

## Ayudantía 2

# DATOS OBSERVACIONALES, INFERENCIA CAUSAL Y ESTRATEGIAS DE IDENTIFICACIÓN

Cristian Candia-Castro Vallejos  
Victor Landaeta Torres  
Melanie Oyarzún Wolf

## CONTENIDOS A REVISAR

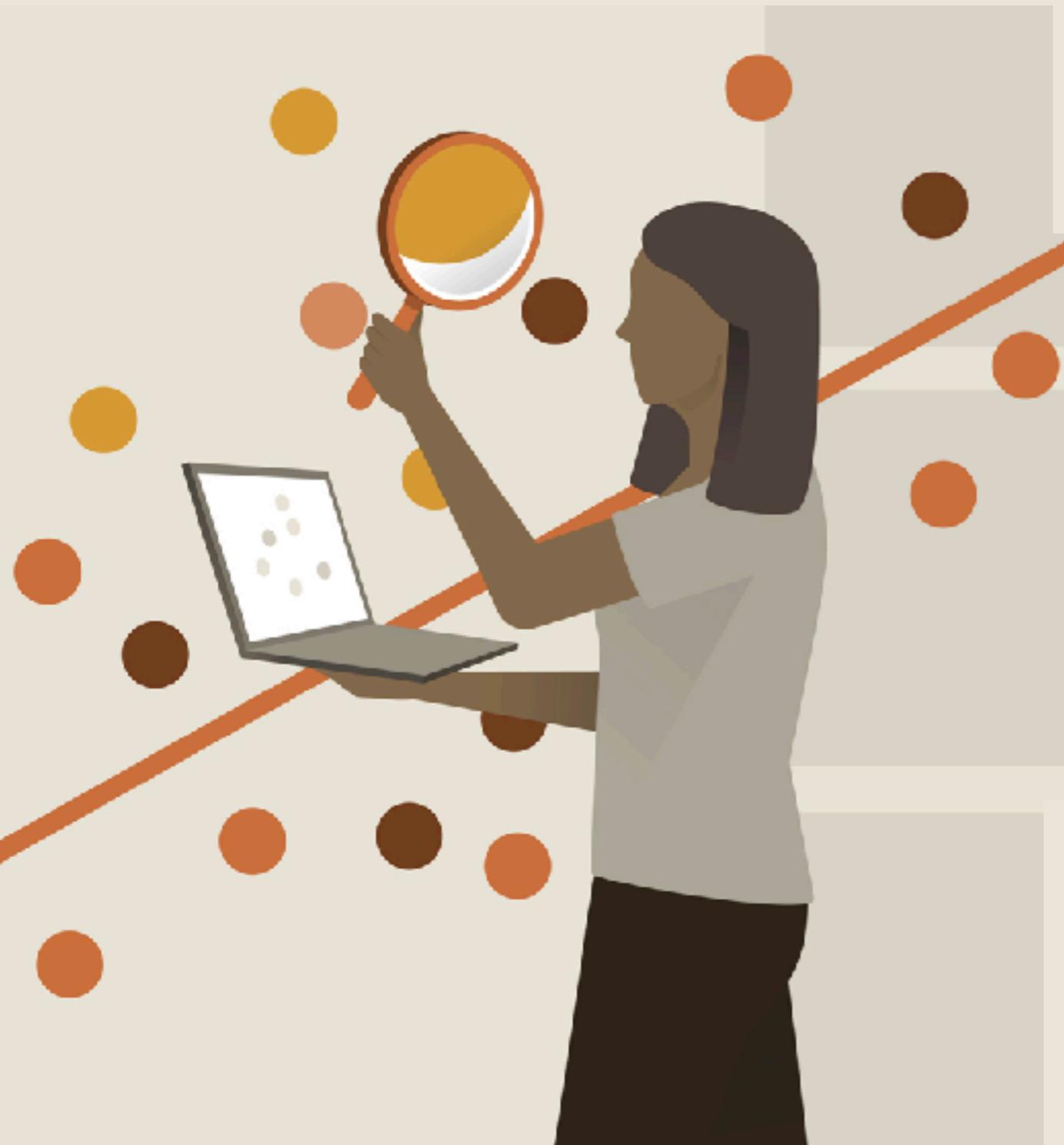
- ▶ Regresiones: recordando y elementos prácticos
- ▶ Causalidad y el problema del contrafactual
- ▶ Sesgo de selección y potential outcomes
- ▶ Diseño experimental
- ▶ Estrategias de identificación
  - ▶ Efectos fijos, aleatorios y modelos jerárquicos-
  - ▶ Método de diferencias en diferencias
  - ▶ Propensity Score Matching
  - ▶ Regresión discontinua
- ▶ Actividad Práctica



## BIBLIOGRAFÍA RECOMENDADA

- ▶ Mostly Harmless Econometrics
- ▶ <https://mixtape.scunning.com/>



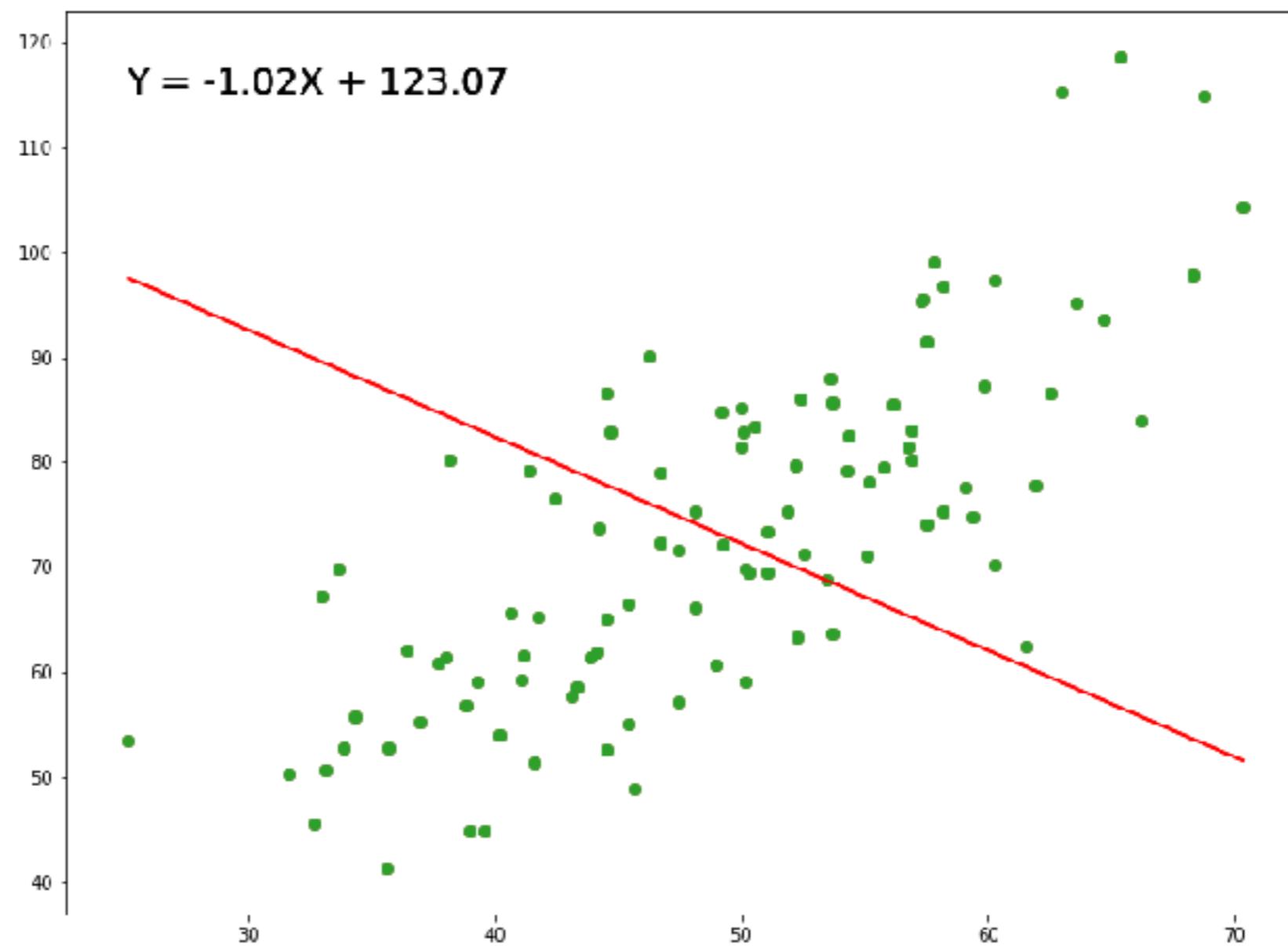


# REGRESIONES REPASO Y DETALLES PRÁCTICOS

# REGRESIONES

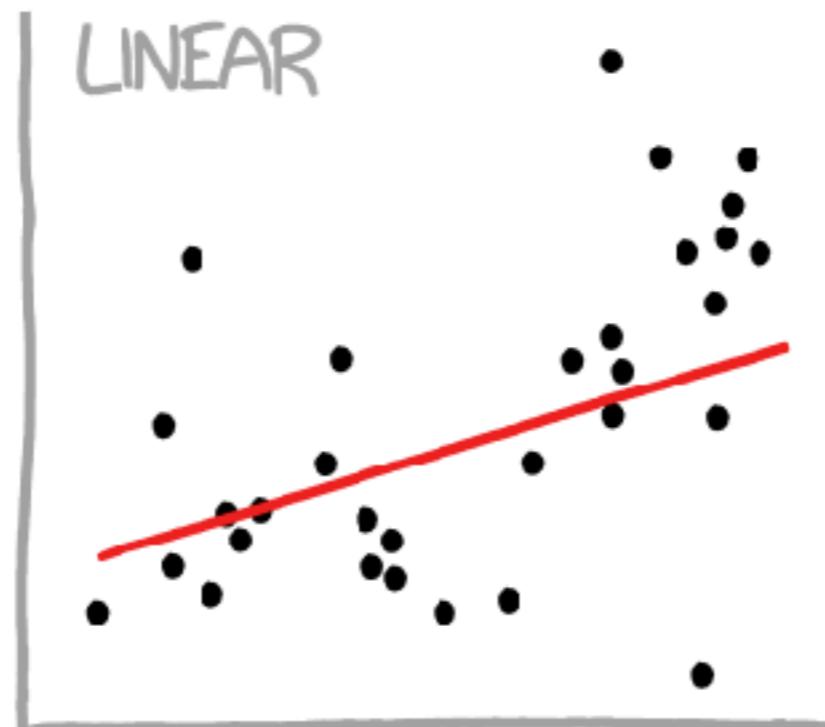
$Y$

# REGRESIONES



# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

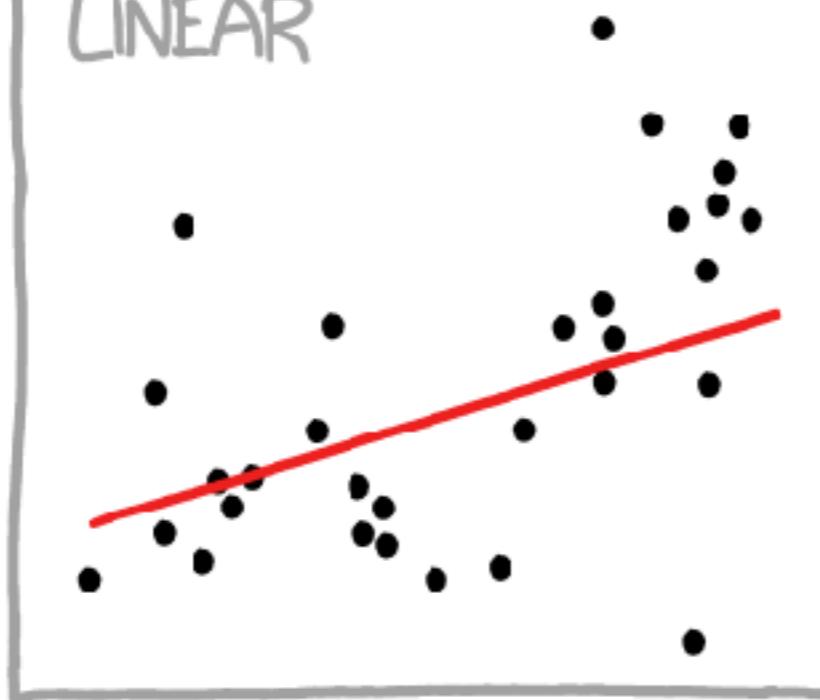
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



"HEY, I DID A  
REGRESSION."

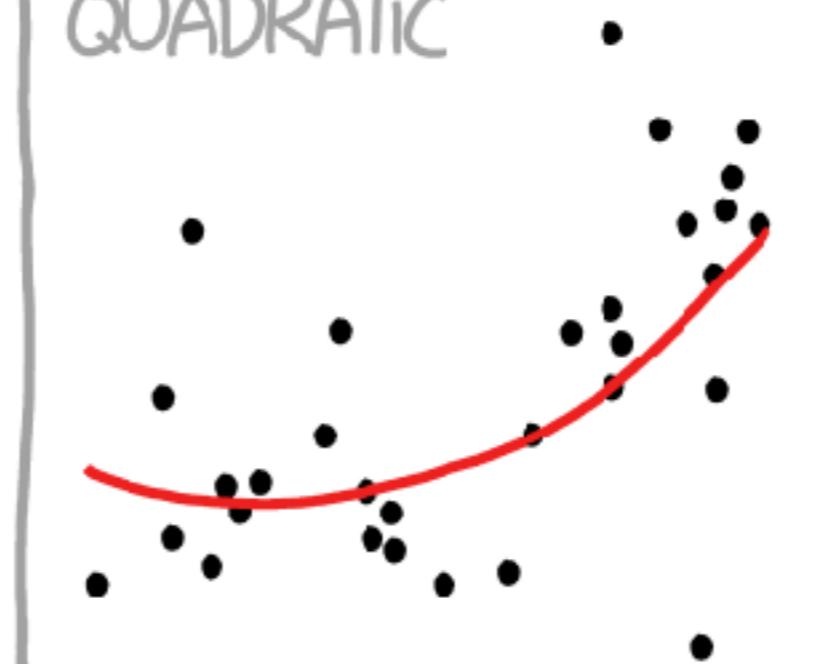
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



"HEY, I DID A  
REGRESSION."

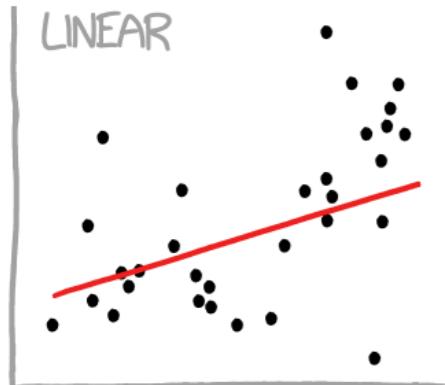
QUADRATIC



"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

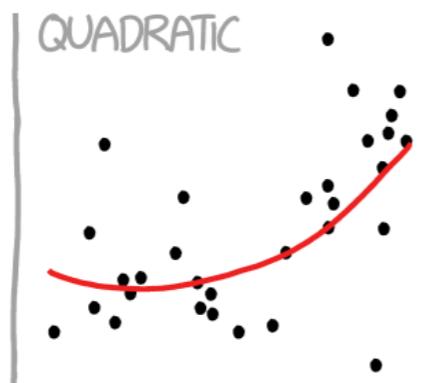
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



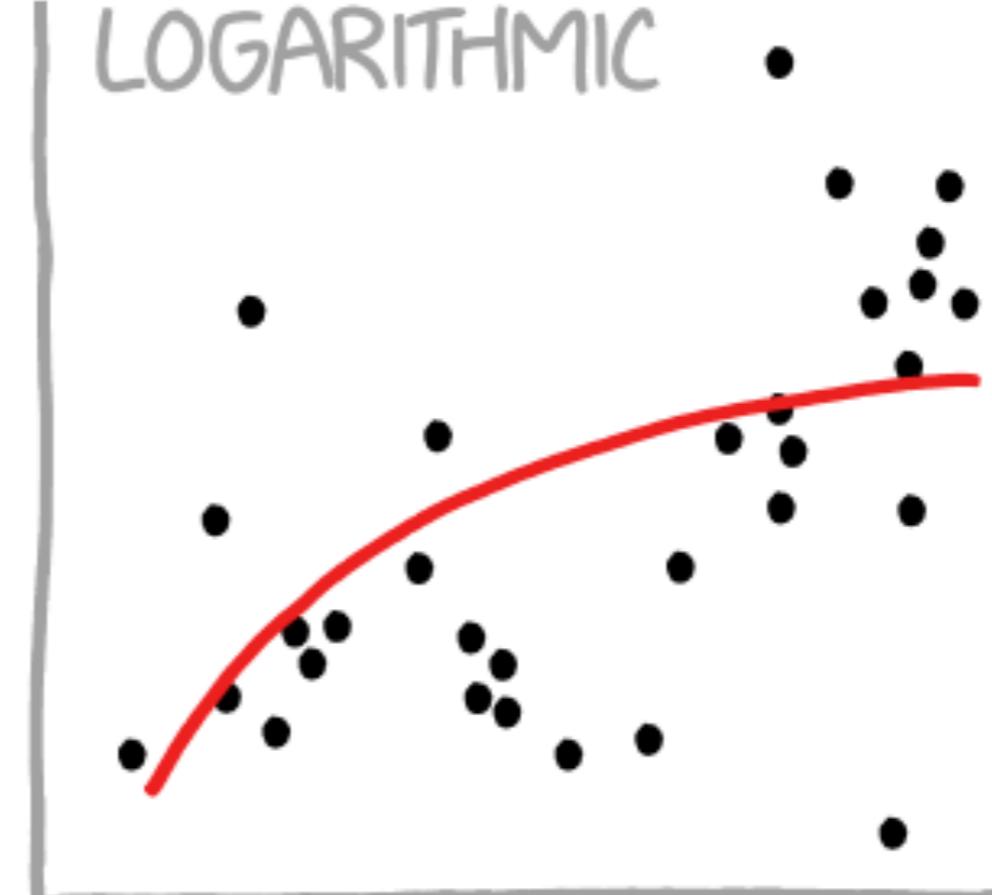
"HEY, I DID A  
REGRESSION."

QUADRATIC



"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

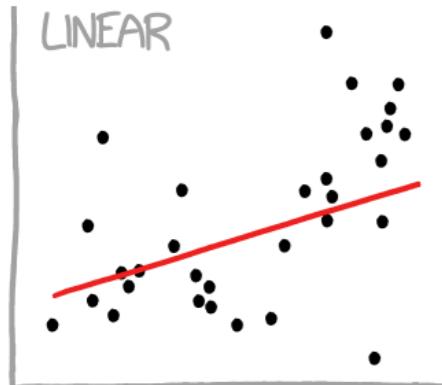
LOGARITHMIC



"LOOK, IT'S  
TAPERING OFF!"

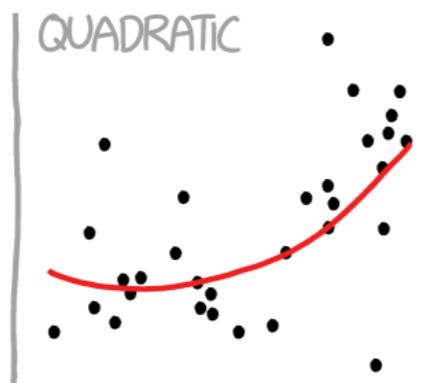
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



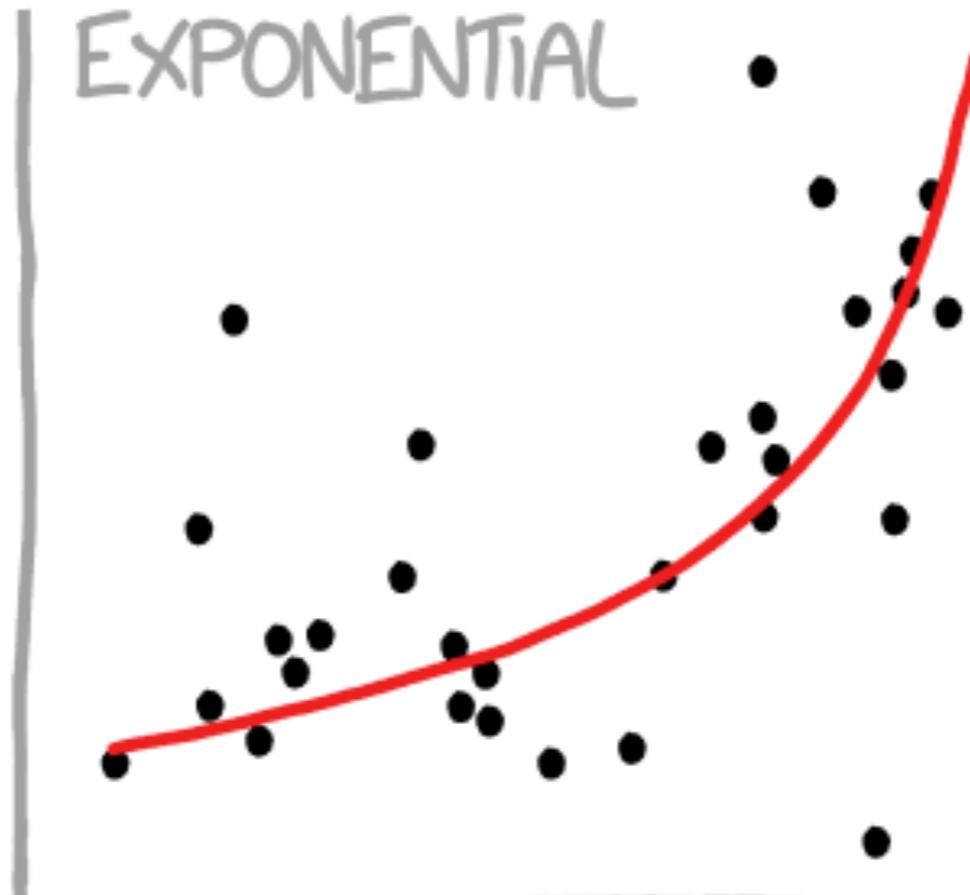
"HEY, I DID A  
REGRESSION."

QUADRATIC



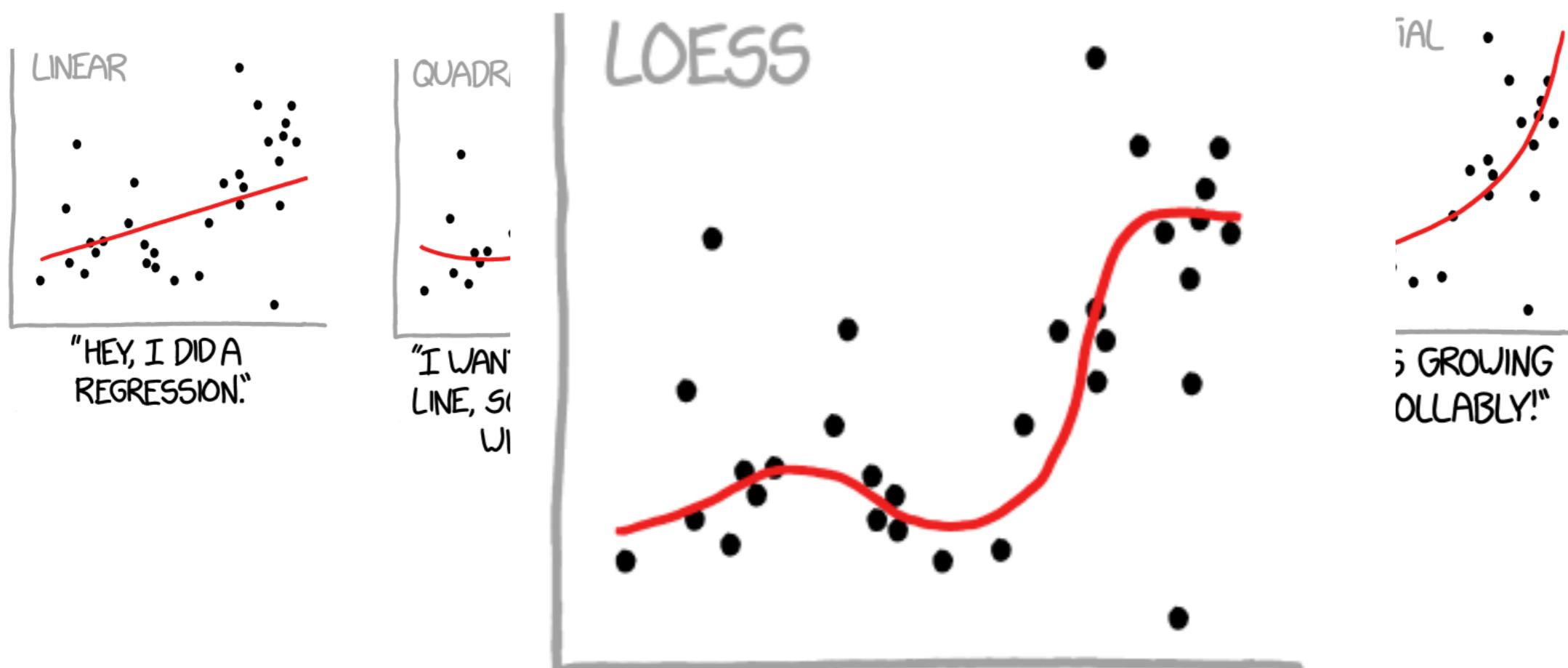
"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

EXPONENTIAL



"LOOK, IT'S GROWING  
UNCONTROLLABLY!"

