sin embargo si estas covariables son muy diferentes entre los grupos, la predicción basada en la función de regresión puede ser sensitiva a cambios en la especificación.
- Para tomar en cuenta este punto la literatura ha avanzado en **estimadores alternativos basados en el emparejamiento de las covariables**.

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto



- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Modelo de Regresión Lineal

- **Remark 3.**
- Para que controlar por los factores observables también controle por los no observables necesitamos que los grupos de tratamiento y control sean lo suficientemente grandes.
- Con pocas observaciones en los dos grupos es menos probable que esto suceda.
- ¿Cómo se determina el tamaño de los grupos?

	<b>Prince Charles</b> <ul style="list-style-type: none"><li>■ Male</li><li>■ Born in 1948</li><li>■ Raised in the UK</li><li>■ Married twice</li><li>■ Lives in a castle</li><li>■ Wealthy &amp; famous</li></ul>
	<small>20102MEME</small> <b>Ozzy Osbourne</b> <ul style="list-style-type: none"><li>■ Male</li><li>■ Born in 1948</li><li>■ Raised in the UK</li><li>■ Married twice</li><li>■ Lives in a castle</li><li>■ Wealthy &amp; famous</li></ul>

# Agenda

## 1 Modelo de Inferencia Causal

- ¿Qué es la Inferencia Causal?
- Modelo de Resultados Potenciales (Rubin, 1974)
- Solución Estadística del PFIC
- El Director de Capacitación perfecto
- Modelo de Regresión Lineal Simple
- Modelo de Regresión Lineal Múltiple

## 2 Modelo de inferencia causal: Resumen

- Modelo de inferencia causal: supuestos
- Modelo de inferencia causal: identificación

## 3 Otras Medidas de Impacto

- Efecto promedio del tratamiento sobre los tratados

## 4 Determinación del Tamaño Muestral

- Tamaño de los grupos de tratamiento y control
- Programa *Entrenamiento para el Empleo (EPEM)*

# Programa *Entrenamiento para el Empleo (EPEM)*

- El EPEM está dirigido a proporcionar entrenamiento a desempleados y subempleados jóvenes con el fin de que obtengan las calificaciones necesarias para poder desempeñarse en puestos vacantes de algunas empresas y elevar así sus posibilidades de inserción laboral.
- La población objetivo del EPEM está conformada por individuos desempleados y subempleados de 18 a 29 años, que se encuentren en la búsqueda activa de empleo, cuenten con al menos tres años de escolaridad, y cumplan con el perfil básico que necesita la empresa demandante de personal.
- Los destinatarios del programa deben estar interesados en recibir entrenamiento para adquirir conocimientos y desarrollar habilidades para desempeñarse en un puesto de trabajo por un período aproximado de tres meses, el cual idealmente se les ofrece al concluir el entrenamiento en función de su desempeño.



# Programa *Entrenamiento para el Empleo (EPEM)*

- La evaluación de impacto tiene como objetivo principal proporcionar evidencia en cuanto a si el entrenamiento para los jóvenes desempleados es un método factible para mejorar las oportunidades de empleo de los beneficiarios del EPEM.
- En particular, la evaluación trata de determinar **cuál es la mejora que perciben los individuos que participan en el programa**, en términos de algunas variables de interés (denominadas variables de resultado): **probabilidad de conseguir empleo y rapidez para hacerlo, ingreso laboral real, y probabilidad de estar empleado con beneficios sociales.**
- **No se cuenta con datos experimentales.**
- No se cuenta con la información de una línea de base generada por el diseño del programa que se pretende evaluar y solo se cuenta con datos recolectados después de la implementación del mismo.

# EPEM: diagnóstico de la información disponible

- Para realizar la evaluación de impacto se cuenta con una muestra aleatoria de 1151 jóvenes que ingresaron al programa EPEM durante 2008 y 2009 (**grupo de tratamiento**) y una muestra independiente de 792 jóvenes seleccionados aleatoriamente del Servicio Público Privado de Intermediación Laboral (SPPIIL) (**grupo de control**).
- El SPPIIL es un sistema de vinculación entre la oferta y demanda de empleo, que funciona como un servicio de intermediación laboral articulado a través de una Bolsa Electrónica de Trabajo (BET).
- Se supone que los individuos registrados en la Bolsa de Trabajo deberían tener un perfil adecuado para participar en el EPEM y por lo tanto ser potenciales beneficiarios del programa.
- Debido a que estos jóvenes no participaron en el entrenamiento, constituyen potencialmente una población con características comparables con aquellos jóvenes que ingresaron al EPEM durante 2008 y 2009.

# EPEM: determinación del tamaño muestral

- El primer paso en una evaluación de impacto es la determinación del tamaño muestral con el que realizarla.
- Los **cálculos de potencia** sirven para determinar el tamaño muestral requerido para que la evaluación estime en forma precisa el impacto del programa.
- Los ejercicios de potencia permiten determinar el tamaño muestral mínimo necesario para implementar la evaluación de impacto y responder convincentemente la pregunta de política relevante.
- Estos ejercicios indican cuál sería el tamaño muestral más pequeño con el que es posible medir el impacto de un programa.