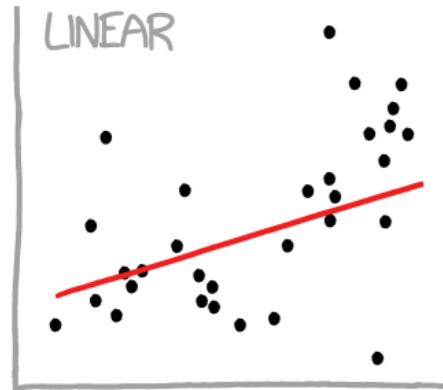
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



"I'M SOPHISTICATED, NOT  
LIKE THOSE BUMBLING  
POLYNOMIAL PEOPLE."

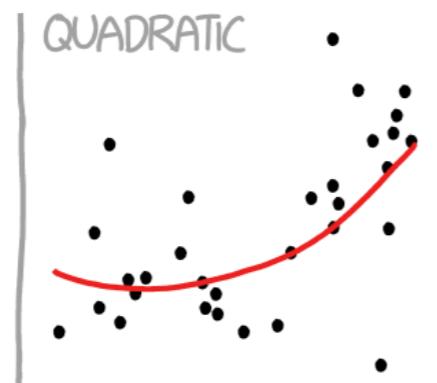
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



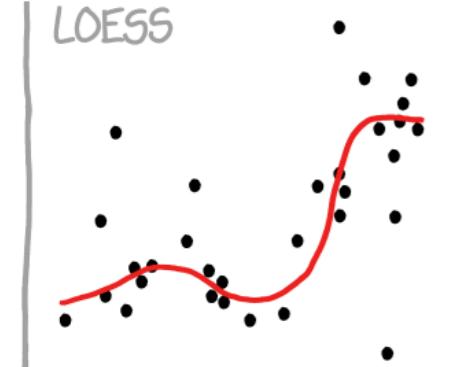
"HEY, I DID A  
REGRESSION."

QUADRATIC



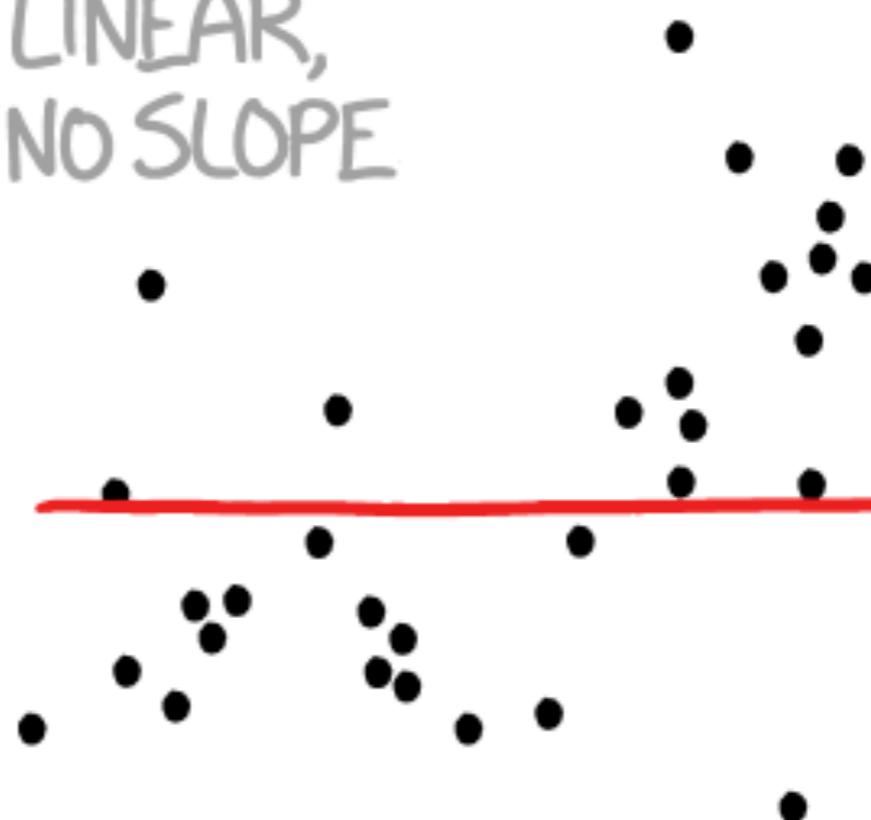
"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

LOESS



"I'M SOPHISTICATED, NOT  
LIKE THOSE BUMBLING  
POLYNOMIAL PEOPLE."

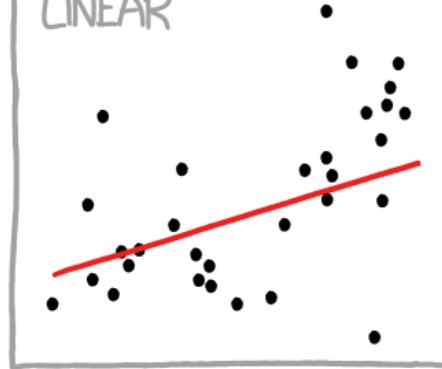
LINEAR,  
NO SLOPE



"I'M MAKING A  
SCATTER PLOT BUT  
I DON'T WANT TO."

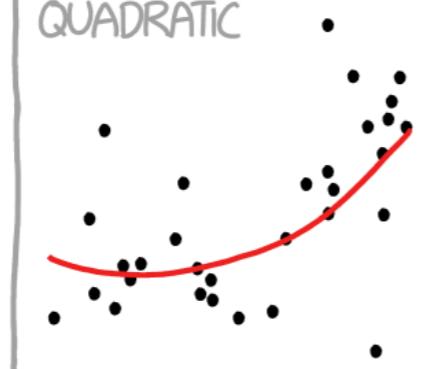
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



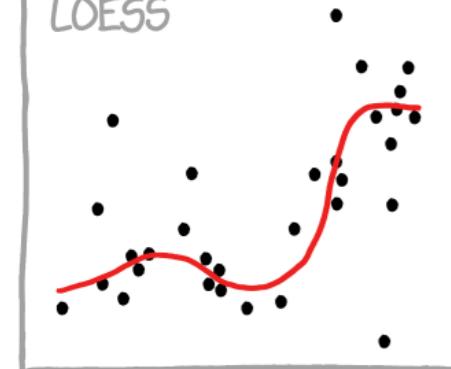
"HEY, I DID A  
REGRESSION."

QUADRATIC



"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

LOESS



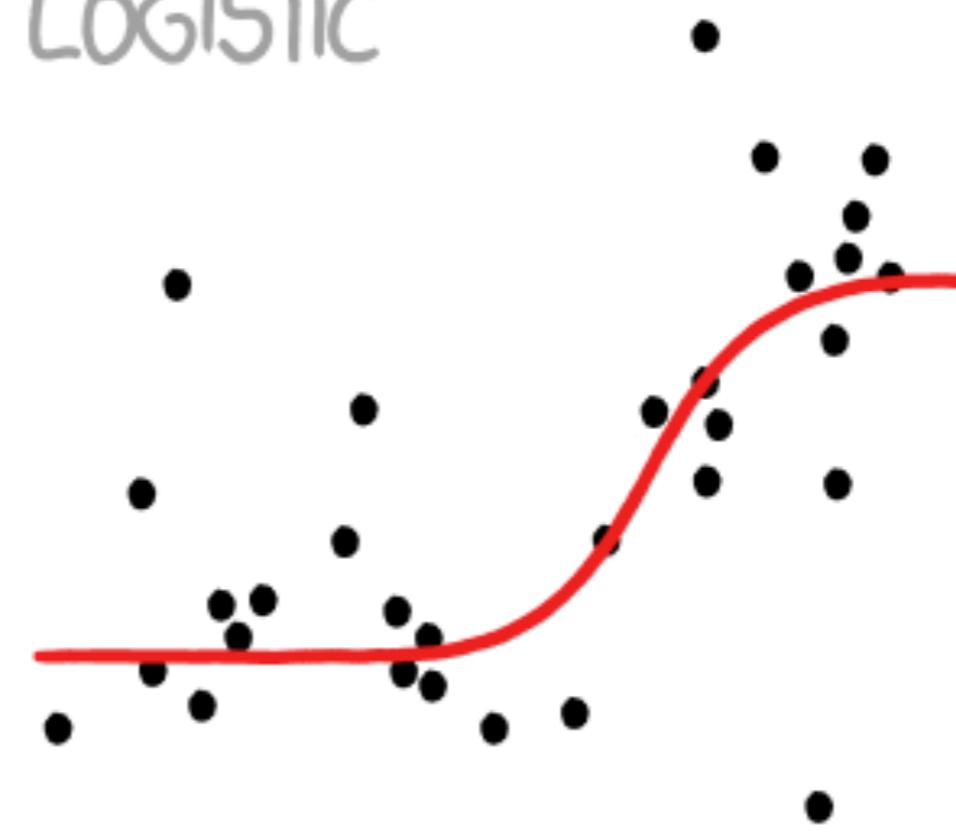
"I'M SOPHISTICATED, NOT  
LIKE THOSE BUMBLING  
POLYNOMIAL PEOPLE."

LINEAR,  
NO SLOPE



"I'M MAKING A  
SCATTER PLOT BUT  
I DON'T WANT TO."

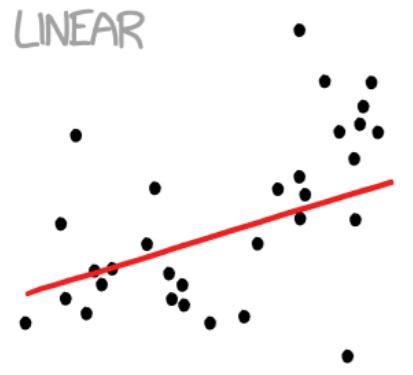
LOGISTIC



"I NEED TO CONNECT THESE  
TWO LINES, BUT MY FIRST IDEA  
DIDN'T HAVE ENOUGH MATH."

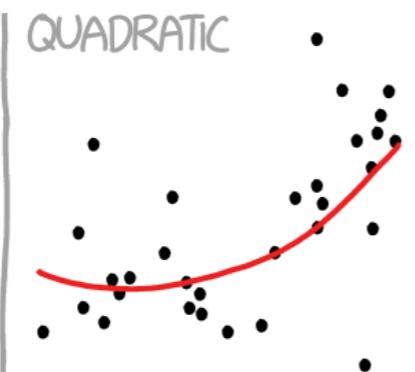
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



"HEY, I DID A  
REGRESSION."

QUADRATIC



"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

LOESS



"I'M SOPHISTICATED, NOT  
LIKE THOSE BUMBLING  
POLYNOMIAL PEOPLE."

LINEAR,  
NO SLOPE



"I'M MAKING A  
SCATTER PLOT BUT  
I DON'T WANT TO."

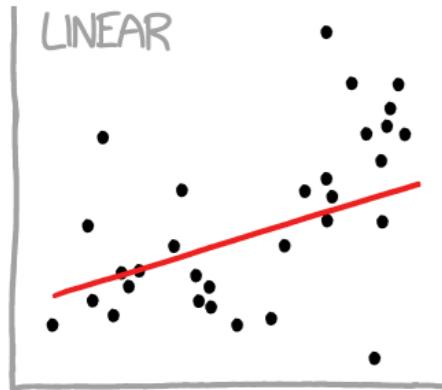
CONFIDENCE  
INTERVAL



"LISTEN, SCIENCE IS HARD.  
BUT I'M A SERIOUS  
PERSON DOING MY BEST."

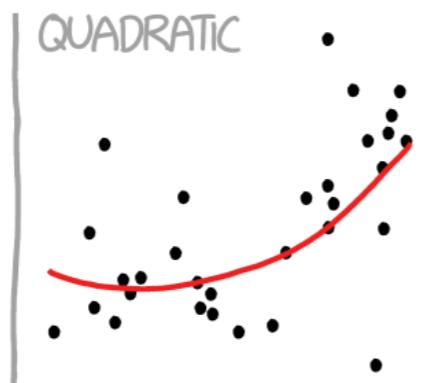
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LINEAR



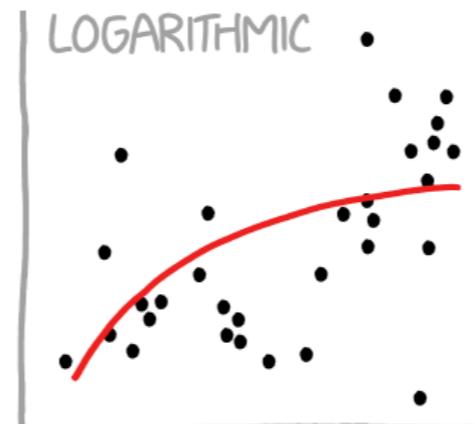
"HEY, I DID A  
REGRESSION."

QUADRATIC



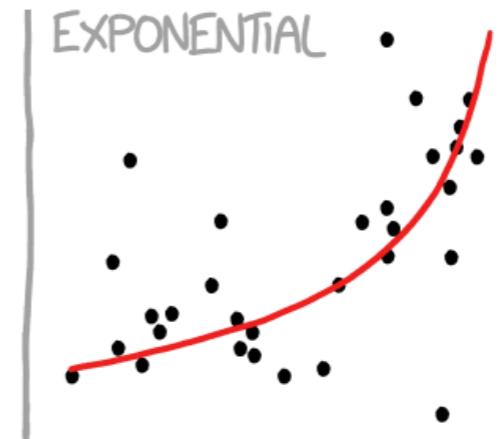
"I WANTED A CURVED  
LINE, SO I MADE ONE  
WITH MATH."

LOGARITHMIC



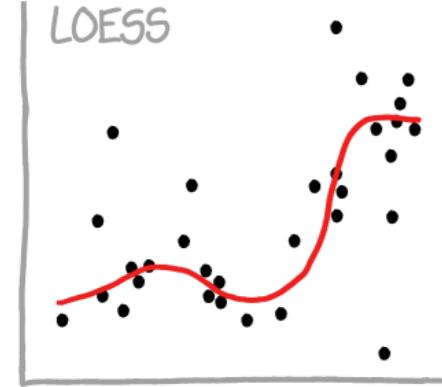
"LOOK, IT'S  
TAPERING OFF!"

EXPONENTIAL



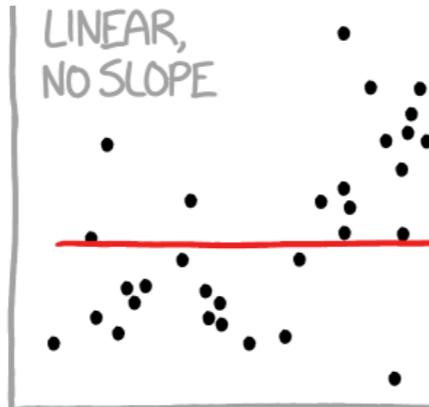
"LOOK, IT'S GROWING  
UNCONTROLLABLY!"

LOESS



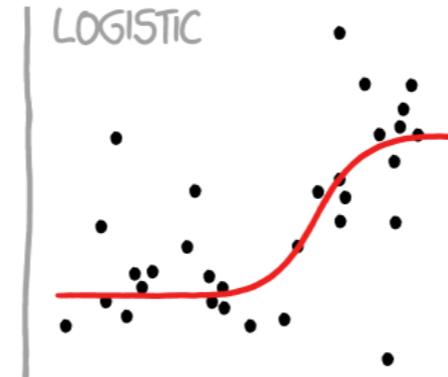
"I'M SOPHISTICATED, NOT  
LIKE THOSE BUMBLING  
POLYNOMIAL PEOPLE."

LINEAR,  
NO SLOPE



"I'M MAKING A  
SCATTER PLOT BUT  
I DON'T WANT TO."

LOGISTIC



"I NEED TO CONNECT THESE  
TWO LINES, BUT MY FIRST IDEA  
DIDN'T HAVE ENOUGH MATH."

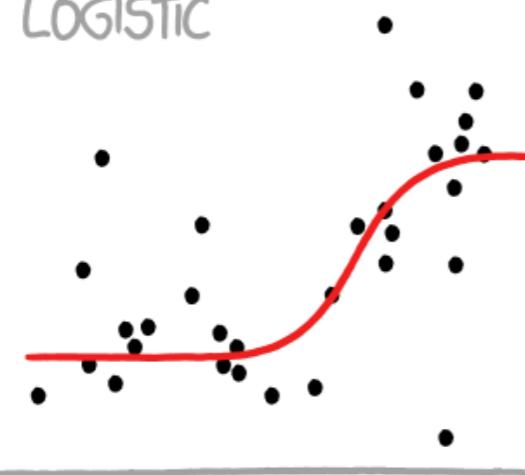
CONFIDENCE  
INTERVAL



"LISTEN, SCIENCE IS HARD.  
BUT I'M A SERIOUS  
PERSON DOING MY BEST."

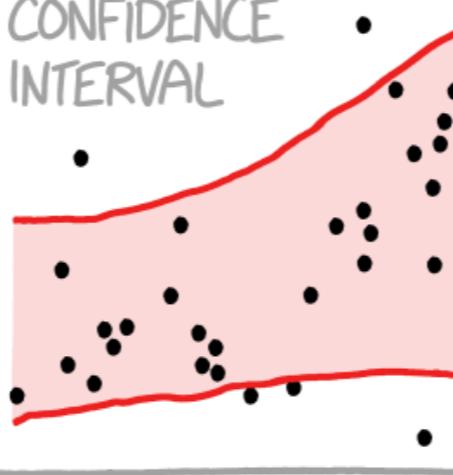
# CURVE-FITTING METHODS AND THE MESSAGES THEY SEND

LOGISTIC



"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."

CONFIDENCE INTERVAL



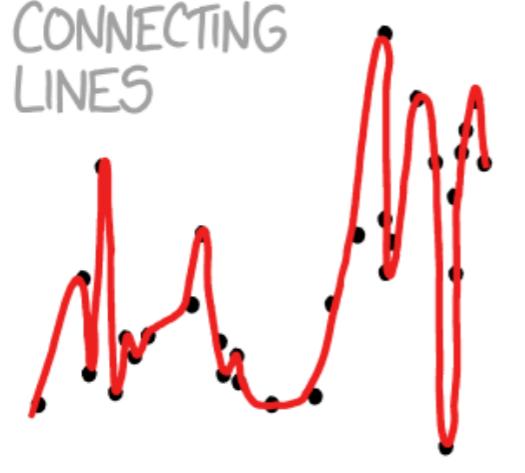
"LISTEN, SCIENCE IS HARD.  
BUT I'M A SERIOUS  
PERSON DOING MY BEST."

PIECEWISE



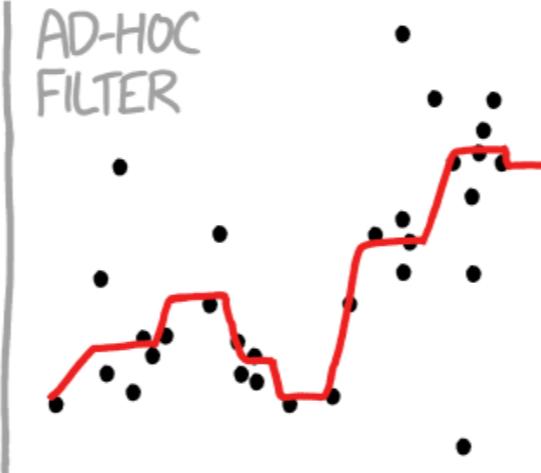
"I HAVE A THEORY,  
AND THIS IS THE ONLY  
DATA I COULD FIND."

CONNECTING LINES



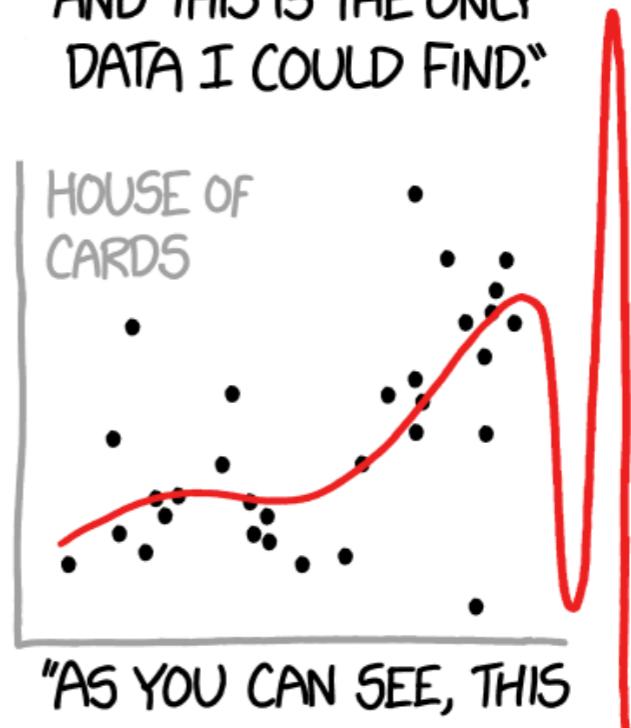
"I CLICKED 'SMOOTH  
LINES' IN EXCEL."

AD-HOC FILTER



"I HAD AN IDEA FOR HOW  
TO CLEAN UP THE DATA.  
WHAT DO YOU THINK?"

HOUSE OF CARDS



"AS YOU CAN SEE, THIS  
MODEL SMOOTHLY FITS  
THE- WAIT NO NO DON'T  
EXTEND IT AAAAAA!!"

# REGRESIONES LINEALES

## LINEAR REGRESSION

The thing we want  
to explain

DEPENDENT  
VARIABLE

↓  
*y*

i.e. 77% of the variance in *y* is  
explained by *x*. Below c.30% means  
they're hardly connected. Above 95%  
and they're practically the same.

$R^2 = 0.77$

DATA  
POINT

If you only had data on *x*, this line  
provides your best estimate of *y*. If the  
fit is strong and no major outliers, *x* could  
be used as a surrogate or forecast of *y*.

← LINE OF BEST FIT

OUTLIER

95% CONFIDENCE BAND

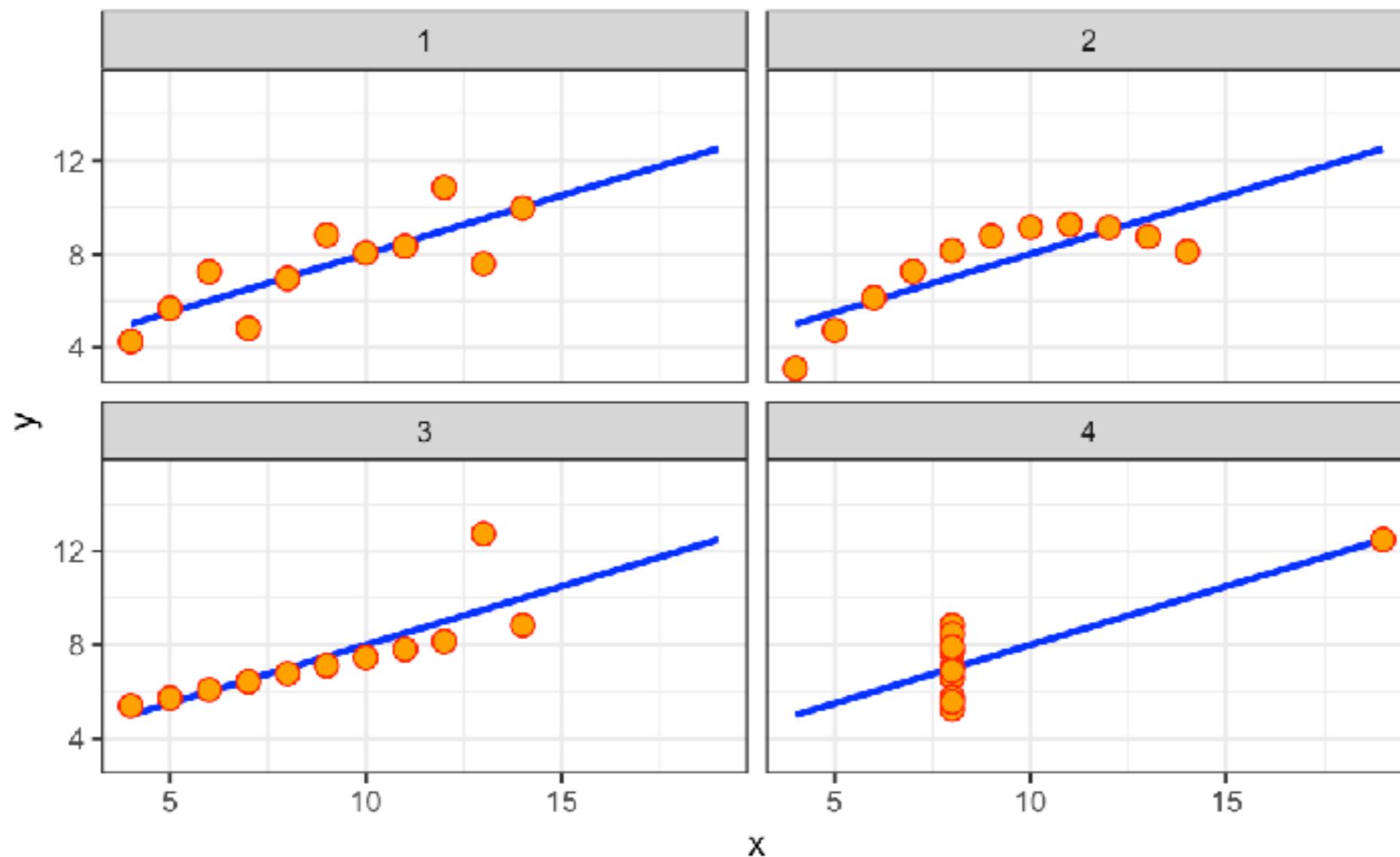
If a data point falls outside these  
lines, you're 95% sure there is  
something special about it causing it  
to do better or worse than others -  
an 'outlier' worth understanding

INDEPENDENT  
VARIABLE  
← *x*

The factor we think  
might influence the  
dependent variable

## IMPORTANCIA DE LA FORMA FUNCIONAL Y CUIDADO CON LAS MÉTRICAS

- Un buen ejemplo de una aplicación de regresión lineal sin pensarlo es el cuarteto de Ascombe.



# ACTIVIDAD DE EJEMPLO

- ▶ Vamos a empezar trabajando con un ejemplo que desarrollaremos los diferentes contenidos.
- ▶ Trabajaremos con tres datasets (ficticios):
  - ▶ Data 1:
  - ▶ Data 2:
  - ▶ Data 3:



# ACTIVIDAD DE EJEMPLO

- ▶ Primero, importemos los data sets y unámoslos usando su id.
- ▶ Vamos a seguir un tutorial que imita a los pasos en una investigación real
  - ▶ Importación y preparación de los datos
  - ▶ Estadística descriptiva



## INTERPRETAR UN COEFICIENTE

- ▶ El coeficiente de la regresión indica cuanto cambia en promedio Y en relaciones a variaciones parciales de X.
- ▶ Es en realidad un ratio entre las covarianzas estandarizadas en unidades de la otra variable:
- ▶ ¿Es entonces una **correlación o una causalidad?**

# Causalidad

## Causa y efecto

- Muchas veces, el objetivo de una pregunta o agenda de investigación está en entender alguna relación causal de interés.

*RAE: De causal.*

1. *f.* Causa, origen, principio.
2. *f.* *Fil.* Ley en virtud de la cual se producen efectos.

*Florían B., Víctor - Diccionario de filosofía: Panamericana Editorial, 2012:*

*La causalidad es la "relación que se establece entre causa y efecto. Se puede hablar de esa relación entre acontecimientos, procesos, regularidad de los fenómenos y la producción de algo".*

# El problema de la estimación causal

- Para ilustrar el problema de la estimación causal hagamos con un ejemplo.
- Imaginemos que queremos saber si los hospitales afectan la salud de las personas.
  - Podríamos simplemente: realizar una encuesta en una población de gente adulta de los cuales algunos asistieron al hospital y otros no y comparar la salud que reportan.
  - La salud es autoreportada como excelente (1), muy buena (2), buena (3), normal (4), o mala (5)

Grupo	Muestra	Promedio salud	Error estándar
Hospital	7,774	2.79	0.014
No hospital	90,049	2.07	0.003

¿Porqué es esto una mala idea?

# El problema del contrafactual

## Sesgo de selección

- ¿Porque no podemos simplemente comparar los grupos?
  - Lo que ocurre se suele llamar el problema de selección, esto quiere decir que quienes están en un grupo u otro no es al azar.
  - En nuestro ejemplo, las personas que asisten al hospital, probablemente sea por algo y tengan peor salud en promedio. Eso no quiere decir que ir al hospital empeore su salud.
  - Esto hace que comparar estos grupos tienen elementos **inobservables** que están afectando el resultado que queremos analizar y que no estamos midiendo.

# El problema del **contrafactual**

- A simple vista, los que asisten al hospital tienen peor salud.
- El problema es que estamos comparando peras con manzanas.

Grupo	Muestra	Promedio salud	Error estándar
Hospital	7,774	2.79	0.014
No hospital	90,049	2.07	0.003

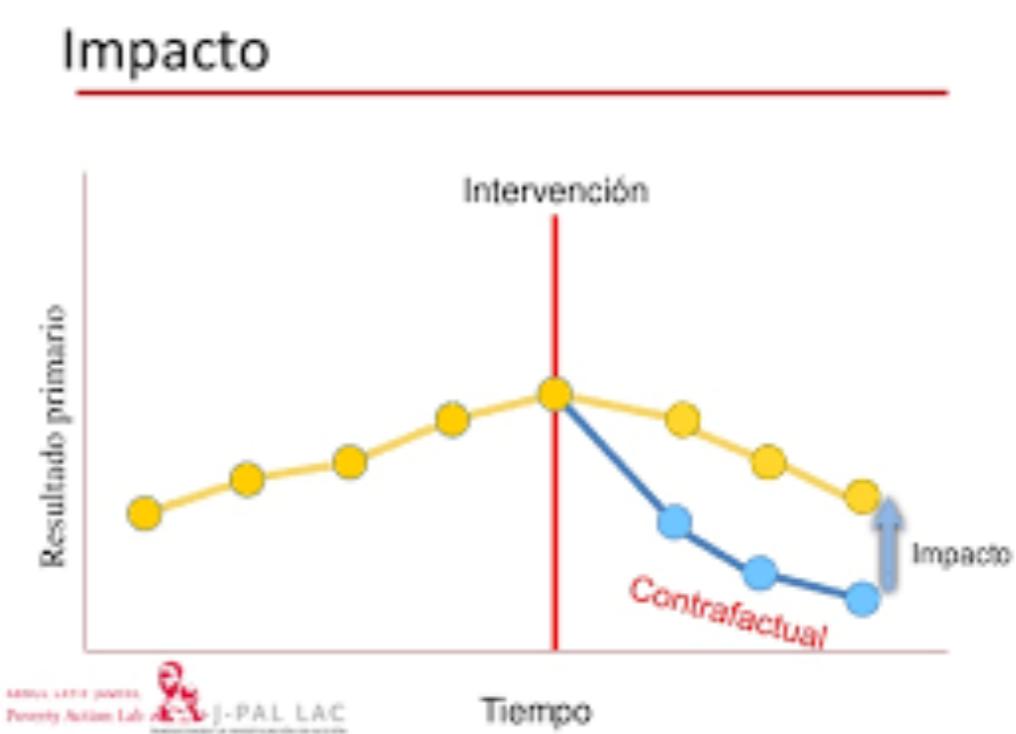
- Ya que quisiéramos saber el impacto de asistir al hospital, deberíamos comparar qué habría pasado con la salud de la persona A si no hubiese ido al hospital y cuando esa MISMA persona si va al hospital.
  - Esa diferencia es el efecto causal, pero es IMPOSIBLE de observar.
- Esta otra misma persona es lo que se llama el **contrafactual**
- Problema fundamental de inferencia causal (Holland 1986): observamos a un individuo en uno de los dos escenarios.

## EL PROBLEMA DEL CONTRAFACTUAL

- ▶ ¿Qué habría ocurrido con los sujetos **en ausencia** del programa/intervención en cuestión



# EL PROBLEMA DEL CONTRAFACTUAL



# Interpretación causal de un modelo

- Sabemos que correlación no implica causalidad. . . PERO:
  - las correlaciones importan
  - hay correlaciones robustas y no robustas

Considerando el modelo:  $y = \alpha + \beta x + u$

- ¿Cómo saber si un parámetro  $\beta$  mide una correlación o una causalidad?
  - Depende de si hay **estrategia de identificación** para superar el problema de selección y su calidad.
  - Cuando una variable está sometida a estos problemas, se le llama **endógena**

# Interpretación causal de un modelo

- La interpretación de un modelo y sus parámetros estimados es una de las partes más importantes de la investigación empírica
- Muchos tipos de investigación:
  - teórica sin datos (ej. Heckscher-Ohlin)
  - teórica con datos (ej. calibración, modelo estructural)
  - descriptiva sin regresiones (ej. comparar promedios entre grupos)
  - descriptiva con regresiones (ej. estimar correlaciones parciales)
  - causal con experimentos naturales (o cuasi-experimentos)
  - causal con experimentos aleatorios

# Explicación vs predicción

Modelos explicativos en perspectiva:

**Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation**

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

The rows highlight where we focus our attention (on either specific features that might affect an outcome of interest, or directly on the outcome itself), whereas the columns specify what types of situations we are modelling (a ‘fixed’ world in which no changes or interventions take place, or one in which features or inputs are actively manipulated or change owing to other uncontrolled forces).

Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, et al. 2021.  
“Integrating Explanation and Prediction in Computational Social Science.”  
*Nature* 595 (7866): 181–88.



# Interpretación causal de un modelo

## Endogeneidad

Generalmente estamos en situaciones en que el coeficiente  $\beta$  asociado a  $x$  en una regresión es endógeno o  $x$  es una variable endógena si estamos en alguno de los siguientes casos:

### 1. Variable omitida

- debe estar correlacionada con  $x$
- debe explicar variable dependiente  $y$

### 2. Causalidad reversa

- $x$  no solo afecta a  $y$ , también  $y$  afecta  $x$

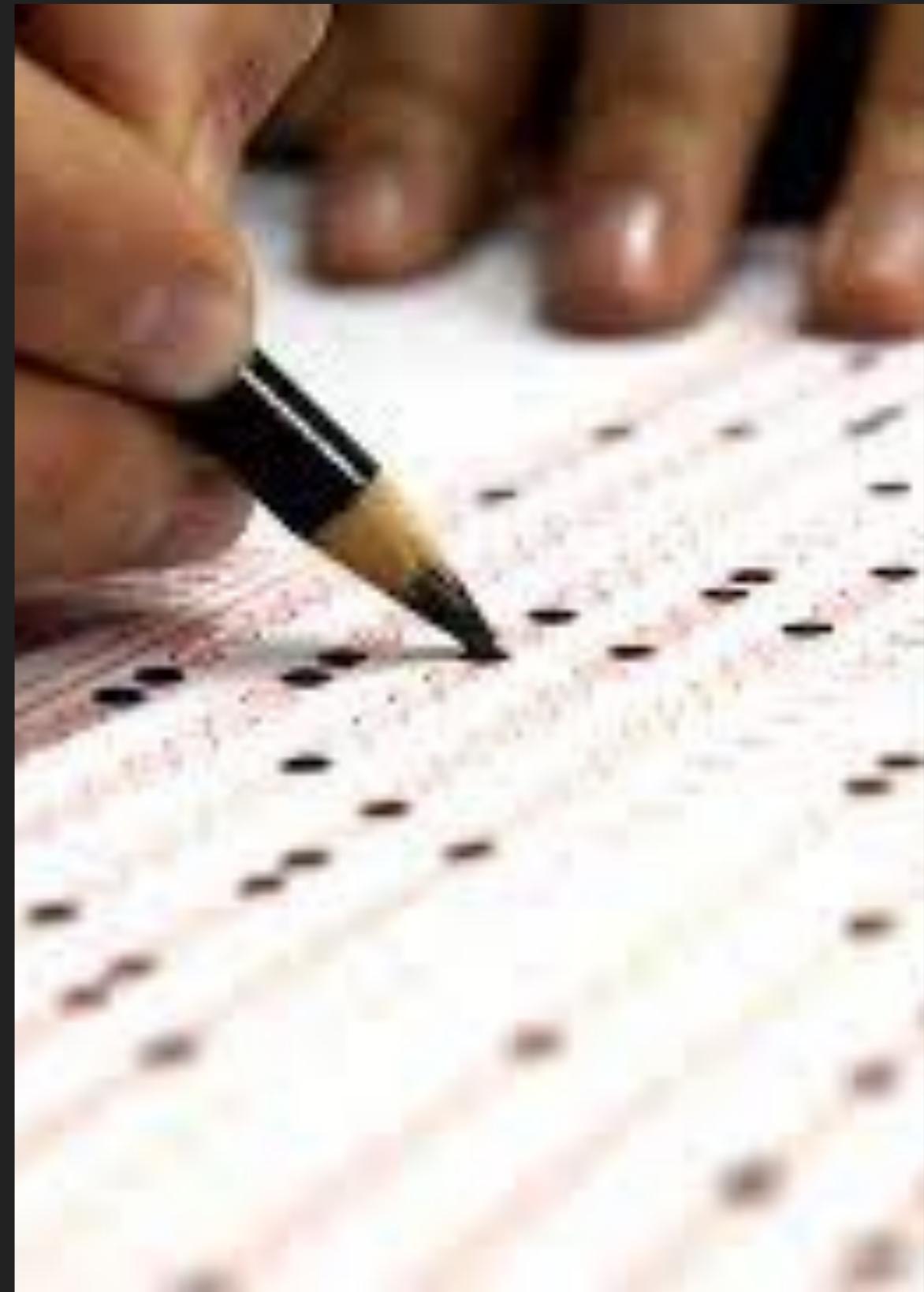
### 3. Bad control

- control está afectado por  $x$
- posible reinterpretación de  $\beta$

Si una variable no es endógena, decimos que es exógena y la estimación directa por OLS da el efecto causal.

# ACTIVIDAD DE EJEMPLO

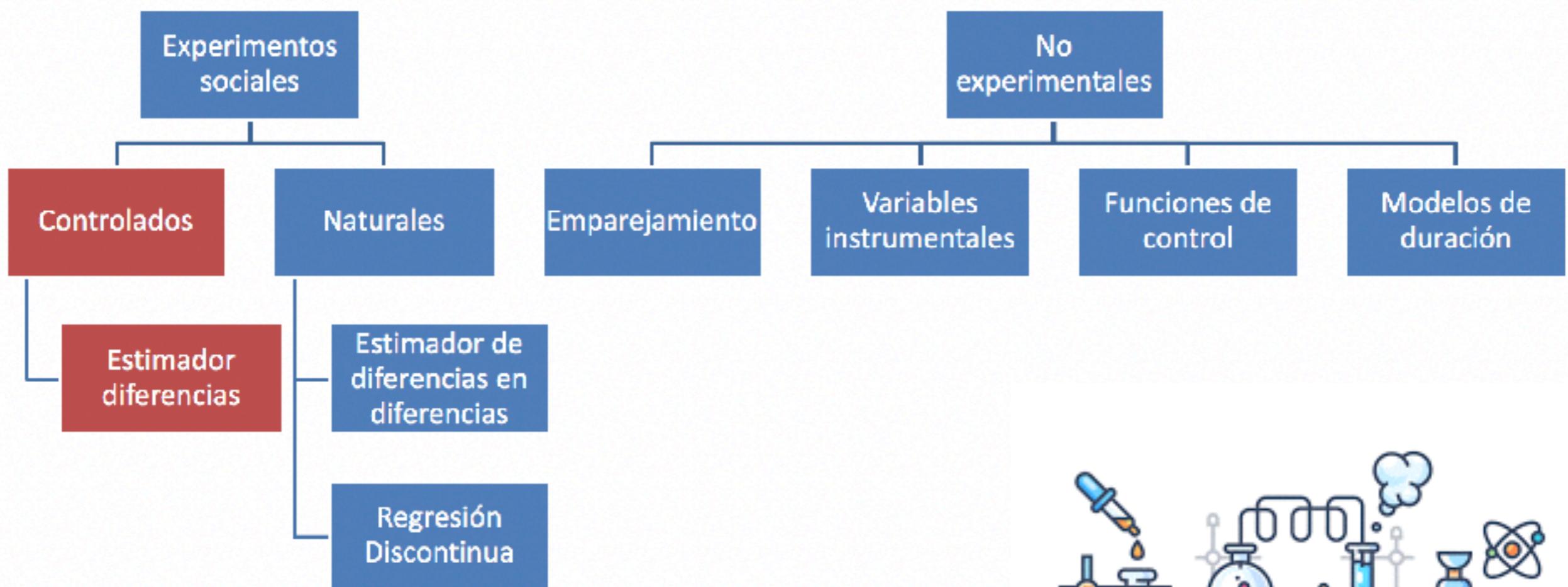
- ▶ Plantee un modelo que relacione los resultados en el colegio y la asistencia la escuela de verano.
- ▶ Estime el modelo e interprete los coeficientes.
- ▶ ¿Es nuestra estimación causal?
- ▶ ¿Cuál sería un contrafactual válido para este caso?
- ▶ ¿Qué problemas que generan endogeneidad podrían estar presentes?



## IDENTIFICACIÓN:

- ▶ Cuando la estimación de un parámetro es consistente, decimos que el parámetro está identificado
- ▶ El conjunto de los datos, la especificación, y la variación en los datos que se utiliza para estimar el parámetro de interés se conoce como estrategia de identificación
- ▶ Si la estrategia de identificación es convincente, decimos que se usa variación exógena para identificar el parámetro

# METODOLOGÍA USADAS PARA COMBATIR LA ENDOGENEIDAD



# Investigación Causal

- ▶ Partes de una investigación empírica que intenta estimar una relación causal de interés para el investigador:
  1. Establecer relación causal de interés (teórica o a-teórica)
  2. El experimento ideal (revela endogeneidad)
  3. Desarrollar una estrategia de identificación
    1. experimento natural (usa datos observacionales)
    2. experimento aleatorio controlado (RCT)
  4. Explicitar la inferencia (población de interés, datos a usar, cómo se calcular los errores estándar)

# EXPERIMENTOS Y CUASI EXPERIMENTOS

- ▶ Un **experimento** es diseñado e implementado de manera consciente por un grupo de investigadores.
  - ▶ Asignación aleatoria a grupos de control y tratamiento
  - ▶ Imitan a los experimentos clínicos.
- ▶ Un **cuasi experimento** o **experimento natural** se basa en un hecho que genera una fuente de variación exógena *como si* fuera aleatoria, sin embargo no es la acción consciente de ningún humano.
- ▶ **Evaluación de impacto** se preocupa de entender los efectos de una política, programa o causa.



# EXPERIMENTOS Y POTENTIAL OUTCOMES

# EXPERIMENTOS CONTROLADOS EN ECONOMIA

- ▶ Experimentos ideales nos entregan la mejor aproximación al efecto causal de algún tratamiento.
- ▶ Los experimentos son caros, pero influenciarles en la política y entorno.
- ▶ Ejemplos:
  - ▶ Medicamentos: ¿Este compuesto baja el colesterol?
  - ▶ Capacitación laboral
  - ▶ Efecto del tamaño de clase

# OUTCOME POTENCIAL Y EL EXPERIMENTO IDEAL

- ▶ Un tratamiento tiene un efecto causal para un determinado individuo:
  - ▶ Si le das el tratamiento, pasa algo que habría sido distinto en la ausencia de ese tratamiento.
- ▶ Un **outcome potencial** es el resultado del individuo bajo un posible estado de tratamiento.
- ▶ Para un individuo vamos a definir el **efecto causal** como la diferencia entre los resultados potenciales si se recibe o no el tratamiento.
- ▶ Este efecto no puede ser observado, nunca!

## OUTCOME POTENCIAL Y EL EXPERIMENTO IDEAL

- ▶ Supongamos que queremos estudiar el efecto causal que tienen los hospitales ( $X$ ) en la salud de las personas ( $Y$ )
- ▶ Tenemos datos de una población de gente adulta, algunos de los cuales usan la sala de emergencia del hospital (ER). La salud es autoreportada como excelente (1), muy buena (2), buena (3), normal (4), o mala (5)
- ▶ Se sugiere comparar la salud de personas que han ido al hospital con la salud de personas que no han ido en el último tiempo

---

Grupo	Muestra	Promedio salud	Error estándar
Hospital	7,774	2.79	0.014
No hospital	90,049	2.07	0.003

## OUTCOME POTENCIAL Y EL EXPERIMENTO IDEAL

- Sea  $D_i = \{0, 1\}$  ir o no al hospital;  $Y_i$  una medida de salud:

$$\text{potential outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \quad (1)$$

- El outcome observado puede escribirse como potential outcomes:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i \quad (2)$$

- Problema fundamental de inferencia causal (Holland 1986): observamos a un individuo en uno de los dos escenarios

# EFECTO DE TRATAMIENTO PROMEDIO

- ▶ Cada persona tiene un diferente efecto al ser tratado o participar de un programa.
- ▶ El **efecto de tratamiento promedio (ATE)** es el promedio de los efectos de tratamientos individuales.

## EL SESGO DE SELECCIÓN

- ▶ Si comparamos a personas que fueron al hospital con aquéllas que no fueron  $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ :

$$= \underbrace{E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)}_{\text{Efecto de ir al hospital en los que fueron}}$$

$$+ \underbrace{E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)}_{\text{Sesgo de selección}}$$

- ▶ El primer término se conoce como efecto del tratamiento en los tratados (ATT, average treatment on the treated en inglés)

## LA ASIGNACIÓN ALEATORIA ELIMINA ESE SESGO

- ▶ La asignación aleatoria de un tratamiento elimina el sesgo de selección porque la aleatoriedad hace que  $D_i$  sea independiente del potential outcome:

$$\begin{aligned}E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= \\&= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\&= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)\end{aligned}$$

- ▶ Lo que se puede simplificar aun más:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(Y_{1i} - Y_{0i})$$

# EXPERIMENTOS CONTROLADOS

Gráfico 3.1 El clon perfecto

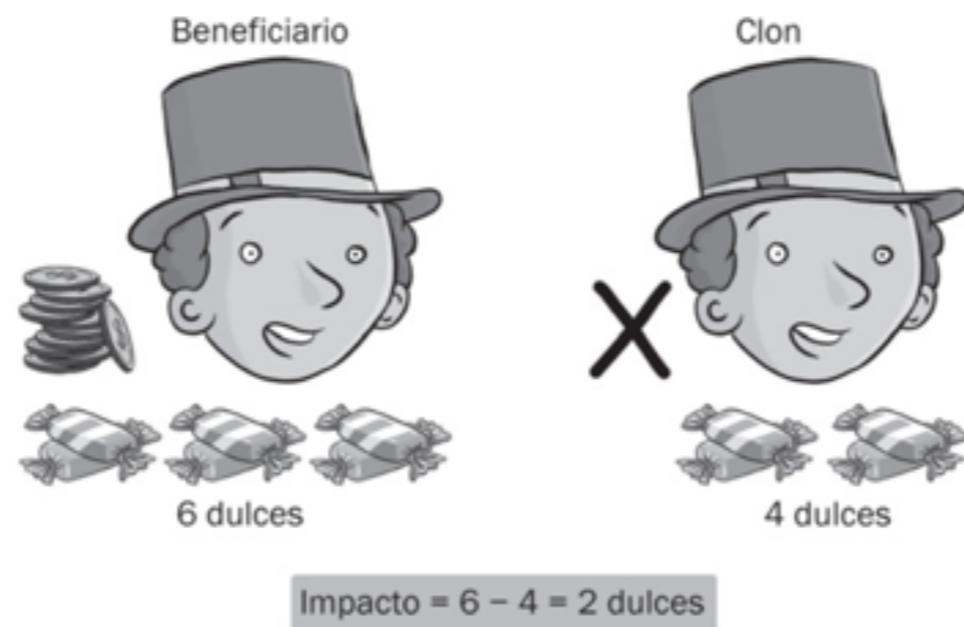
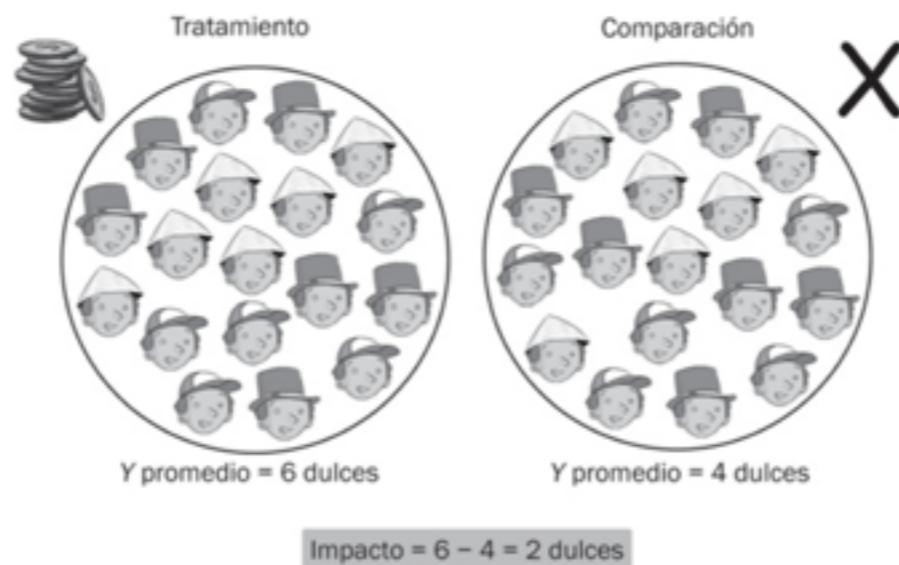


Gráfico 3.2 Un grupo de comparación válido



## EXPERIMENTOS CONTROLADOS

- ▶ Al grupo de personas que recibieron el tratamiento le llamamos tratados, a los que no recibieron el tratamiento le llamamos controles, quienes representan el contrafactual
- ▶ Si la aleatorización es exitosa, deberíamos observar que tratados y controles son similares antes de recibir el tratamiento
- ▶ La medición de características entre ambos grupos se hace con cuestionarios de base y se reporta típicamente en una tabla con comparaciones de medias. Se le llama balance en observables

## ESTIMACIÓN DEL EFECTO TRATAMIENTO

- ▶ ¿Cómo estimamos el efecto tratamiento en los tratados?

$$Y_i = \alpha + \beta D_i + \varepsilon; \quad (5)$$

$$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon; \quad (6)$$

- ▶ Si  $D_i$  fue asignado de manera aleatoria, entonces  $\beta$  representa el efecto causal de  $D_i$  en  $Y_i$ . Cuando  $D_i$  es binaria,  $\alpha$  representa el valor promedio de  $Y_i$  en el grupo de control ( $Y_{0i}$ )
- ▶ Esta regresión podemos estimarla simplemente usando mínimos cuadrados ordinarios (podríamos también simplemente mostrar comparaciones de medias entre grupos usando gráficos)
- ▶ ¿Cuál es el rol las variables de control?