# EPEM: determinación del tamaño muestral

- Para la determinación del tamaño muestral hay que definir el nivel de significación,  $\alpha\%$ , con el que se quiere trabajar (la probabilidad de rechazar una hipótesis nula verdadera) y el nivel de potencia,  $(1 - \beta)\%$ , uno menos la probabilidad de aceptar una hipótesis nula falsa ( $\beta$ ).
- En la práctica estos niveles se ubican en 5% para el nivel de significación y 80 o 90% para la potencia.
- Supongamos que  $\mu_T$  y  $\mu_C$  son las **medias de la variable de resultado** en los grupos de tratamiento y de control, respectivamente.
- Entonces, para la determinación del tamaño muestral se parte de las siguientes hipótesis:

$$H_0 : \mu_T = \mu_C$$

$$H_1 : \mu_T > \mu_C$$

# EPEM: determinación del tamaño muestral

- El estadístico de contraste es:

$$\frac{\bar{y}_T - \bar{y}_C}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} = \frac{\bar{y}_T - \bar{y}_C}{\sqrt{\frac{1}{n_T}} \sqrt{\sigma_T^2 + \frac{\sigma_C^2}{r}}}$$

con  $r = n_C/n_T$ .

- La distribución muestral del estadístico de contraste es,

$$\frac{\bar{y}_T - \bar{y}_C}{\sqrt{\frac{1}{n_T}} \sqrt{\sigma_T^2 + \frac{\sigma_C^2}{r}}} \sim N \left( \frac{\mu_T - \mu_C}{\sqrt{\frac{1}{n_T}} \sqrt{\sigma_T^2 + \frac{\sigma_C^2}{r}}}, 1 \right) \equiv N(\mu^*, 1)$$

# EPEM: determinación del tamaño muestral

- Entonces, la probabilidad de rechazar la hipótesis nula con un nivel de significación del  $\alpha\%$  es,

$$P[N(\mu^*, 1) > Z_\alpha] = P[N(0, 1) > Z_\alpha - \mu^*]$$

donde  $Z_\alpha$  es el valor crítico de una normal estándar al  $\alpha\%$  de significación.

- Para alcanzar una potencia de  $(1 - \beta)\%$  se define:

$$Z_\alpha - \mu^* = -Z_\beta$$

- Despejando  $n_T$  de la ecuación anterior se tiene,

$$n_T = \frac{(\sigma_T^2 + \frac{\sigma_C^2}{r})(Z_\alpha + Z_\beta)^2}{(\mu_T - \mu_C)^2}$$

$$\text{y } n_C = r \times n_T.$$

# EPEM: determinación del tamaño muestral

- Note que para determinar el tamaño muestral de ambos grupos hay que establecer el valor de parámetros poblacionales desconocidos.
- $\mu_T, \mu_C, \sigma_T^2$  y  $\sigma_C^2$ .
- En general, la determinación del tamaño muestral involucra cuatro preguntas:
  - 1 Cuál es el indicador de **resultado**?
  - 2 Cuál es el **nivel de impacto mínimo** que justificaría la inversión que se hará en el programa/política?
  - 3 Cuál sería un **nivel de potencia** razonable para la evaluación que se hará?
  - 4 Cuáles son **la media y la varianza de la variable de resultado** antes de la implementación de la política/programa?

# EPEM: determinación del tamaño muestral

- La primera pregunta se relaciona con el objetivo de la evaluación.
- La segunda es una pregunta de política más que técnica. Vale la pena implementar el programa si éste incrementa el ingreso laboral en 5%, 10%, 15%?
- Puesto de otra manera: cuál es el nivel de impacto debajo del cual, el programa se consideraría un fracaso?
- Responder esta última pregunta no solo depende del costo del programa y el tipo de beneficios que brinda sino también del costo de oportunidad de no invertir los fondos en un programa alternativo.



# EPEM: determinación del tamaño muestral

- En el caso del EPEM...
- **Econometrista:** cuál es la variable de resultado?
- **Gobierno:** el ingreso laboral. El econometrista nota que en los registros del programa el ingreso laboral promedio reportado, antes de la implementación del mismo, es de 3000 lempiras.
- **Econometrista:** cuál sería el nivel mínimo de impacto que justificaria la intervención?
- **Gobierno:** es una pregunta difícil de responder...podríamos analizar en cuanto varia el tamaño muestral para un incremento de 200, 400 y 600 lempiras?
- **Econometrista:** cuál seria un nivel de potencia razonable? en general se utiliza 80 o 90%.
- **Econometrista:** analizando los registros, el desvío estándar del ingreso laboral es de 1500 lempiras.

# EPEM: determinación del tamaño muestral

- Entonces, aplicando nuestra fórmula:

$$\begin{aligned}n_T &= \frac{(\sigma_T^2 + \frac{\sigma_C^2}{r})(Z_\alpha + Z_\beta)^2}{(\mu_T - \mu_C)^2} = \frac{(1500^2 + 1500^2)(1.6448 + 1.2815)^2}{(3200 - 3000)^2} \\ &= 963.43 \simeq 964\end{aligned}$$

y  $n_C = n_T / 1 = n_T = 964$

- En Stata, el comando es `sampsi`,

```
sampsi #1 #2 [, options]
```

donde #1 y #2 son las medias de la variable de resultado en el grupo de control y en el de tratamiento, respectivamente.

- En `options` hay que poner los desvíos estándar, `sd1 (#)` y `sd2 (#)`, de la variable de resultado en el control y tratamiento; el nivel de significación elegido, `alpha (#)`, la potencia, `p (#)` y si el test es de una cola `onesided`.

# EPeM: determinación del tamaño muestral

```
. sampsi 3000 3200, sd1(1500) sd2(1500) alpha(0.05) p(0.9) onesided
```

Estimated sample size for two-sample comparison of means

Test  $H_0: \mu_1 = \mu_2$ , where  $\mu_1$  is the mean in population 1  
and  $\mu_2$  is the mean in population 2

Assumptions:

```
alpha =    0.0500 (one-sided)
power =    0.9000
m1 =      3000
m2 =      3200
sd1 =      1500
sd2 =      1500
n2/n1 =     1.00
```

Estimated required sample sizes:

```
n1 =      964
n2 =      964
```