## EL ESTIMADOR DE DIFERENCIAS

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

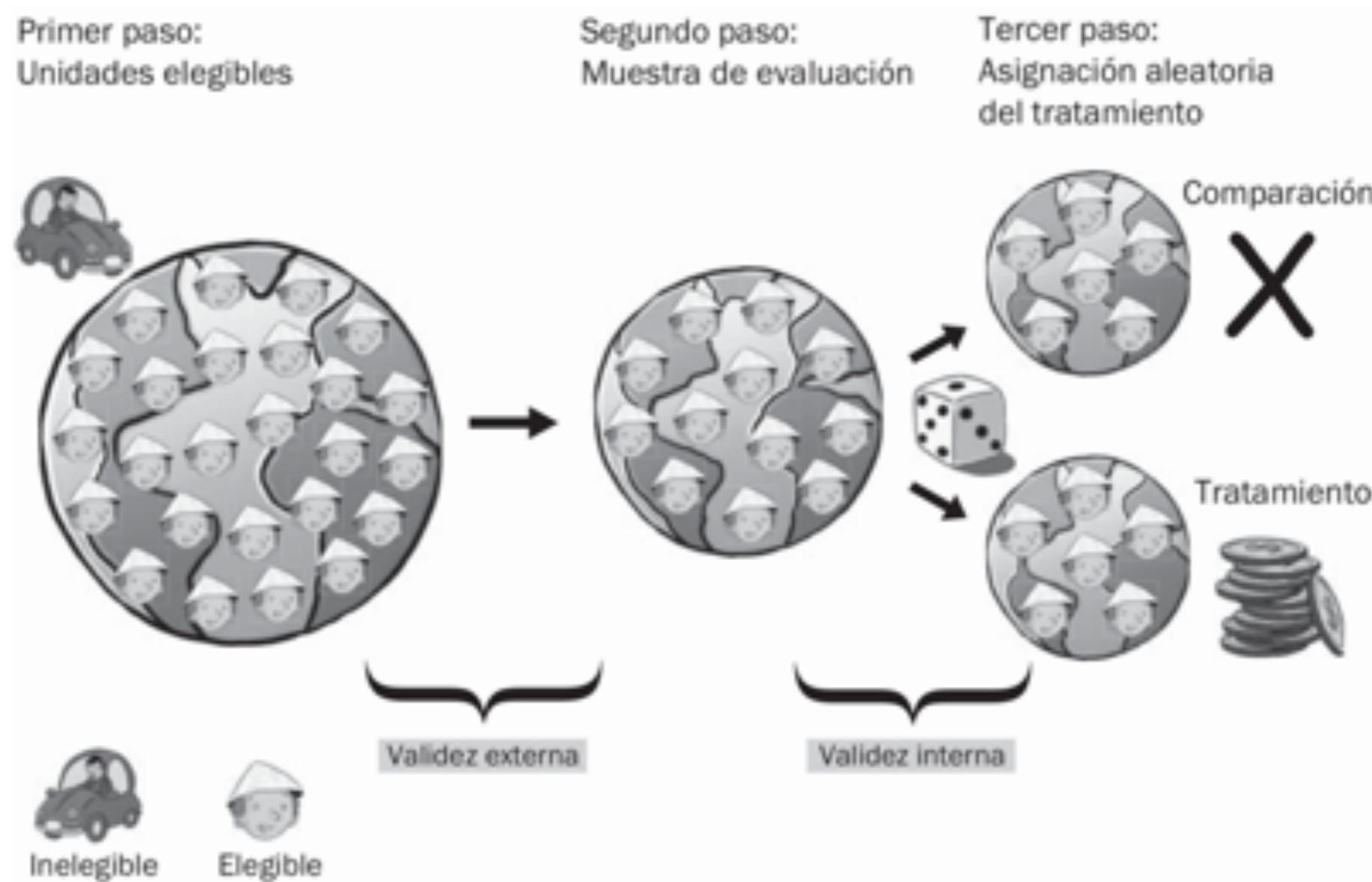
- ▶ Cuando el tratamiento es binario, el estimador va a ser simplemente la diferencia entre el promedio del grupo de tratamiento y de control.

$$\bar{Y}^{treated} - \bar{Y}^{control}$$

- ▶ Este estimador, también es llamado el estimador de diferencias y en este caso corresponde a una estimación del ATE.

# ASIGNACIÓN ALEATORIA

Gráfico 4.3 Pasos para la asignación aleatoria del tratamiento



# EJEMPLO DE APLICACIÓN

- ▶ Sigamos con nuestro ejemplo.
- ▶ Suponga que a un grupo de estudiantes del colegio se les invitó a participar (gratuitamente) de la escuela de verano.
- ▶ Plantee el modelo de diferencias y estímelo directamente.
- ▶ Interprete sus resultados.
- ▶ Obtenga la tabla de balance.



### VALIDEZ INTERNA

- ▶ Los experimentos suelen ser especialmente fuertes en su validez interna. Es decir, la capacidad de obtener un estimador causal para EL GRUPO EXPERIMENTAL:
- ▶ Algunas amenazas son:
  - ▶ **Falla al aleatorizar**
  - ▶ **Falla al entregar el tratamiento**, tratamiento parcial o *partial compliance*
    - ▶ Solución: usar el tratamiento como un instrumento!!
  - ▶ **Atrición**

# VALIDEZ EXTERNA

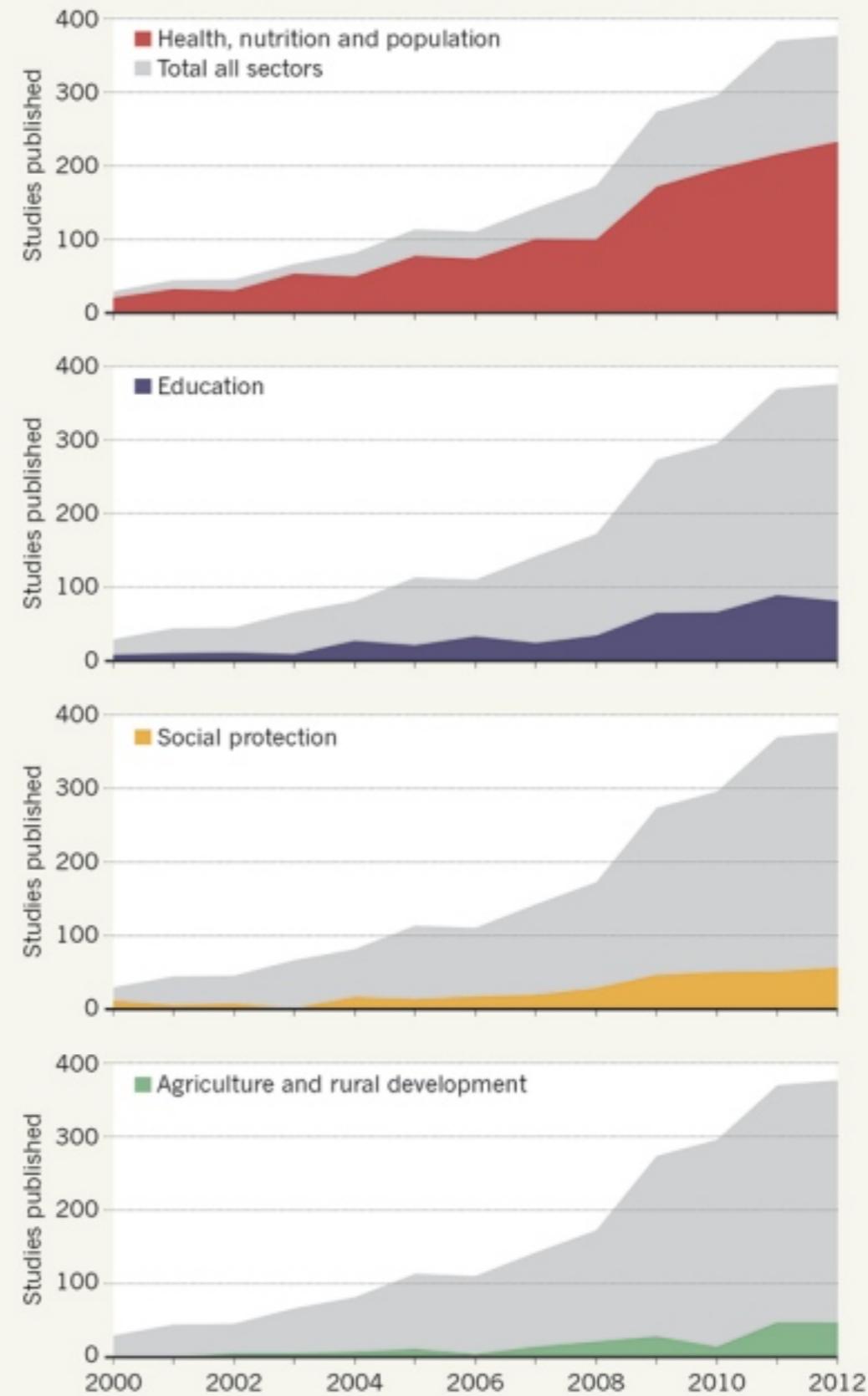
- ▶ **Generalización: ¿puede el resultado en muestra A extrapolarse a B?**
- ▶ Los experimentos suelen tener problemas de validez externa:
  - ▶ **Muestra no representativa**
  - ▶ **Tratamiento no representativo**
  - ▶ **Efectos de equilibrio parcial y general**
    - ▶ Los experimentos pueden no capturar efectos de equilibrio general
    - ▶ Depender de la escala
- ▶ **Variables contextuales**
  - ▶ Tendencia actual: meta-análisis

## SITUACIÓN ACTUAL:

- ▶ Actualmente es uno de los métodos más usados.
- ▶ “Estandar de oro” o benchmark para encontrar efectos causales.
- ▶ Principal herramienta en

### SCALE THE HEIGHTS

The growing influence of the randomized controlled trial in economic spheres can be seen in the number of studies published each year. Most of the increase is in four sectors — although many studies overlap.



## ¿TODO EXPERIMENTO ASEGURA UN EFECTO CAUSAL?

- ▶ Muchas veces podemos usar un diseño experimental para medir y que no asegure una estimación causal.
- ▶ Para que la estimación sea causal necesitamos una **asignación aleatoria sea del tratamiento**.
- ▶ Es decir, que aquello a lo que queremos encontrarle un efecto causal sea lo asignado.
- ▶ No siempre es posible realizarlo ejemplos complicados: educación, sexo, prosocialidad, etc.



# INTRODUCCIÓN A DIFERENCIAS EN DIFERENCIAS

## EXPERIMENTOS NATURALES

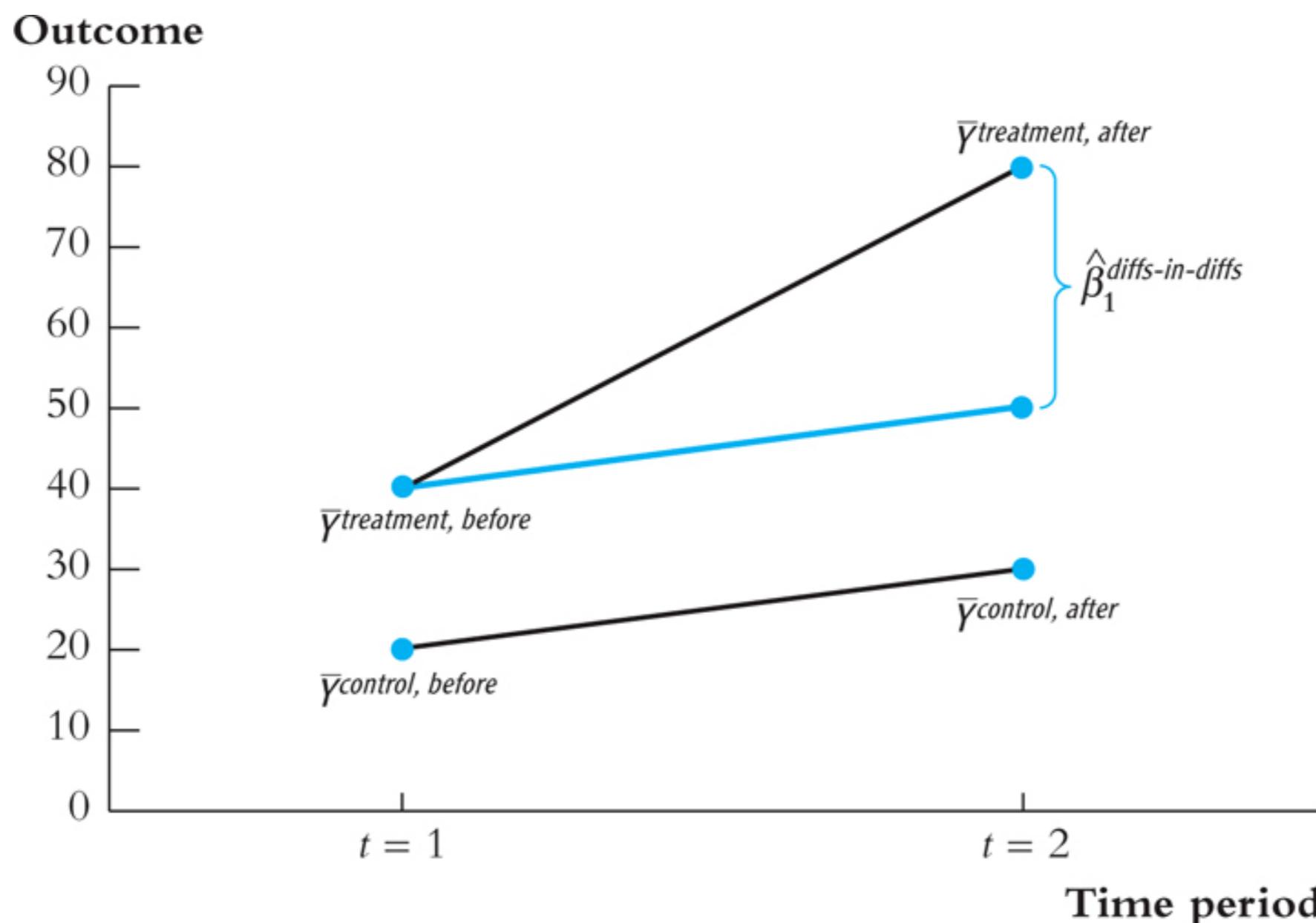
- ▶ Un **experimento natural o cuasi experimento** corresponde a algún **hecho fortuito** que genera que ciertos individuos sean expuestos a un tratamiento y a otros no.
- ▶ Dos tipos de cuasi experimentos:
  - ▶ El tratamiento es asignado como si fuera aleatorio (o aleatorio en otras variables)
  - ▶ Una variable Z que influye en el tratamiento es la que es asignada aleatoriamente (Podemos usar IV)

# ESTIMADOR DE DIFERENCIAS EN DIFERENCIAS

- ▶ Consideremos el tratamiento (D) **como si fuera** asignado aleatoriamente: MCO
- ▶ El **estimador de diferencias en diferencias** usa medidas antes y después del tratamiento de Y y estima el efecto como la diferencia entre las diferencias pre y post tratamiento para los tratados y controles.

$$\hat{\beta}_1^{diffs-in-diffs} = (\bar{Y}^{treat, after} - \bar{Y}^{treat, before}) - (\bar{Y}^{control, after} - \bar{Y}^{control, before})$$

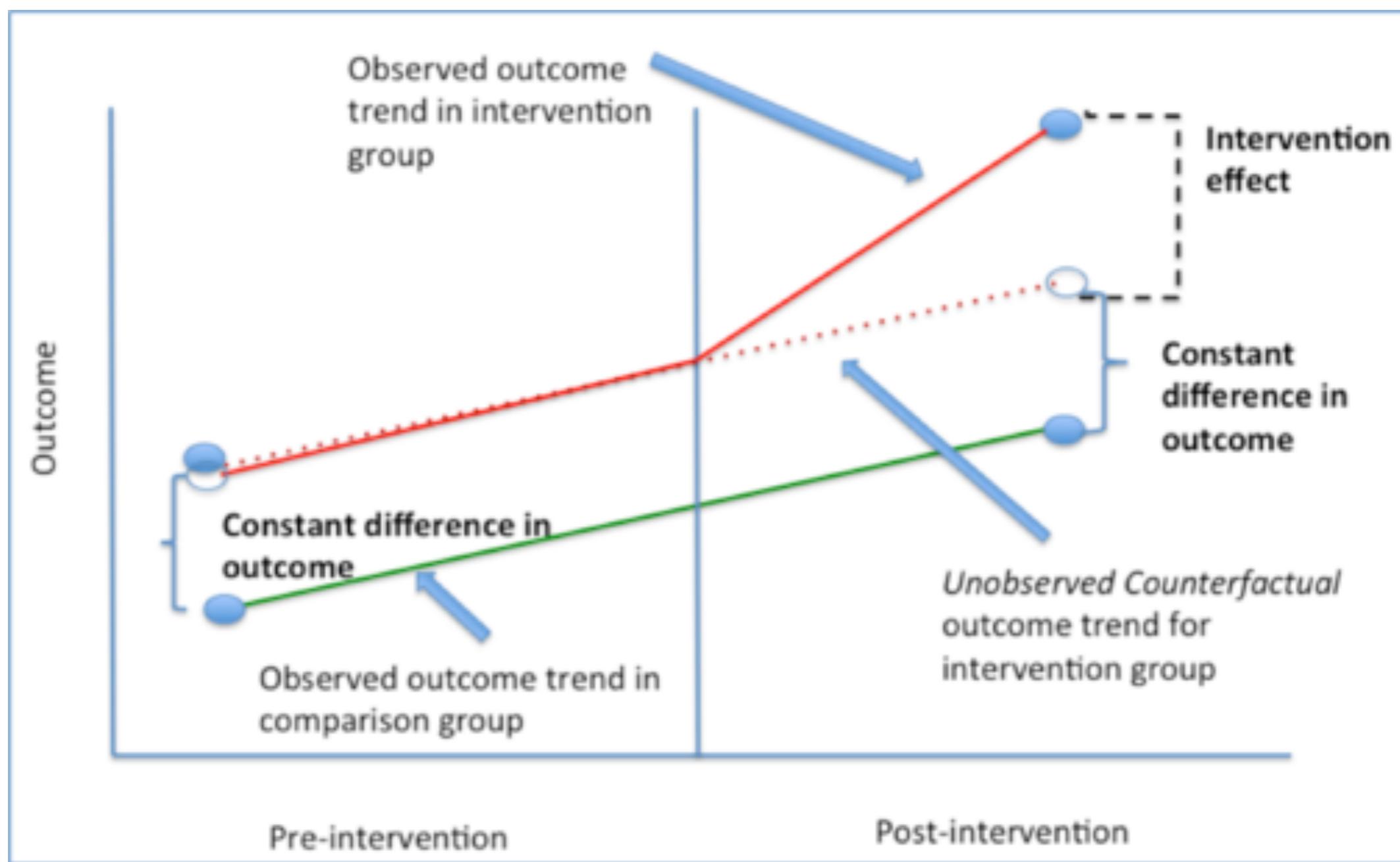
# ESTIMADOR DE DIFERENCIAS EN DIFERENCIAS



## SUPUESTO DE IDENTIFICACIÓN:

- ▶ Esta estrategia tiene un supuesto identificación clave: El cambio en el **grupo de control es un contrafactual válido** para el grupo de tratamiento.
  - ▶ Cuando el panel tiene solo 2 períodos no hay forma de decir algo con respecto al supuesto de identificación
  - ▶ Cuando el panel tiene más de 2 períodos, deberíamos al menos observar que los outcomes de grupos de control y tratamiento se mueven de forma similar *antes del tratamiento*
  - ▶ Por esta razón, a este supuesto a veces se le llama supuesto de tendencias paralelas

# TENDENCIAS PARALELAS



## ESTIMADOR DE DIFERENCIAS EN DIFERENCIAS

- ▶ Imagine que tenemos *varios cortes transversales de las mismas unidades* en el tiempo, i.e. tenemos un panel de datos
- ▶ Cuando un subconjunto de las unidades del panel recibe un tratamiento, podemos potencialmente utilizar al resto de las unidades como grupo de control dinámico
- ▶ Las regresiones más comunes tienen la siguiente forma:

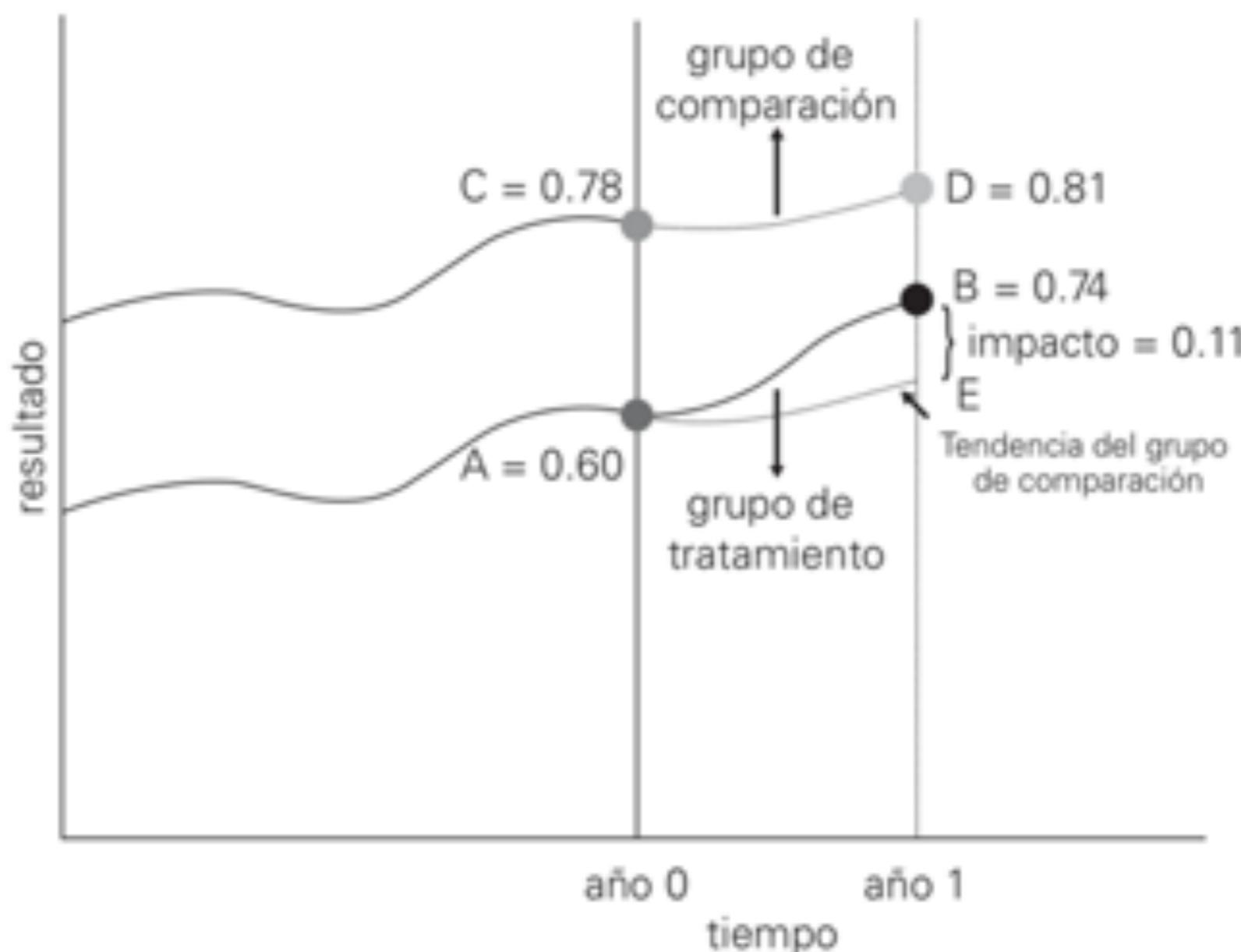
$$Y_{it} = \alpha D_i + \gamma Post_t + \beta D_i \times Post_t + \phi X_{it} + \varepsilon_{it} \quad (1)$$

$$Y_{it} = \alpha_i + \gamma_t + \beta D_i \times Post_t + \phi X_{it} + \varepsilon_{it} \quad (2)$$

$\beta$  corresponde al *estimador de diferencias-en-diferencias*

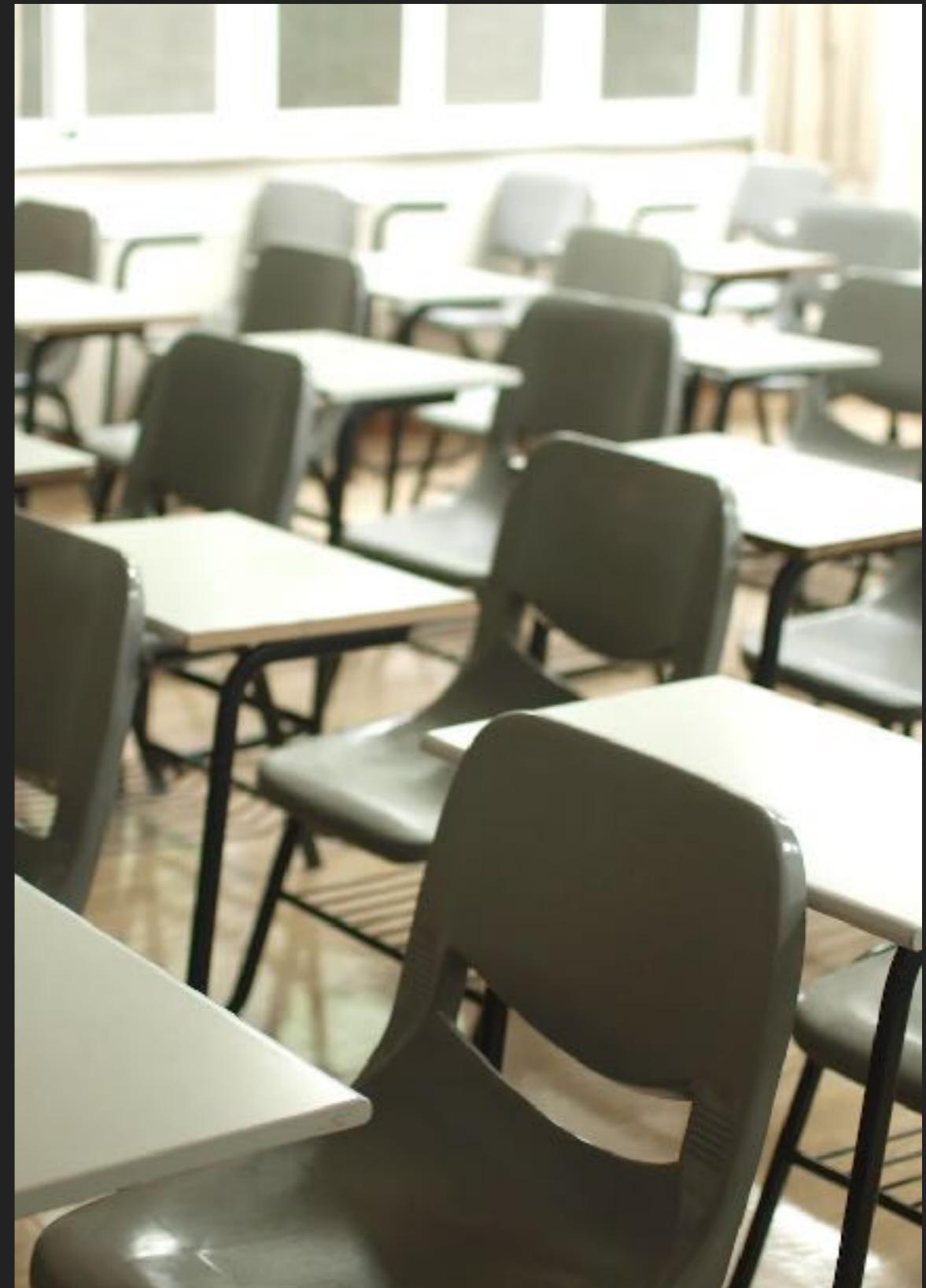
### EJEMPLO:

Gráfico 6.1 Diferencias en diferencias



# SIGAMOS NUESTRO EJEMPLO

- ▶ Considere que tiene los datos de las pruebas aplicadas a todos ANTES de que fueran al curso de verano .
- ▶ Plantee un modelo Dif-Dif y estímelo.
- ▶ Interprete los coeficientes.
- ▶ ¿Podríamos esperar que se cumpla el supuesto de tendencias paralelas?



## DIFERENCIAS EN DIFERENCIAS NO PARAMÉTRICO

- ▶ Para decir algo respecto al supuesto de tendencias paralelas, los autores típicamente estiman una regresión semiparamétrica:

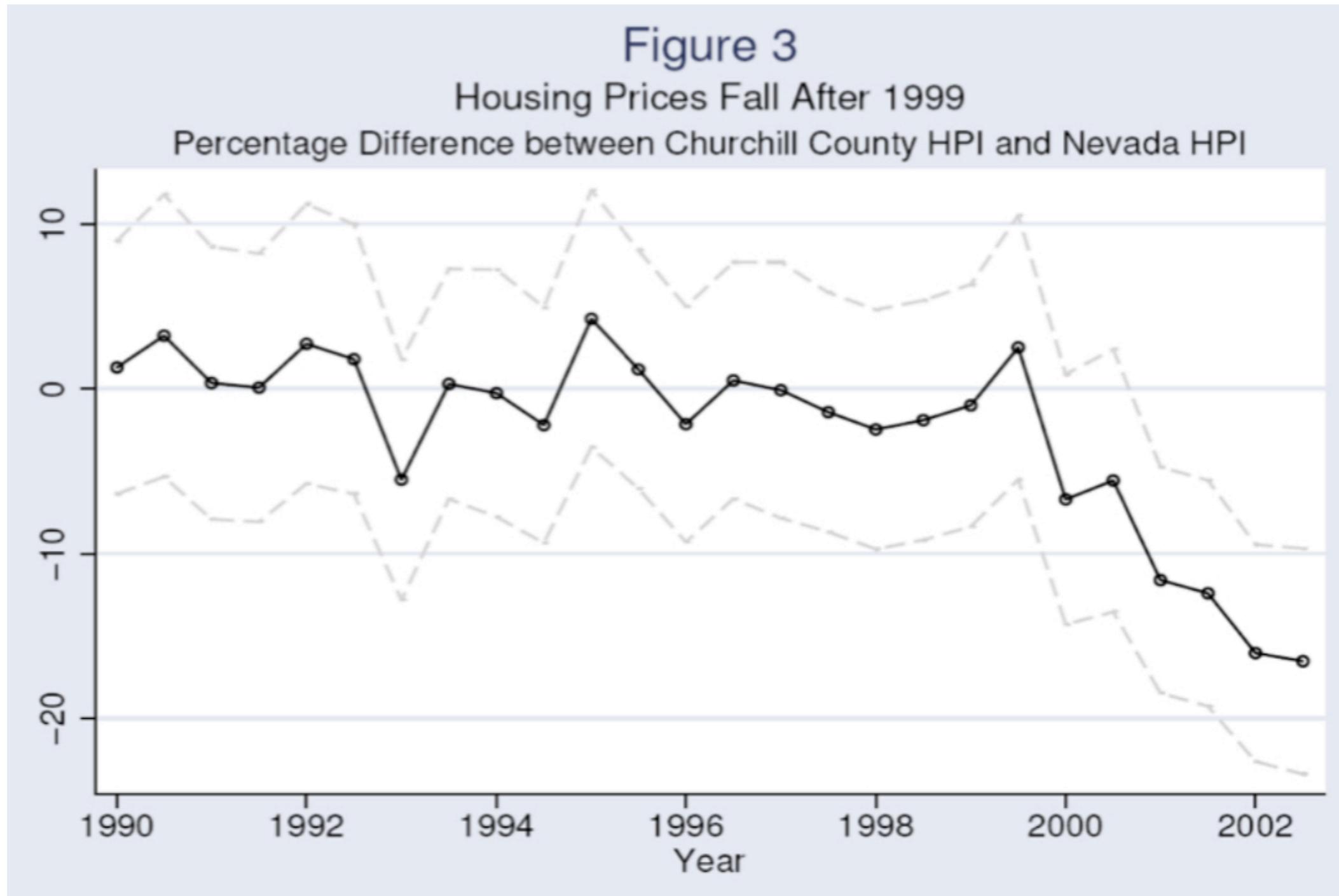
$$Y_{it} = \alpha_i + \gamma_t + \sum_{k=t_0}^{\tau} \beta_k (D_i \times T_k) + \varepsilon_{it} \quad (3)$$

donde  $\beta_k \approx 0$  cuando  $k < \tau$  y  $\tau$  es el momento del tratamiento

- ▶ Si el tratamiento tuvo algún efecto diferencial en el grupo de tratamiento, debiésemos observar que  $\beta \neq 0$  para  $k > \tau$
- ▶ Veamos un ejemplo donde el supuesto de identificación *parece* cumplirse y otro clásico ejemplo en que no se cumple...

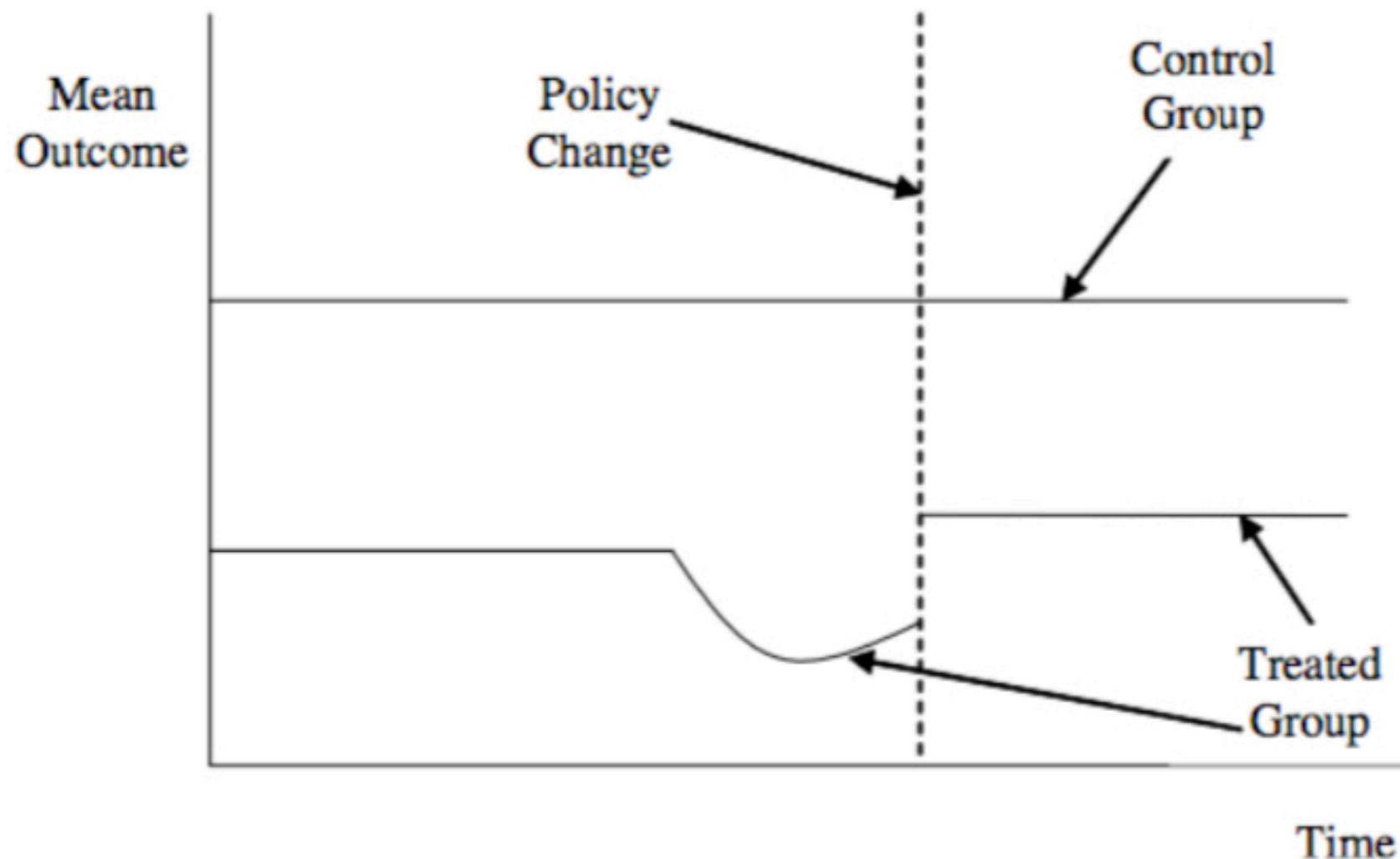
## DIFERENCIAS EN DIFERENCIAS

---



## RIESGOS A ESTA METODOLOGÍA

Figura: El “Dip de Ashenfelter”



# VARIABLES INSTRUMENTALES

---

OTROS  
MÉTODOS

### COMBATIR LA ENDOGENEIDAD

- ▶ Imagine que queremos estimar el efecto *causal* de una variable  $X$  en una variable  $Y$  mediante la siguiente regresión lineal:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (1)$$

- ▶ Lamentablemente sospechamos que la variable  $X$  es endógena y, por lo tanto, el estimador MICO no tendrá interpretación causal
- ▶ Una potencial solución para estimate el efecto causal de interés es utilizar una variable instrumental  $Z$

## COMBATIR LA ENDOGENEIDAD

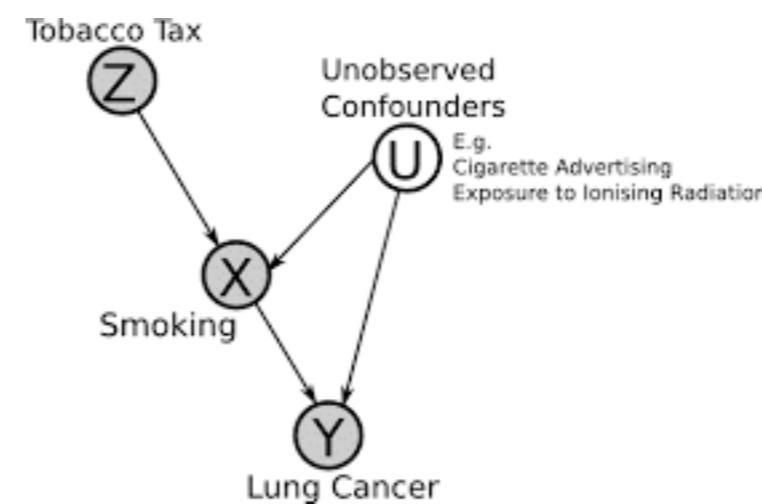
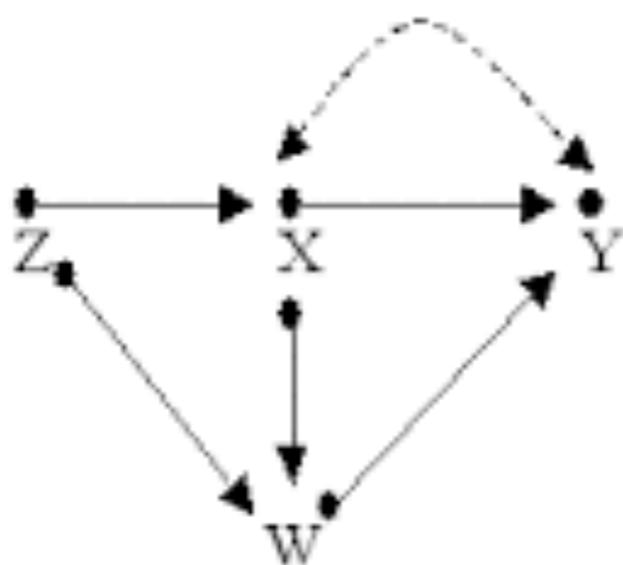
- Un caso común es sospechar que hay una variable omitida  $W$  que está correlacionada con  $X$  y que explica  $Y$ :

$$Y_i = \alpha + \beta X_i + \underbrace{\delta W_i}_{\varepsilon_i} + \eta_i \quad (2)$$

- Como no observamos  $W$ , no podemos incluirla como control y se encuentra en el término de error, lo que implica que  $E[\varepsilon_i | X_i] \neq 0$
- Dicho de otra manera, tenemos el problema que:

$$\frac{\partial Y}{\partial X} = \beta + \frac{\partial \varepsilon}{\partial X} \neq \beta \quad (3)$$

### UNA SOLUCIÓN: VARIABLE INSTRUMENTAL



- ▶ Un variable instrumental  $Z$  que cumpla con las siguientes condiciones teóricas y econométricas:
  1. Primera etapa:  $Z$  debe explicar *suficiente* variación en  $X$
  2. Restricción de exclusión:  $Z$  debe afectar a  $Y$  sólo a través de  $X$

Outcomes  
variables

y



variable  
in  
endogenous

X

$$\text{Cov}(x, z) \neq 0$$

zero  
 $\text{Corr}(x, z) = 0$

extra noise

$$x = f(z)$$

$$F > 3$$

$$E[\mu | z] = 0$$

exogenous  
inst.

## UNA SOLUCIÓN: VARIABLE INSTRUMENTAL

- ▶ Notar que bajo las condiciones anteriores podemos estimar sin problemas la primera etapa y la llamada forma reducida:

$$X_i = \nu + \gamma Z_i + \eta_i \quad (6)$$

$$Y_i = \tau + \delta Z_i + \mu_i \quad (7)$$

- ▶ Esto significa que podemos usar  $Z$  para estimar el efecto causal de interés  $\beta$  de la siguiente manera:

$$\frac{\partial Y}{\partial X} = \frac{\partial Y / \partial Z}{\partial X / \partial Z} = \frac{\delta}{\gamma} = \beta \quad (8)$$

- ▶ Aquí se ve la importancia de que  $\gamma \neq 0$

## UNA SOLUCIÓN: VARIABLE INSTRUMENTAL

- Notar que bajo las condiciones anteriores podemos estimar sin problemas la primera etapa y la llamada forma reducida:

$$\xrightarrow{\text{Primera etapa}} X_i = \nu + \gamma Z_i + \eta_i \quad (6)$$

$$\xrightarrow{\text{forma reducida}} Y_i = \tau + \delta Z_i + \mu_i \quad (7)$$

$$\xrightarrow{\text{Endog}} Y_i = \alpha + \beta X_i + \varepsilon_i \xrightarrow{\text{MC}} \hat{\beta} \text{ signo } \epsilon[\hat{\beta}] \#$$

- Esto significa que podemos usar  $Z$  para estimar el efecto causal de interés  $\beta$  de la siguiente manera:

$$\frac{\partial Y}{\partial X} = \frac{\partial Y / \partial Z}{\partial X / \partial Z} = \frac{\delta}{\gamma} = \beta \quad (8)$$

- Aquí se ve la importancia de que  $\gamma \neq 0$

### UNA SOLUCIÓN: VARIABLE INSTRUMENTAL

- ▶ Un variable instrumental  $Z$  que cumpla con las siguientes condiciones teóricas y econométricas:
  1. Primera etapa:  $Z$  debe explicar *suficiente* variación en  $X$ 
    - ▶ esta condición es econométrica y se puede chequear
    - ▶ si explica poca → problema de instrumentos débiles
    - ▶ si explica mucha →  $Z$  no difiere de  $X$  y es endógena
  2. Restricción de exclusión:  $Z$  debe afectar a  $Y$  sólo a través de  $X$ 
    - ▶ esta condición es teórica y corresponde a un supuesto
    - ▶ debe defenderse con argumentos
    - ▶ evidencia puede apoyar argumentos pero nunca probarlos

## ESTIMACIÓN:

- ▶ La estimación es trivial. En vez de estimar:

$$\hat{\beta}_{MICO} = (X'X)^{-1}X'Y \quad (9)$$

- ▶ Estimamos en cambio:

$$\hat{\beta}_{VI} = (Z'X)^{-1}Z'Y \quad (10)$$

- ▶ En el caso de tener múltiples instrumentos, estimamos;

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y \quad (11)$$

donde  $\hat{X} = \hat{\tau} + \hat{\gamma}Z$

## CONTROLES Y EFECTOS FIJOS

- ▶ Las tres regresiones tienen la misma forma al incluir controles:

$$Y_i = \alpha + \beta X_i + \omega'_1 \text{Controles}_i + \varepsilon_i \quad (12)$$

$$X_i = \nu + \gamma Z_i + \omega'_2 \text{Controles}_i + \eta_i \quad (13)$$

$$Y_i = \tau + \delta Z_i + \omega'_3 \text{Controles}_i + \mu_i \quad (14)$$

- ▶ Notar que todos los controles deben incluirse en todas las regresiones: regresión endógena, primera etapa, y forma reducida
- ▶ La misma lógica aplica con la inclusión de efectos fijos

## MULTIPLES ENDÓGENAS

- En algunas limitadas ocasiones estamos interesados en una regresión que tiene múltiples variables endógenas:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Necesitamos ahora múltiples primeras etapas e instrumentos:

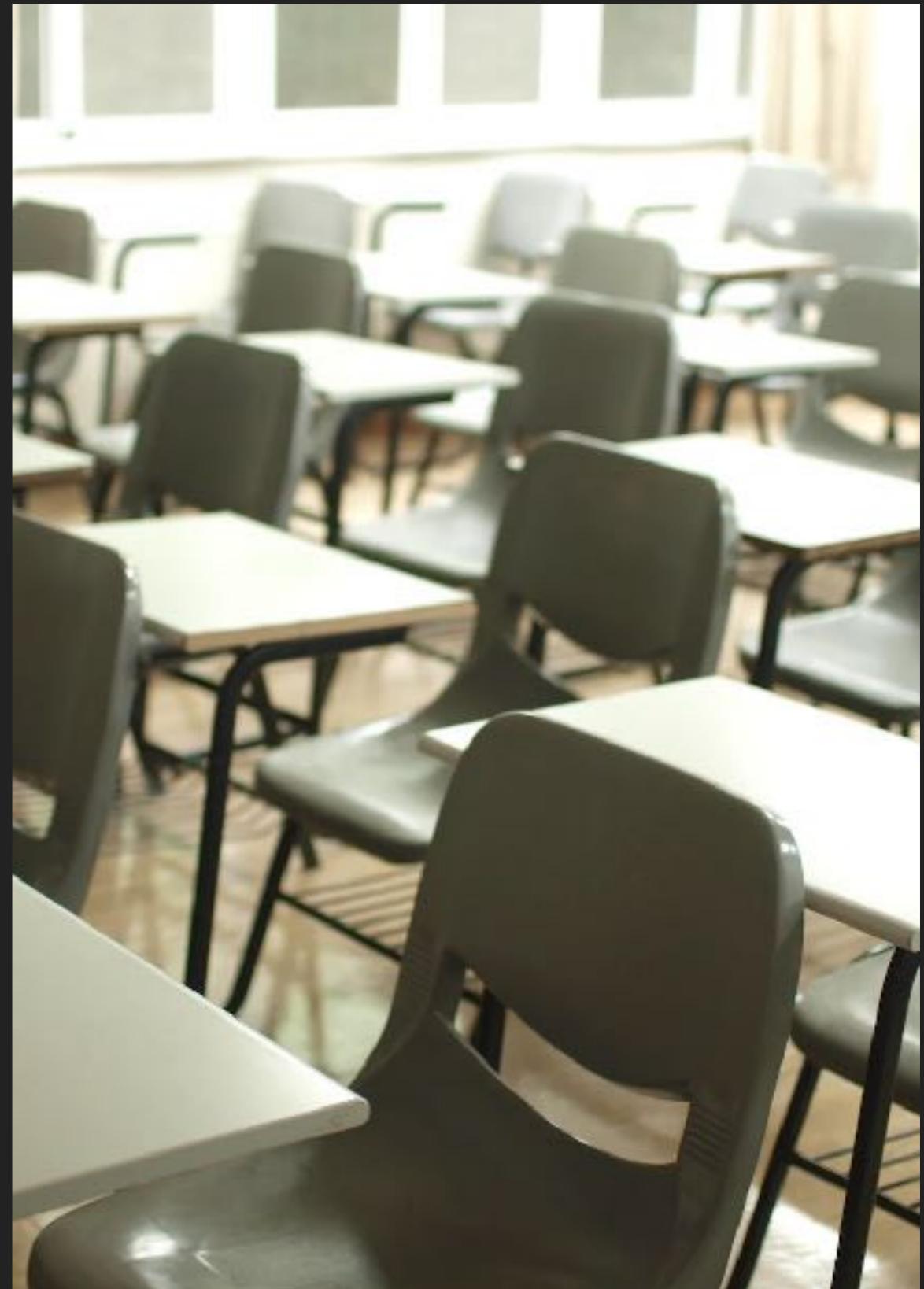
$$X_{1i} = \nu_1 + \gamma_{11} Z_{1i} + \gamma_{21} Z_{2i} + \eta_{1i}$$

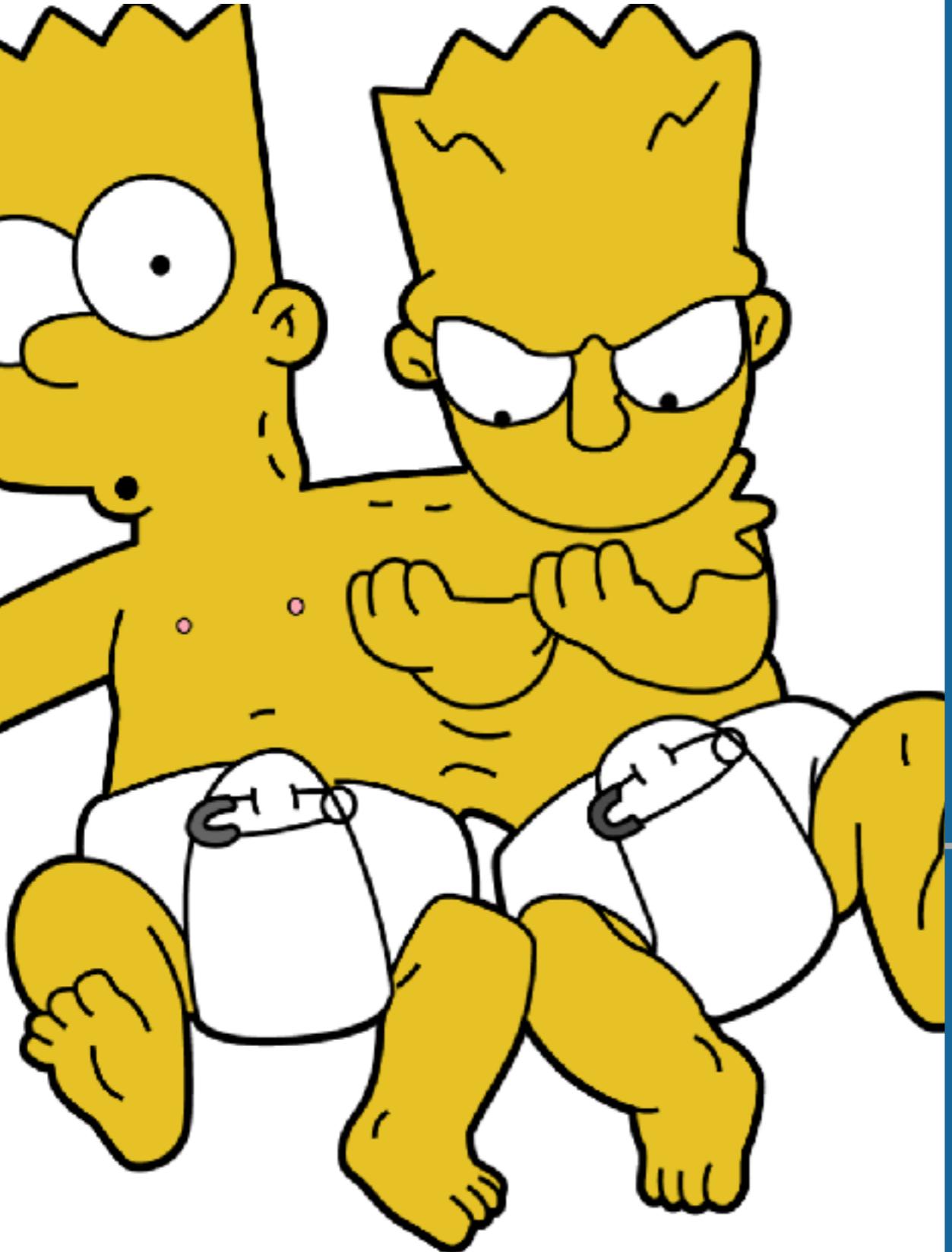
$$X_{2i} = \nu_2 + \gamma_{22} Z_{2i} + \gamma_{12} Z_{1i} + \eta_{2i}$$

- Y para identificar  $\beta$  necesitaremos múltiples condiciones que revisaremos en ayudantía con un artículo de ejemplo

# SIGAMOS NUESTRO EJEMPLO

- ▶ En realidad, nuestro interés no es el efecto causal de la carta sino más bien el impacto del curso de verano en rendimiento.
- ▶ ¿Podríamos usar la carta como un instrumento?
- ▶ Realicemos esta estimación y comparemos.





# MATCHING

---

SELECCIÓN EN  
OBSERVABLES

## MATCHING O EMPAREJAMIENTO

El emparejamiento usa grandes series de datos y técnicas estadísticas complejas para construir el mejor **grupo artificial de comparación** posible para el grupo de tratamiento.

**Gráfico 7.1 Pareamiento exacto con cuatro características**

Unidades tratadas				Unidades no tratadas			
Edad	Género	Meses desempleado	Diploma de secundaria	Edad	Género	Meses desempleado	Diploma de secundaria
19	1	3	0	24	1	8	1
35	1	12	1	38	0	2	0
41	0	17	1	58	1	7	1
23	1	6	0	21	0	2	1
55	0	21	1	34	1	20	0
27	0	4	1	41	0	17	1
24	1	8	1	46	0	9	0
46	0	3	0	41	0	11	1
33	0	12	1	19	1	3	0
40	1	2	0	27	0	4	0

# Matching

## Selección en observables

- Observamos una muestra aleatoria de una población y las siguientes variables para cada una de estas observaciones:  
 $(Y_i, T_i, X_i)$  donde  $Y$  es el outcome de interés,  $T$  un indicador de tratamiento y  $X$  covariables que no son resultado de  $T$ .
- Objetivo: estimar el efecto causal de  $T$  en  $Y$ .
- Problema:  $T$  no fue asignado aleatoriamente.
- Supuesto: la asignación de  $T$  es solo explicada por  $X$ .

**“If the covariate distributions differ substantially by treatment status, conventional regression methods can be sensitive to minor changes in the specification because of their heavy reliance on extrapolation”**

**Imbens, 2014 (Matching Methods in Practice)  
Nobel 2021**

# Controles sintéticos

- Cuando buscamos obtener el efecto de tratamiento, es clave contar con grupos no tratados.
- Seleccionar los controles es un problema cuando los grupos están muy desbalanceados (incluso si no hay sesgo de selección)
- Una opción es **construir controles sintéticos** este es, generar grupos de comparación basados en los observables.
  - Podríamos entender una regresión como que asigna igual peso a los elementos en el control sintético.

# **Frente al problema de selección...**

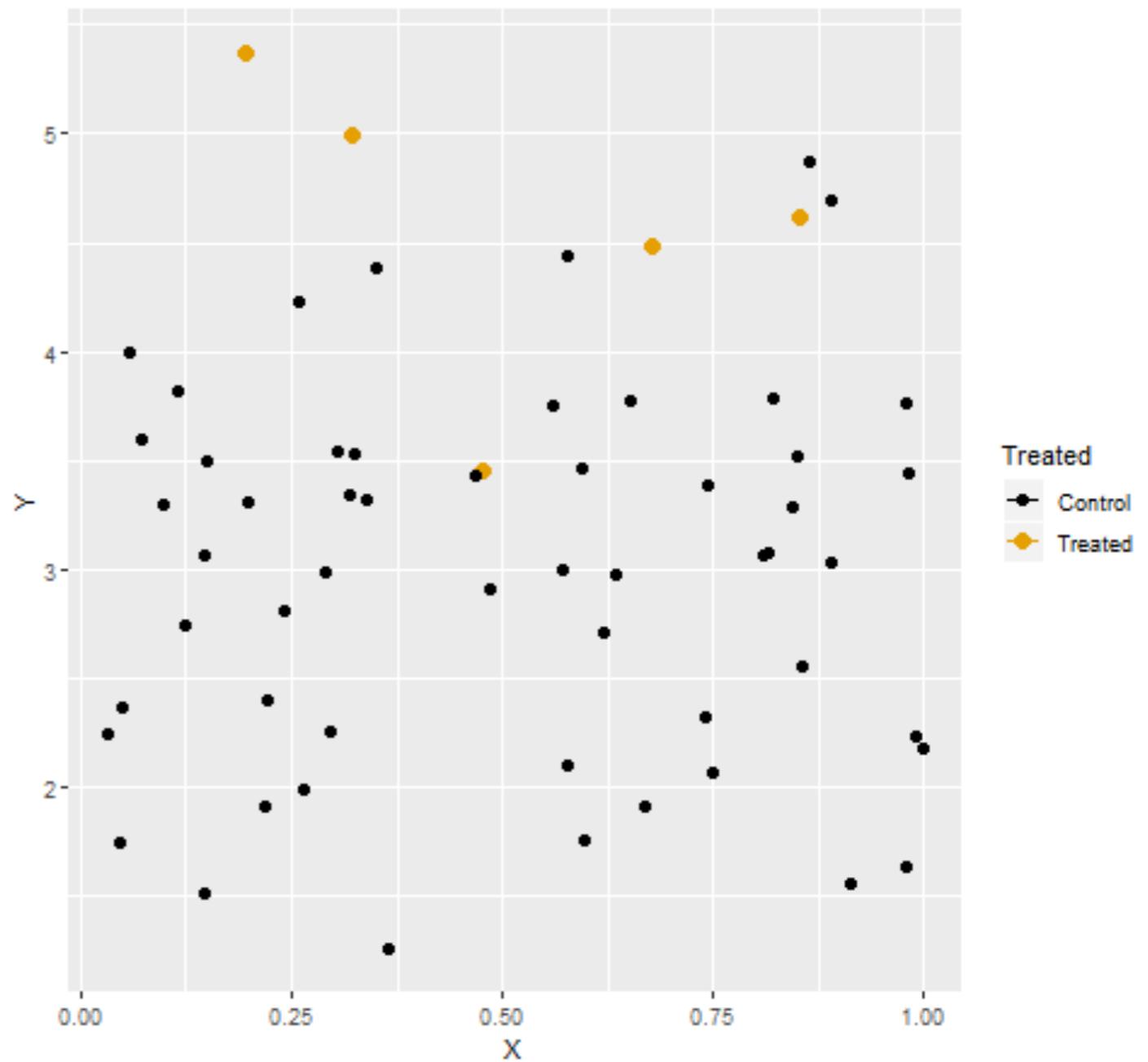
## **Controlar por ser elegido**

- Si elegir participar en un programa genera sesgo por variables omitidas en el estudio, podríamos controlar por la probabilidad de que alguien participe del programa.
- El objetivo de este método es identificar a los individuos en la muestra que son casi idénticos dado un set de observables, tal que su probabilidad de participar en el tratamiento sea igual.
- De estos, identificamos a los efectivamente tratados y a los que no y en ellos hacemos la comparación.

# Ejemplo

## Tratamiento binario

The Effect of Treatment on Y while Matching on X (with a caliper)  
1. Start with raw data.



# Buscamos al clon

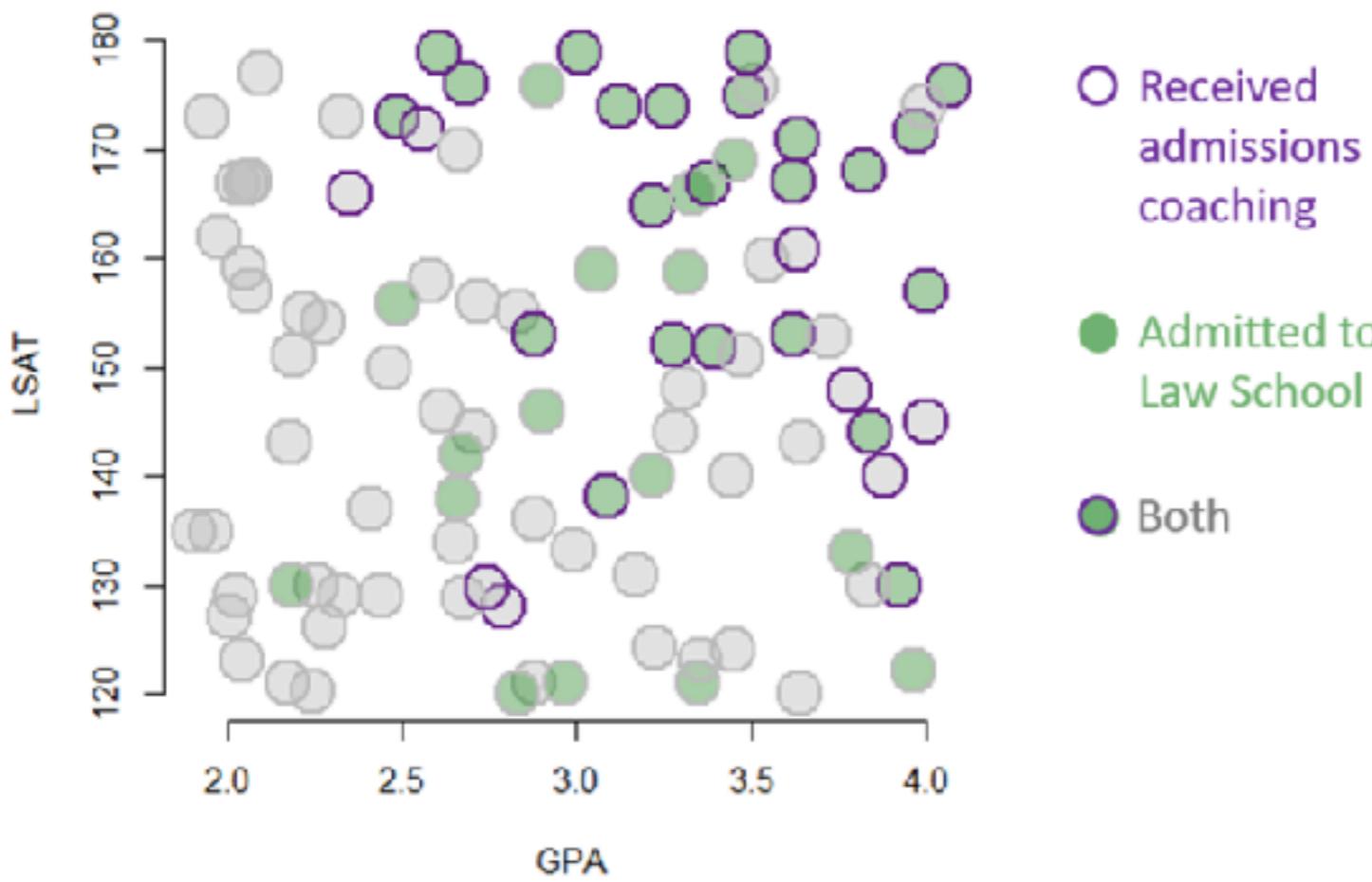


# **Ejemplo:**

## **Coaching para entrar a la escuela de leyes**

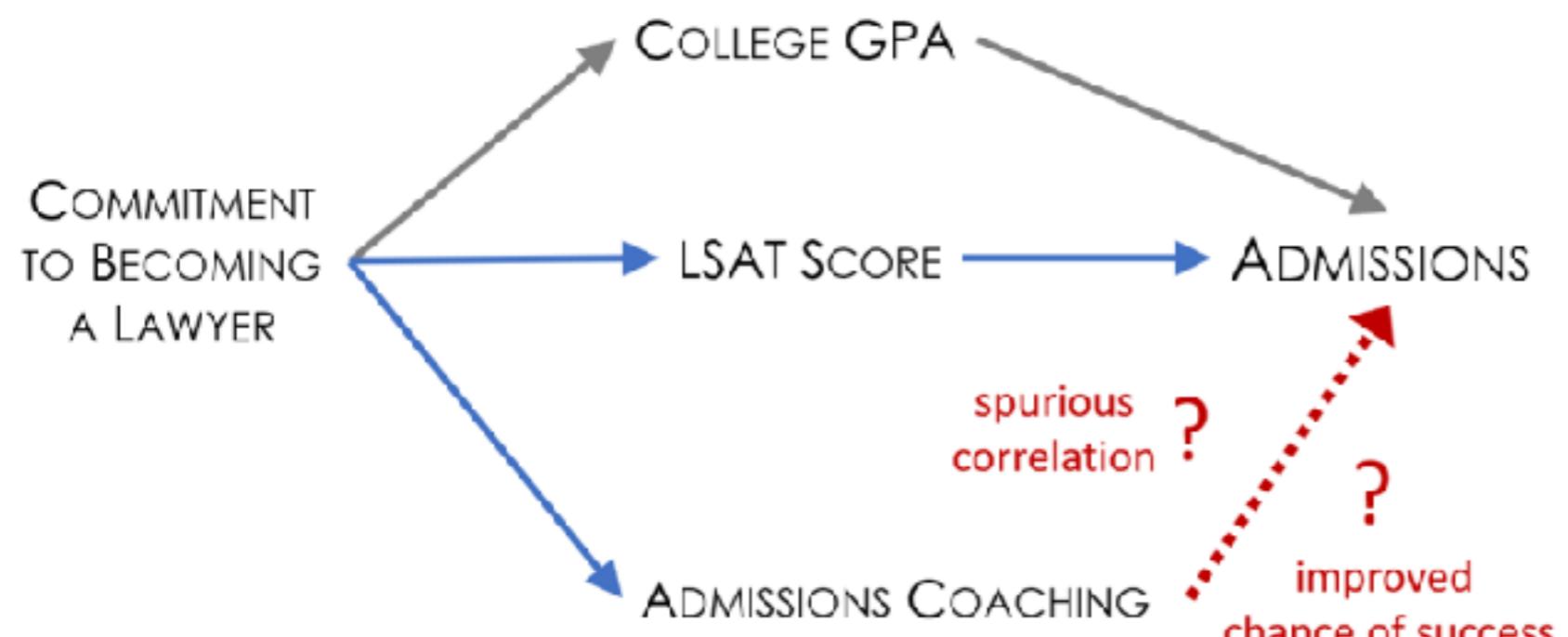
- Imaginemos que queremos evaluar un programa de coaching para entrar a la escuela de leyes.
- Claramente las personas se auto-seleccionan para tomarlo.
  - Probablemente, los que están dispuestos a pagar por un programa así también están dispuestos a hacer otras cosas.
  - Probablemente ya tienen buenas notas, buenos scores en las pruebas, etc.

# Ejemplo



- Podemos ver que hay una alta correlación entre participar en el programa de coaching y ser admitidos a la escuela de leyes.
- Si participar del coaching fuera aleatorio, esto no sería ningún problema.
- Si ademas, entrar a la escuela de leyes solo dependiera de eso tampoco seria problema.
  - Sin embargo, vemos que también se correlaciona con las notas y puntajes.

# Ejemplo



- Incluirlos en la regresión, va a solucionar algunos problemas, pero no soluciona el sesgo de selección:

Ya que no estamos controlando por la motivación.

En este caso, el grupo tratado y control son peras y manzanas.

# Ejemplo

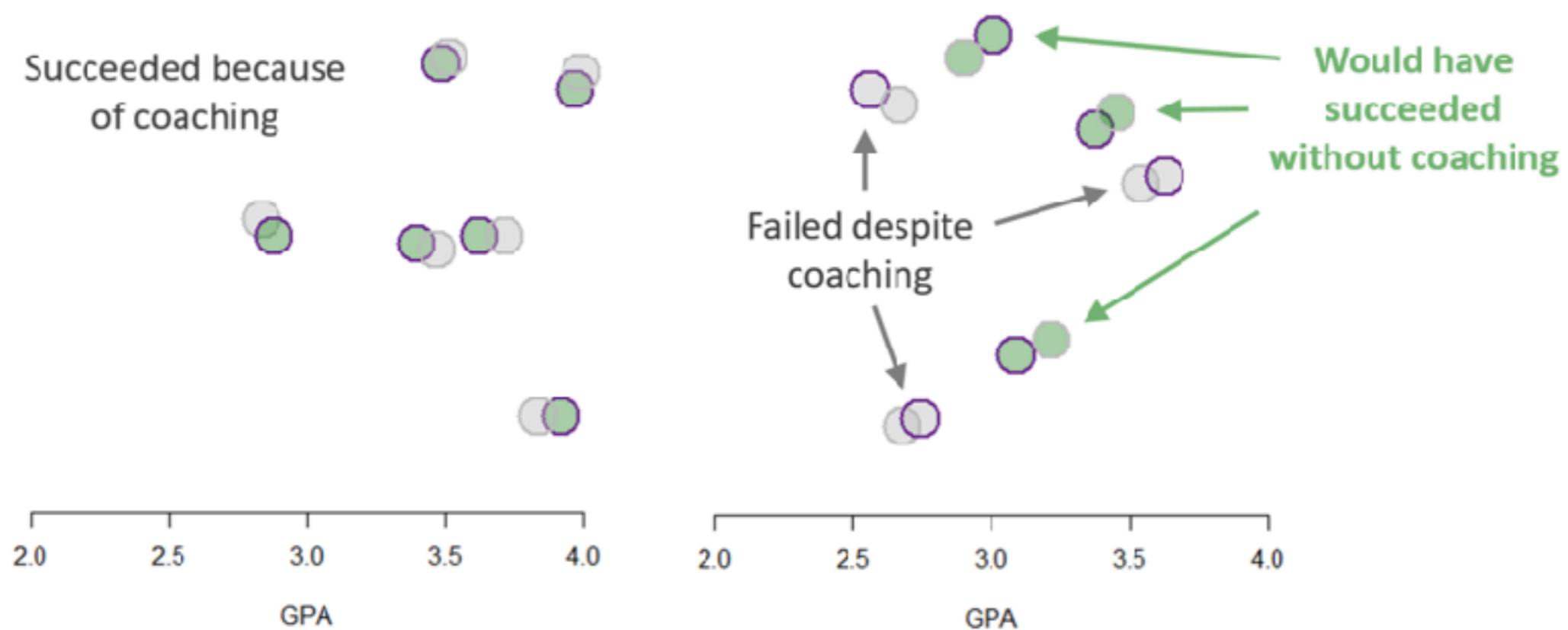
## ¿Qué hacer?

- Como los grupos tratados y control en un momento de tiempo son peras y manzana, podemos recurrir a otra técnica.
  - No podríamos hacer DiD o panel, ya que no tenemos medidas de la tasa de admisión antes y después de acceder al programa.
  - En este caso, solo tenemos post-test.
- En este contexto, matching puede ser una buena estrategia. En esta buscamos identificar “gemelos” en observables, tal que uno haya recibido el tratamiento y el otro no. En este ejemplo, solo tenemos dos controles así que es un enfoque simplista:

# Ejemplo

## ¿Qué hacer?

- Como tratamiento y outcome son binarios, podemos usar inspección visual para identificar el efecto.
- Conceptualmente, cada par es un contrafactual si creemos que la selección en observables es suficiente.



# Ejemplo

## ¿Qué hacer?

- Ahora estimamos en una regresión (en probabilidad lineal por simplicidad)

<i>Dependent variable:</i>		
	Estimados toda la data <sup>admit</sup>	Estimados solo twins
	(1)	(2)
coaching	0.34*** (0.11)	0.50** (0.19)
gpa	0.19** (0.07)	0.13 (0.22)
lsat	0.003 (0.002)	0.003 (0.01)
Constant	-0.78* (0.44)	-0.70 (1.16)
Observations	100	24
R <sup>2</sup>	0.29	0.28
Adjusted R <sup>2</sup>	0.27	0.17

Note: p<0.1; p<0.05; p<0.01

# Propensity scores

## Identificando y construyendo los controles sintéticos

- Un propensity score representa la probabilidad de ser asignado al grupo de tratamiento basado en un set de características observables.
- En este ejemplo sería la probabilidad de que un postulante participara en el servicio de coaching.
- Al emparejar observaciones en grupo tratamiento y control basados en su propensity score, podemos crear tratamientos de control artificiales que son más balanceados que un análisis directo de la data.

# Propensity score

## Construyendo controles sintéticos

- El propensity score es la probabilidad que una unidad sea tratada  $T_i = 1$  condicionada a un valor de las co-variables  $X_i = x_i$
- ¿Cómo estimamos el propensity score?

$$P(Y_i = 1 | X) = f(X_i; \Phi) + \epsilon_i$$

- Recordar que nuestras opciones son: modelo de probabilidad lineal, logit o probit. Típicamente usamos un logit y obtenemos:

$\hat{p}(X_i)$  - nuestra estimación de función de verosimilitud/probabilidad  
(Depende lo que se estime) en base a observables.

Ahora cada individuo en la muestra tiene su propio propensity score.

### SIGAMOS CON NUESTRO EJEMPLO

- ▶ Ahora estimaremos el efecto del curso de verano mediante matching.
- ▶ Estimemos por propensity score y evaluemos esta estrategia.



## Ejemplo 2:

### Evaluando resultados académicos de escuelas publicas y privadas.

- El contexto educacional es un entorno fértil para usar esta técnica, ya que rara vez se asignan los tratamientos aleatoriamente.
- Imaginemos que queremos saber si los estudiantes que asisten a colegios privados tienen mejores resultados en prueba estandarizada de matemáticas que los que fueron a colegios públicos.
- Elegir ir a un colegio privado o público en sí mismo tiene sesgo de selección.  
Supongamos que contamos con estas tres variables:
  - Income: Ingreso familiar
  - occ\_score: score de la ocupación del padre, basada en salario, educación y prestigio de 0 a 100
  - y\_educ: años de educación del padre

## Ejemplo 2:

### Evaluando resultados académicos de escuelas publicas y privadas.

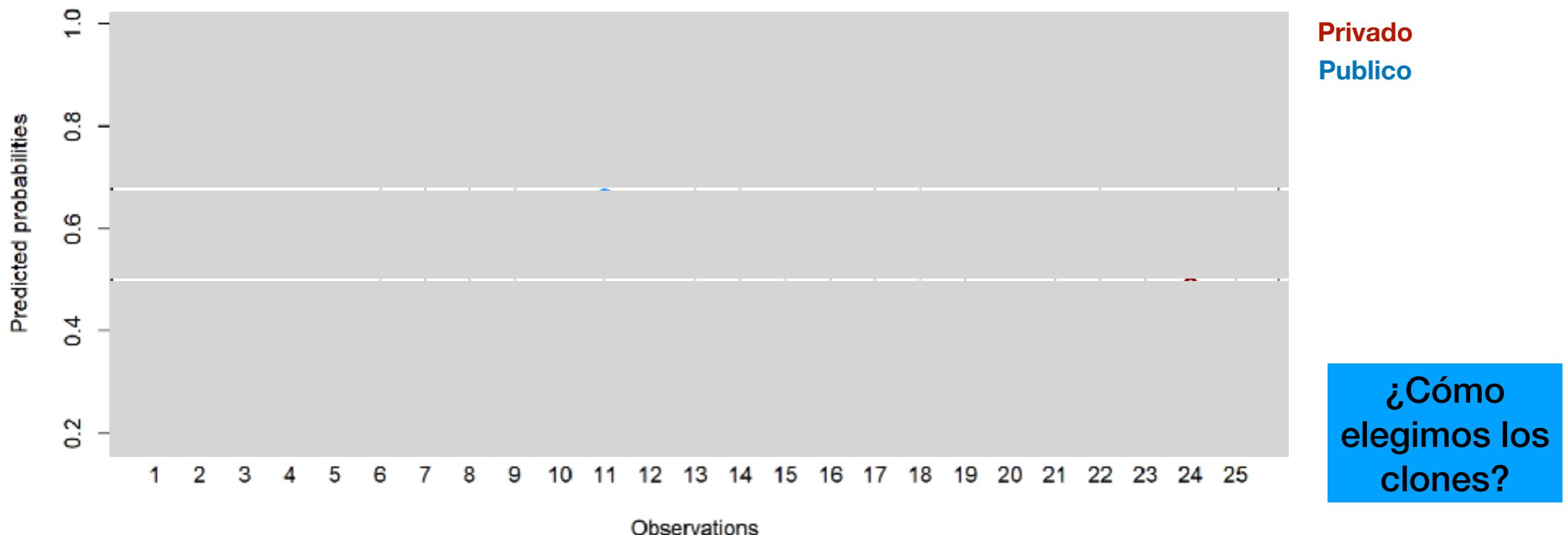
- **Creamos el matching score**
- Este score indica la probabilidad de un individuo de pertenecer al grupo de tratamiento basada en un set de covariables.
- Entonces, hacemos una regresión:

$$pr(privado = 1) = f(income, occ\_score, y\_Educ) + e$$

- Y en base a esto, obtenemos un estimado de la probabilidad, este es el propensity score.

## Ejemplo 2:

Evaluando resultados académicos de escuelas publicas y privadas.



¿Cómo  
elegimos los  
clones?

# Estimación efecto de tratamiento

- **Estimación por bloques:** particionar propensity score en  $J$  bloques y estimar regresión lineal (2) en cada bloque. Esto nos da  $J$  estimadores  $\hat{\beta}_j$ . Luego calculamos el efecto tratamiento usando:

$$\hat{\beta} = \sum_{j=1}^J \frac{N_{cj} + N_{tj}}{N} \hat{\beta}_j$$

- **Matching:** para cada unidad tratada encontramos una o varias unidades con similar  $\hat{p}(X_i)$ . Luego calculamos la diferencia entre los outcomes de estas unidades:

$$\hat{\beta} = \frac{1}{n_1} \sum_{i \in I_1} [Y_{1i} - \sum_{j \in I_1} w(i, j) Y_{0j}]$$

Opciones para match: vecino más cercano, estratificación y kernel.

# Estimación

## Uno-a-uno vs uno-a-muchos

- Lo primero que debemos definir, es si el análisis lo haremos en un match 1:1 o 1:n
  - **1:1:** En el primer caso, emparejaremos un estudiante de escuela privada a uno de escuela publica. Este método obliga a reducir fuertemente la muestra.
  - **1:n:** En este caso, quisiéramos unir un individuo en el grupo tratado (escuela privada) a un conjunto de individuos de escuela pública (control). Este segundo método permitiría tener emparejamientos menos precisos, pero permite mantener un mayor numero de muestra.

# Estimación

## Uno-a-uno vs uno-a-muchos

- Por ejemplo, observemos a los 5 con menor propensity score.
- ¿Quién usamos como su control?
- Hay varios métodos:
  - Vecino más cercano
  - estratificación
  - kernel

# Estimación

## Uno-a-uno vs uno-a-muchos

- Por ejemplo, observemos a los 5 con menor propensity score.
- ¿Quién usamos como su control?
- Hay varios métodos:
  - **Vecino más cercano**
  - estratificación
  - kernel

# Estimación

## Uno-a-uno vs uno-a-muchos

- Por ejemplo, observemos a los 5 con menor propensity score.
- ¿Quién usamos como su control?
- Hay varios métodos:
  - **Vecino más cercano**
  - estratificación
  - kernel

# **Estimación**

## **¿Con o sin reemplazo?**

- Lo siguiente es ver que ocurre con los otros casos.
- ¿Quién usamos como su control?
- Sin reemplazo indica que una vez que un estudiante es asignado como control, no puede volver a ser asignado.

# **Estimación**

## **¿Con o sin reemplazo?**

- Lo siguiente es ver que ocurre con los otros casos.
- ¿Quién usamos como su control?
- Sin reemplazo indica que una vez que un estudiante es asignado como control, no puede volver a ser asignado.

# Estimación

## Greedy and optimal process

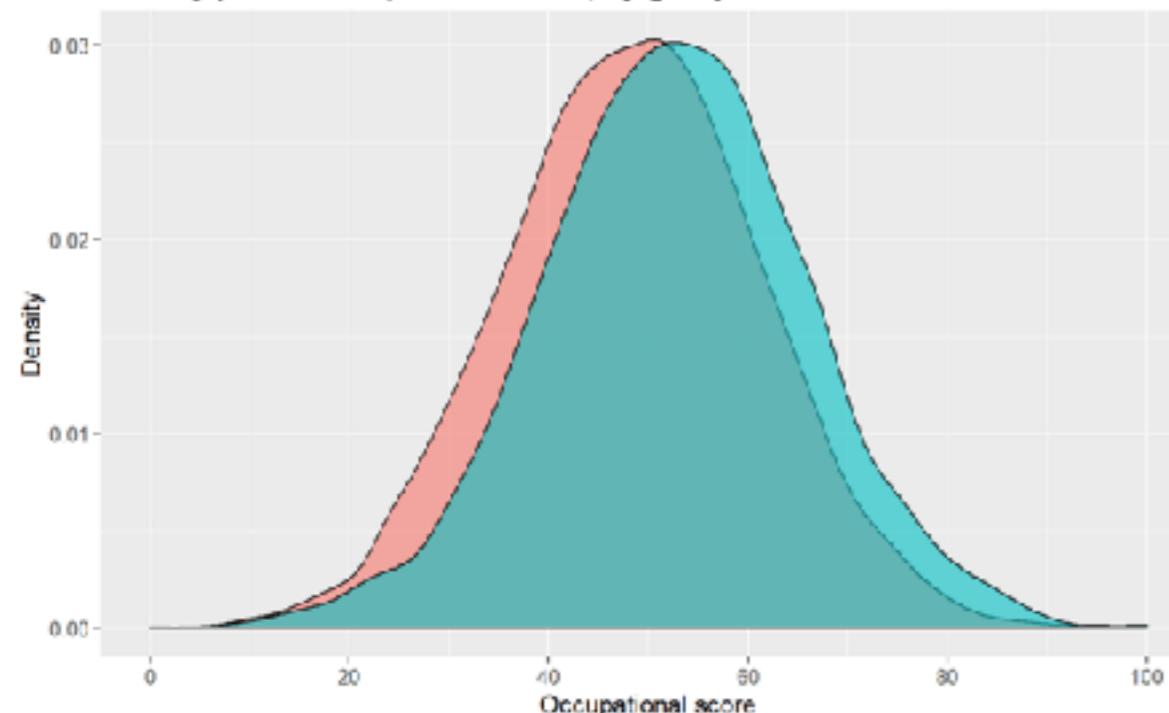
- El proceso que usemos para el matching también altera la asignación entre tratados y grupos no tratados.
- En un proceso **óptimo** primero vemos la distancia total entre las diferentes opciones. Entonces, se puede elegir el punto de partida que minimice la distancia entre dos scores.

# Controles sintéticos

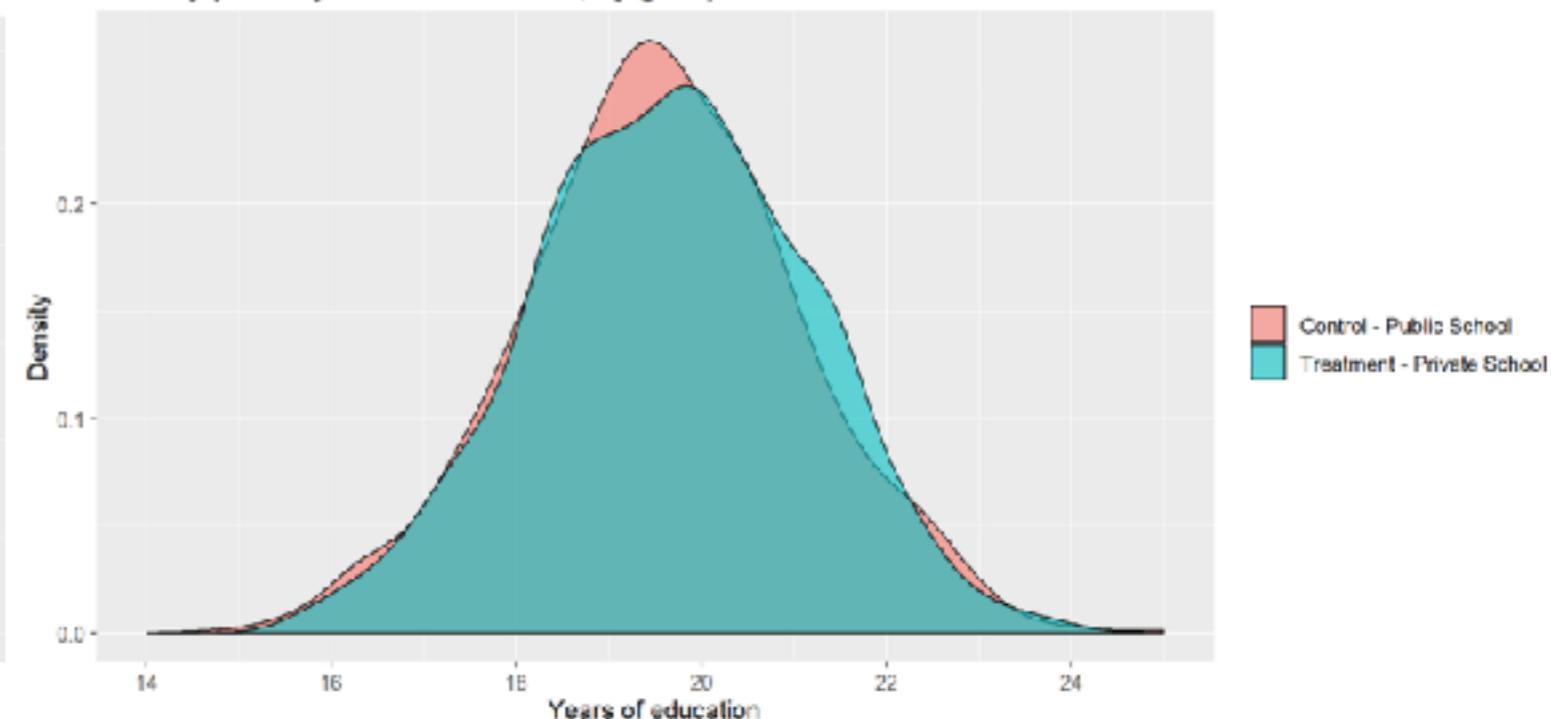
- El supuesto base es que un set con pesos de las unidades de comparación son un contrafactual válido para las unidades tratadas.
  - Selección en observables.
- Un riesgo potencial es overfitting por multicolinelaidad entre las unidades seleccionadas en el control.
- Se puede combinar con otras técnicas: Synthetic DiD
- En R se puede implementar en el paquete **MatchIt** y la función `matchit`

# Controles sintéticos

Density plot of occupational score, by group



Density plot of years of education, by group



# Calculando el efecto tratamiento

- Una vez hemos definido los controles sintéticos para cada miembro del grupo tratado, podemos simplemente hacer un test-t entre ambos o una regresión directamente.



**“(...) (My) point is that unless a particular estimator  
is robust to modest changes in implementation,  
any results should be viewed with suspicion”**

**Imbens, 2014 (Matching Methods in Practice)  
Nobel 2021**

### ESTIMEMOS NOSOTROS ESTE EJEMPLO

- Obtengamos el estimador de matching con los 3 vecinos más cercanos.
- Usemos directamente el paquete matchit.



# EFECTOS FIJOS

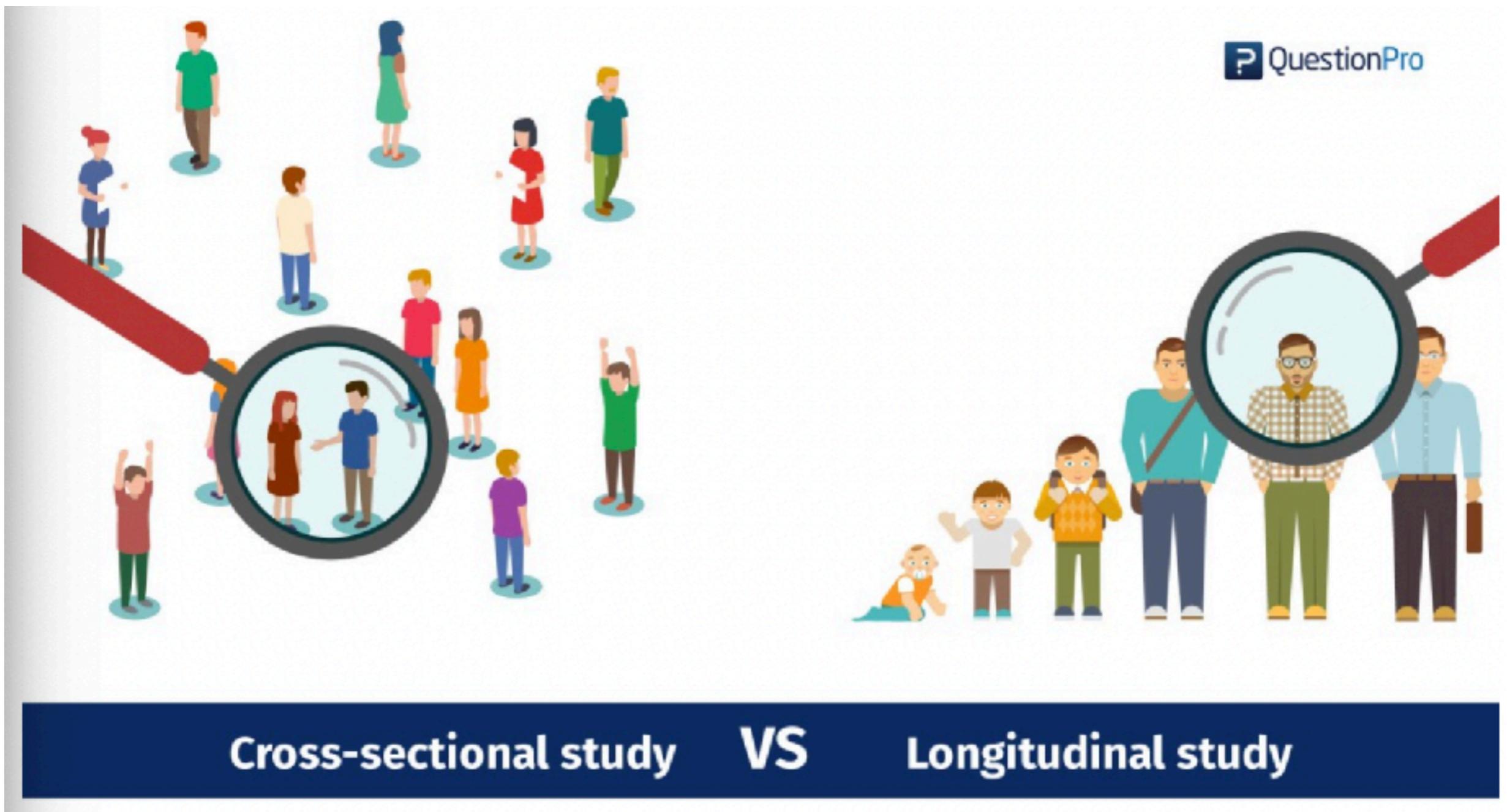
---

OTROS  
MÉTODOS

## A VECES CONTAMOS CON DATOS DE PANEL

- ▶ Hasta ahora hemos hablado principalmente asumiendo que tenemos una base de datos de corte transversal, donde observamos 1 vez a cada unidad de interés
- ▶ Sin embargo, podríamos observar 1 vez a cada unidad pero tener múltiples cortes transversales (ej. múltiples CASEN)
- ▶ En otros casos, podemos observar a la *misma* unidad de interés en múltiples ocasiones, lo que conocemos como datos de panel o datos longitudinales
- ▶ En otras ocasiones, las mismas o distintas unidades de interés se encuentran bajo diferentes categorías, dando origen a múltiples dimensiones

# APROVECHANDO LA RIQUEZA DE LOS DATOS



## EFFECTOS FIJOS

- ▶ ¿Por qué es importante la estructura de los datos? Porque nos permite determinar con mayor detalle cómo comparar unidades y absorber diferencias observables e *inobservables*
- ▶ Un efecto fijo es el coeficiente asociado a un indicador. Por ej., una regresión que incluye efectos fijos por industria  $j = 1, \dots, J$ , contiene 1 indicador por cada industria en los datos:

$$D_j = \begin{cases} 1 & \text{si } j = 1 \\ 0 & \text{si } j \neq 1 \end{cases} \quad \forall j \in \{1, \dots, J\}$$

- ▶ Otro ejemplo: un corte transversal tipo CASEN 2015 puede incluir efectos fijos por comuna, lo que implicaría:
  1. comparación dentro de cada comuna
  2. diferencias entre comunas están “controladas”

## EJEMPLO: ¿QUÉ MODELOS PODRÍAMOS CONSTRUIR ?

Cuadro: Datos de corte transversal de estudiantes

estudiante	escuela	GPA	asistencia
1	1	5,7	90
2	1	4,6	95
3	1	5,9	91
4	2	6,3	88
5	2	4,5	97
6	2	6,6	85
7	2	3,8	84

## NOTACIÓN

- ▶ Suponga que queremos estimar la *relación empírica* entre GPA ( $Y$ ) y asistencia a clases ( $X$ ) usando datos de corte transversal
- ▶ Sea los estudiantes indexados por  $i = 1, \dots, I$  y las escuelas por  $j = 1, \dots, J$ , podemos escribir esta relación empírica como:

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij} \quad (21)$$

- ▶ ¿Qué estamos comparando en esta regresión?

## NOTACIÓN

- ▶ Alternativamente, podríamos incluir efectos fijos por escuela:

$$Y_{ij} = \alpha + \beta X_{ij} + \underbrace{\gamma_1 D_1 + \cdots + \gamma_J D_J}_{\text{dummies por escuela}} + \varepsilon_{ij} \quad (22)$$

$$Y_{ij} = \alpha + \beta X_{ij} + \sum_{j=1}^J \gamma_j D_j + \varepsilon_{ij} \quad (23)$$

$$Y_{ij} = \alpha + \beta X_{ij} + \eta_j + \varepsilon_{ij} \quad (24)$$

donde  $\eta_j \equiv \gamma_1 D_1 + \cdots + \gamma_J D_J$  son los efectos fijos por escuela

- ▶ ¿Podemos incluir los efectos fijos  $\eta_j$  y  $\alpha$  en la regresión?

## LA CATEGORÍA OMITIDA

- ▶ Debido a la multicolinealidad entre los efectos fijos y la constante, tenemos que tomar una decisión
- ▶ Tenemos dos opciones: (1) eliminamos la constante, o (2) dejamos la constante pero eliminamos uno de los indicadores dentro de los efectos fijos:

$$Y_{ij} = \alpha + \beta X_{ij} + \eta_j + \varepsilon_{ij} \quad (25)$$

$$Y_{ij} = \beta X_{ij} + \eta_j + \varepsilon_{ij} \quad (26)$$

- ▶ ¿Cuál es la diferencia entre ambas opciones?

## INTERPRETACIÓN

- ▶ Considere las dos regresiones discutidas hasta ahora:

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij} \quad (27)$$

$$Y_{ij} = \beta X_{ij} + \eta_j + \varepsilon_{ij} \quad (28)$$

- ▶ ¿Qué están haciendo los efectos fijos?

1. Están absorbiendo toda la variación en  $Y$  que proviene de *las escuelas*, observables e inobservables. ¿Qué pasa si estimo?:

$$Y_{ij} = \beta X_{ij} + \delta W_j + \eta_j + \varepsilon_{ij} \quad (29)$$

2. Al absorber esta variación, pasamos de comparar (i) el GPA de alumnos con distinta asistencia, a comparar (ii) el GPA de alumnos con distinta asistencia en la misma escuela

# MULTIPLES CORTES TRANSVERSALES

Cuadro: Múltiples cortes transversales de estudiantes

estudiante	escuela	GPA	asistencia	year
1	1	5,7	90	2009
2	1	4,6	95	2009
3	1	5,9	91	2009
4	2	6,3	88	2009
5	2	4,5	97	2009
6	2	6,6	85	2009
7	1	5,3	75	2010
8	1	3,9	90	2010
9	1	4,3	98	2010
10	2	6,7	82	2010
11	2	5,1	98	2010
12	2	6,2	81	2010

## CORTES TRANSVERSALES REPETIDOS

- ▶ Hemos agregado una nueva dimensión a los datos, la cual llamaremos  $t = 2009, 2010$ . Con estos nuevos datos podemos estimar la misma regresión anterior:

$$Y_{ijt} = \beta X_{ijt} + \eta_j + \varepsilon_{ijt} \quad (30)$$

- ▶ Pero si estamos preocupados que el GPA es solo comparable dentro de una misma escuela en el mismo año, podríamos mejorar la regresión y estimar:

$$Y_{ijt} = \beta X_{ijt} + \eta_{jt} + \varepsilon_{ijt} \quad (31)$$

## IMPORTANCIA DE LA NOTACIÓN

- ▶ La notación es importante para ver cuál es la fuente de variación en los datos que se está utilizando para estimar el parámetro de interés. Considere las regresiones:

$$Y_{ijt} = \beta X_{ijt} + \eta_{jt} + \varepsilon_{ijt} \quad (32)$$

$$Y_{ijt} = \beta X_{ijt} + \eta_j + \omega_t + \varepsilon_{ijt} \quad (33)$$

- ▶ ¿Cuál es la diferencia? ¿Qué estoy comparando en una y qué estoy comparando en la otra? ¿Hay alguna de estas regresiones que prefiera por sobre la otra? ¿En qué casos? ¿Por qué?

## DATOS DE PANEL

estudiante	escuela	GPA	asistencia	year
1	1	5,7	90	2009
2	1	4,6	95	2009
3	1	5,9	91	2009
4	1	6,3	88	2009
1	1	4,5	97	2010
2	1	6,6	85	2010
3	1	5,3	75	2010
4	1	3,9	90	2010
1	1	4,3	78	2011
2	1	6,7	82	2011
3	1	5,1	79	2011
4	1	6,2	81	2011

## DATOS DE PANEL

- ▶ Los efectos fijos son particularmente útiles cuando utilizamos datos de panel porque nos permiten absorber toda la variación entre, por ejemplo, estudiantes
- ▶ En este caso, los efectos fijos estarían ayudándonos a controlar por todas las diferencias observables e inobservables entre estudiantes que no cambian en el tiempo
- ▶ La especificación más usada incluye efectos fijos por unidad, en nuestro caso un estudiante, y por tiempo, en nuestro caso un año

$$Y_{ijt} = \beta X_{ijt} + \theta_i + \omega_t + \varepsilon_{ijt} \quad (34)$$

- ▶ ¿Qué estamos comparando en este caso?

## DATOS DE PANEL

- ▶ Otras especificaciones pueden ayudarnos a chequear si hay variación a través de grupos de estudiantes en el tiempo que pueda estar confundiendo la estimación de interés:

$$Y_{ijt} = \beta X_{ijt} + \theta_i + \omega_{jt} + \varepsilon_{ijt} \quad (35)$$

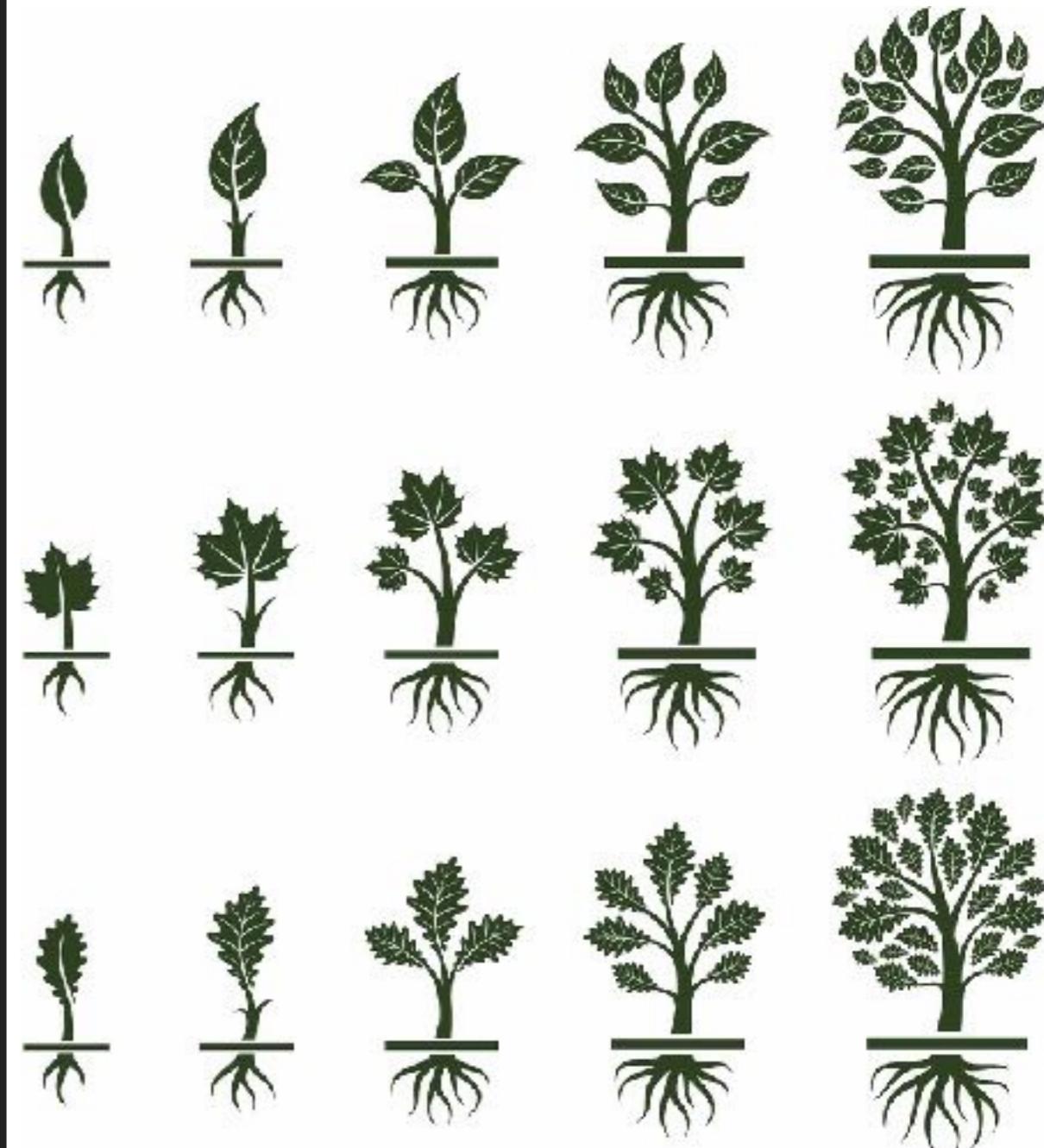
- ▶ ¿Cuál especificación prefiere si quiere estimar el efectos causal de  $X$  en  $Y$ ? ¿Una proveniente de múltiples cortes transversales o una proveniente de datos de panel?

$$Y_{ijt} = \beta X_{ijt} + \eta_{jt} + \varepsilon_{ijt} \quad (36)$$

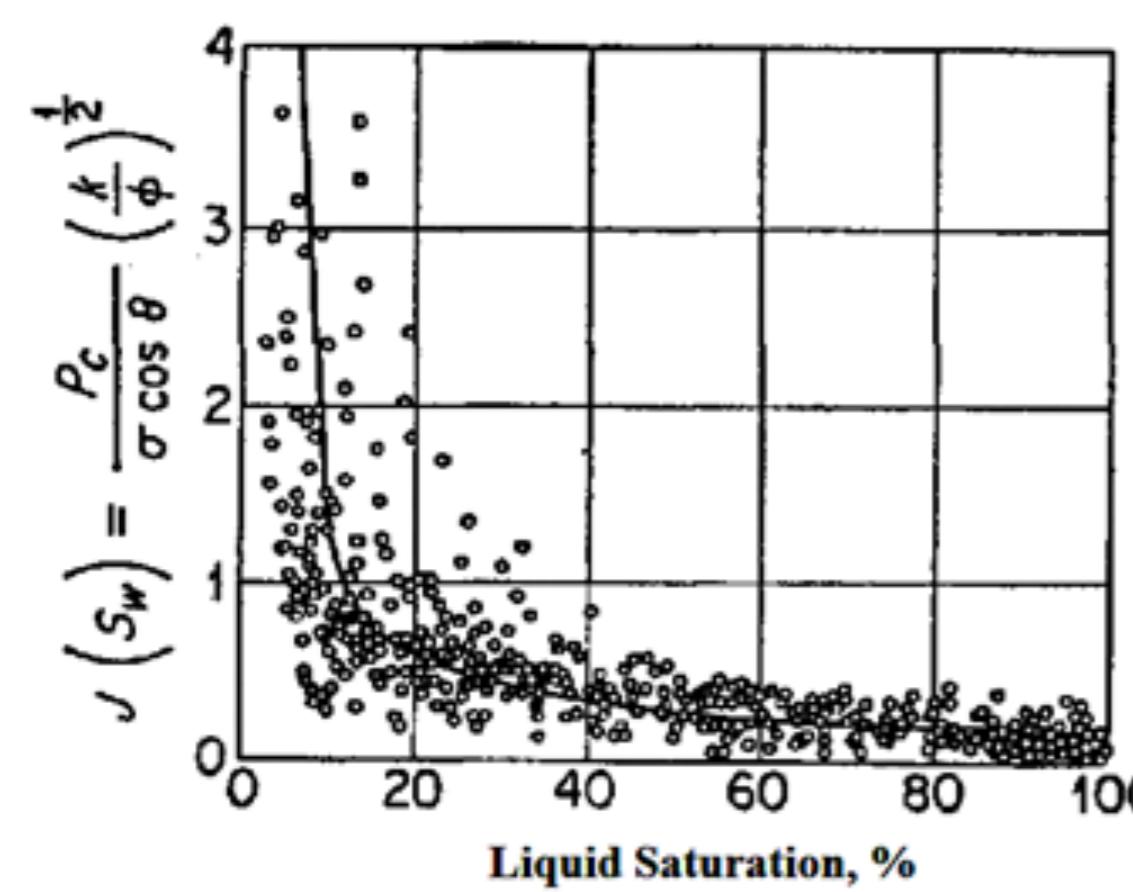
$$Y_{ijt} = \beta X_{ijt} + \theta_i + \omega_{jt} + \varepsilon_{ijt} \quad (37)$$

# AGREGUEMOS EFECTOS FIJOS

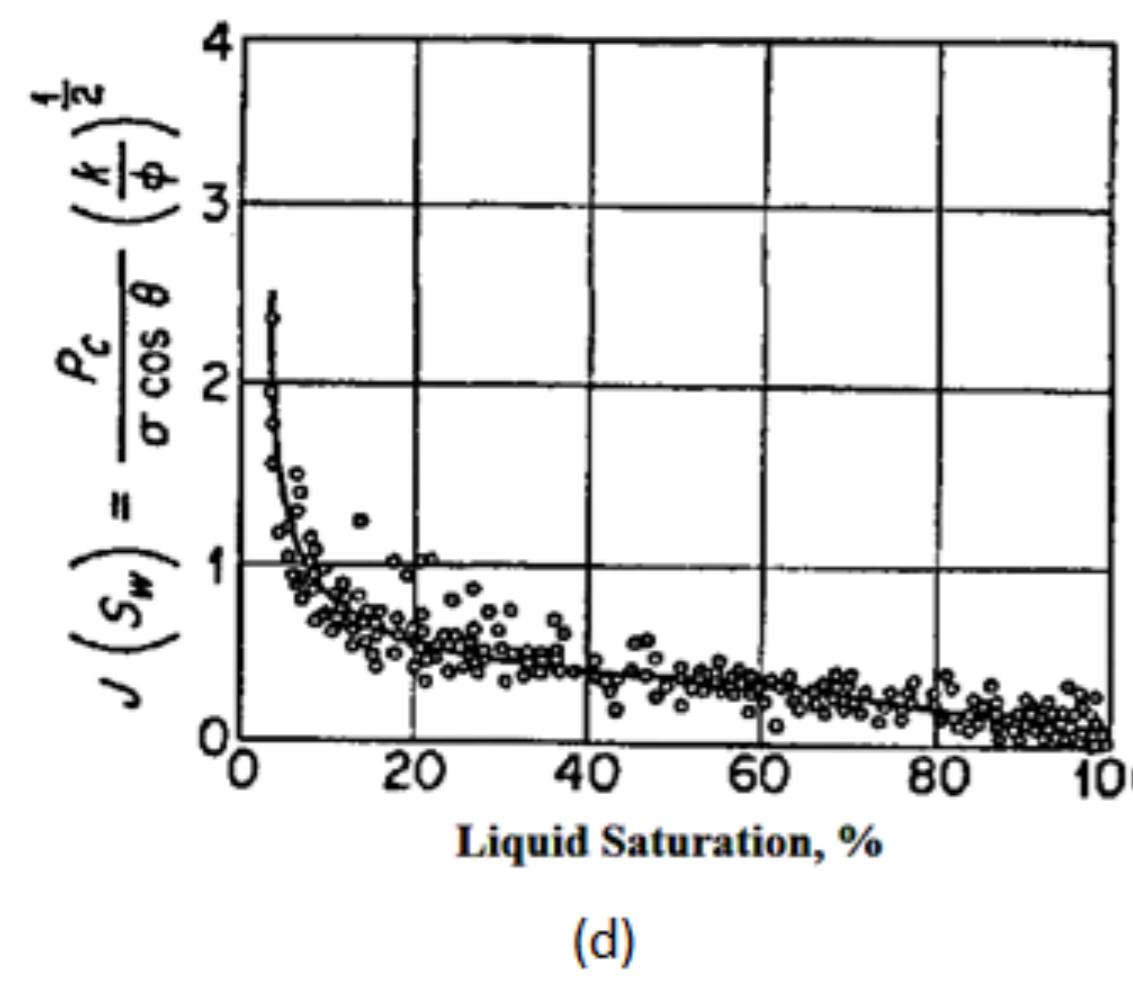
- ▶ ¿Qué elemento podríamos incluir efectos fijos en nuestro modelamiento?
- ▶ Plantee los modelos y comente las diferentes interpretaciones.
- ▶ Estime ambos efectos fijos por separados y en conjunto. Compare sus estimaciones.



# OTROS DETALLES: FORMA FUNCIONAL Y NO LINEALIDADES



(b)



(d)

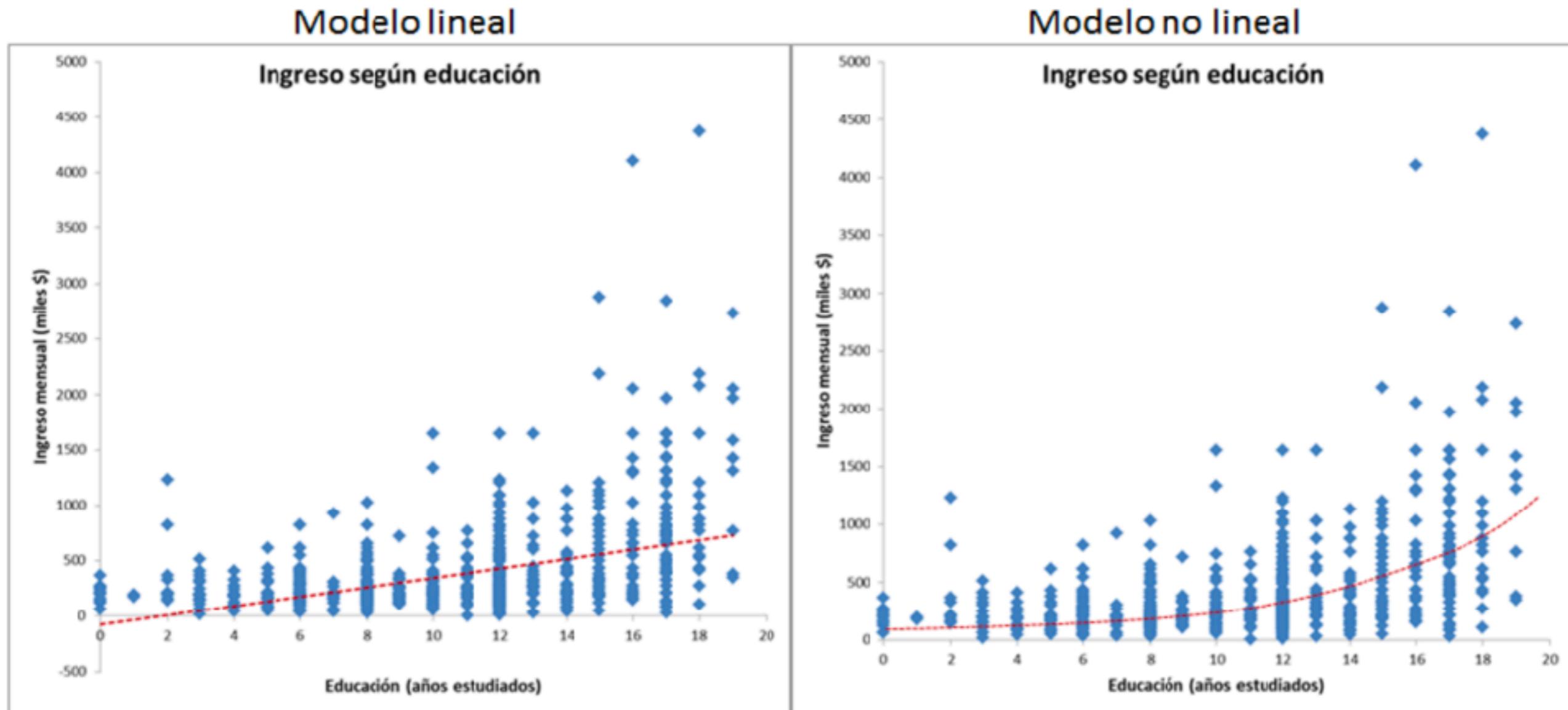
## LA MOTIVACIÓN DE UN MODELO MÚLTIPLE

- Queremos estimar el **efecto causal** de un cambio en  $X$  sobre  $Y$ .
  - La causalidad es un tema muy complejo
  - En este curso definimos un efecto causal como el efecto que se podría medir en un experimento aleatorio controlado.
- En este sentido: tiene sentido mejorar el modelo e incluir múltiples variables.
  - Enriquecer el modelo
  - Efectos no constantes y no linealidades
  - Reducir problemas de endogeneidad y sesgo

## LA IMPORTANCIA DE LA FORMA FUNCIONAL

- ¿Cómo se abordan relaciones económicas no son lineales con regresión lineal?
- Regresión lineal → lineal en los parámetros
- Se pueden transformar las variables de manera no-lineal

# LA IMPORTANCIA DE LA FORMA FUNCIONAL



$$\text{Ingreso} = -73.879 + 42.264 \text{esc}$$

$$\text{Ingreso} = 114.22 \exp(0.0837 \text{esc})$$

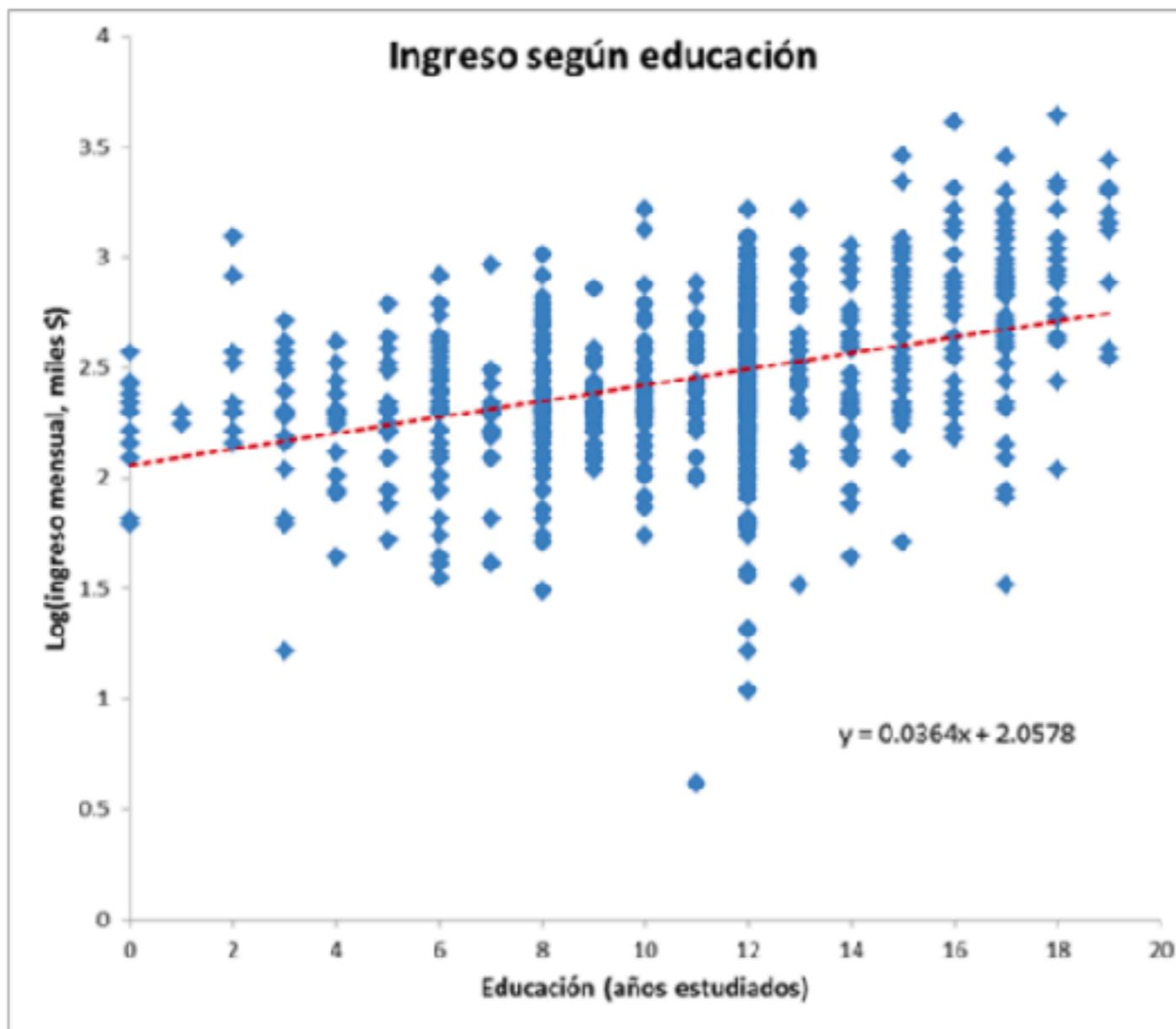
## LA IMPORTANCIA DE LA FORMA FUNCIONAL

- ¿Cómo expresar forma exponencial en regresión lineal?

$$\text{ingreso} = e^{(\beta_0 + \beta_1 \text{esc} + u)}$$

$$\log(\text{ingreso}) = \beta_0 + \beta_1 \text{esc} + u$$

# LA IMPORTANCIA DE LA FORMA FUNCIONAL



$$\widehat{\log(Ing)} = 2.0578 + 0.0364esc$$

Interpretación:

# LA IMPORTANCIA DE LA FORMA FUNCIONAL

## Forma funcional

Source	SS	df	MS	Number of obs	=	158
Model	93.1477523	1	93.1477523	F( 1, 156)	=	100.30
Residual	144.876268	156	.928694025	Prob > F	=	0.0000
Total	238.02402	157	1.51607656	R-squared	=	0.3913
				Adj R-squared	=	0.3874
				Root MSE	=	.96369

loging	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
esc	.1668695	.016662	10.01	0.000	.1339573 .1997818
cons	4.510175	.2158137	20.90	0.000	4.083879 4.936467

Log (natural)  
del ingreso  
mensual

Recordemos que:

$$\begin{aligned}\Delta \log(y) &= \widehat{\beta}_1 \Delta x; \\ 100\Delta \log(y) &= (100\widehat{\beta}_1)\Delta x; \\ \% \Delta y &= (100\widehat{\beta}_1)\Delta x; \\ \frac{\% \Delta y}{\Delta x} &= 100\widehat{\beta}_1\end{aligned}$$

1 año adicional de escolaridad  $\Leftrightarrow$  16.7% más ingreso.

# MODELOS CON LOGARITMOS

## Resumen de las formas funcionales en las que se emplean logaritmos

Modelo	Variable dependiente	Variable independiente	Interpretación de $\beta_1$
Nivel-nivel	$y$	$x$	$\Delta y = \beta_1 \Delta x$
Nivel-log	$y$	$\log(x)$	$\Delta y = (\beta_1 / 100) \% \Delta x$
Log-nivel	$\log(y)$	$x$	$\% \Delta y = (100 \beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 \% \Delta x$

## EFFECTOS NO CONSTANTES

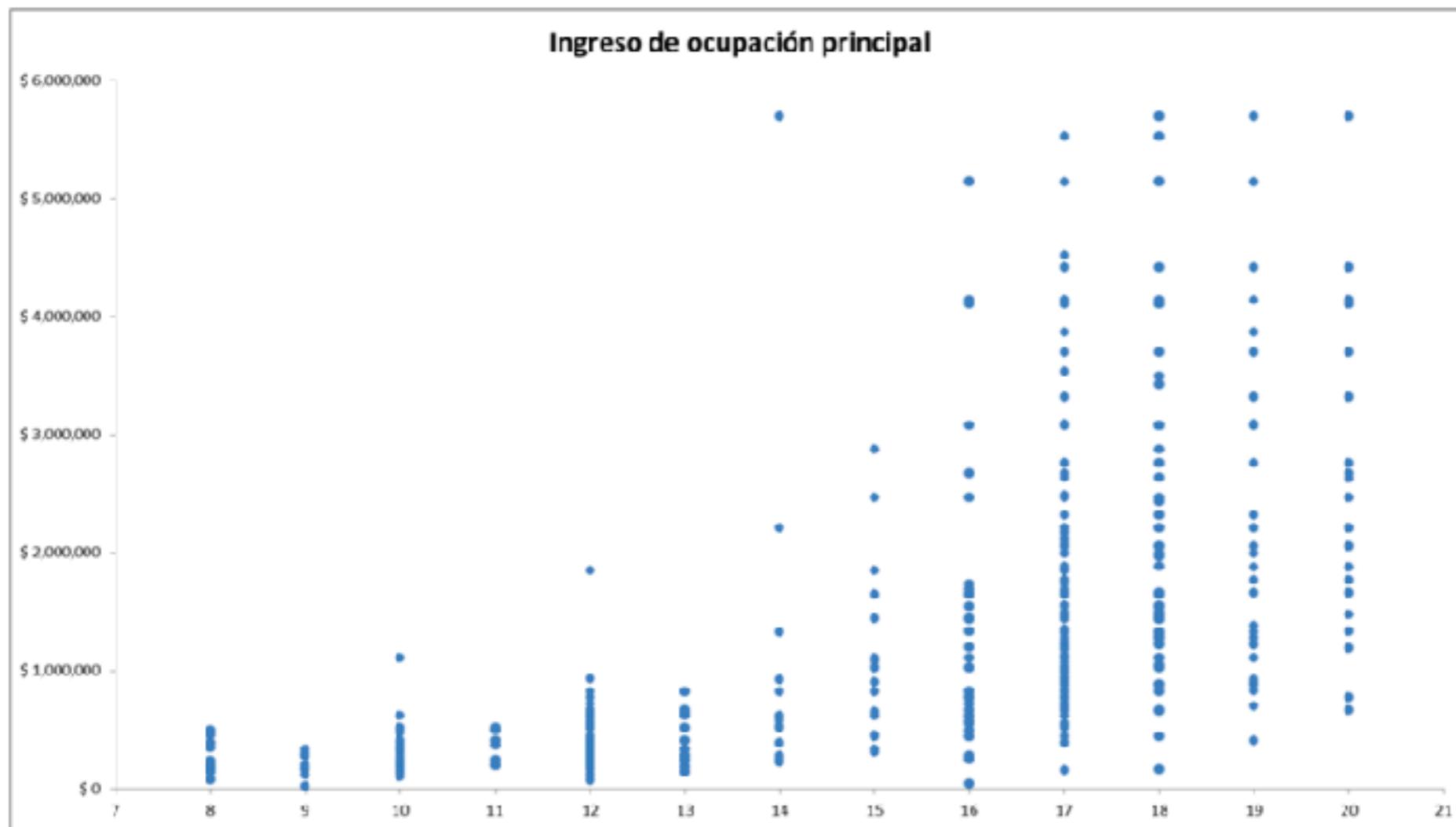
- También podemos tratar de explicar mejor los datos, incluyendo efectos no constantes:

$$\text{salario} = \beta_0 + \beta_1 \text{educacion} + \beta_2 \text{educacion}^2 + u$$

- ¿Cómo se ve ahora el efecto de la educación en el salario?
- Esto no rompe el supuesto de linealidad: el modelo debe ser lineal en los coeficientes  $\beta_j$  no en las variables  $x_j$

# EFFECTOS NO CONSTANTES

- Consideremos el caso del salario y educación....





# INFORMACIÓN CUALITATIVA

## INFORMACIÓN CUALITATIVA

Muchas veces queremos incorporar al análisis información que no es de carácter numérico, sino que responde a alguna característica o calidad del sujeto:

- Variable binaria: toma 2 valores solamente
  - Ejemplo: sexo (h,m), zona (urbano,rural)

Variable categórica: puede tomar más de 2 valores

- Ejemplo: región, industria, ocupación...

¿Cómo analizamos esta información?

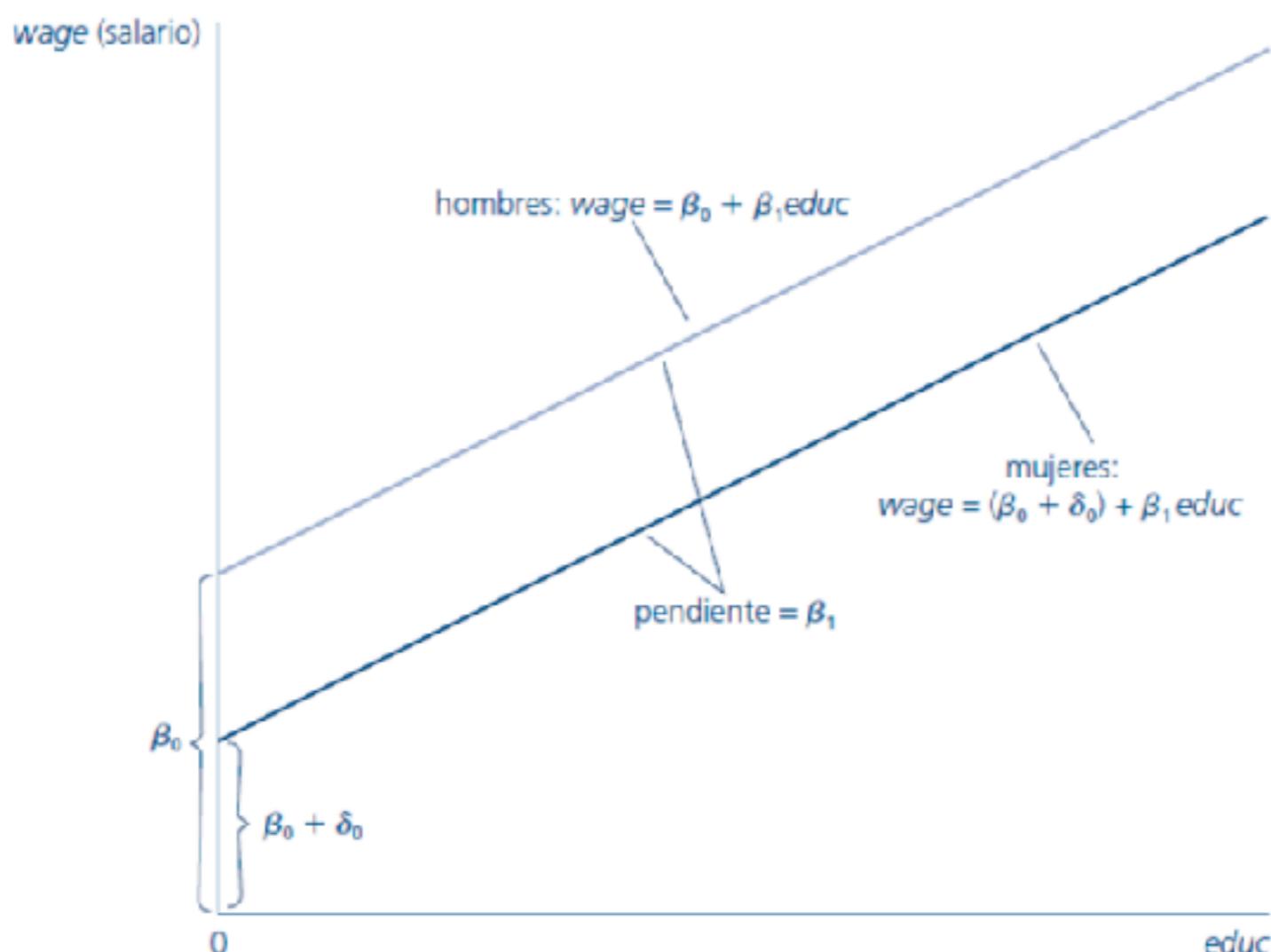
## INFORMACIÓN CUALITATIVA – BINARIAS

### Variable binaria (o “dummy”)

- Útil si se re-define como variable 0-1
- Definir como 1 la categoría de interés
- Es la idea de un “interruptor”
- Ejemplo:  $\text{sexo} = 1$  si hombre, 2 si mujer (Redefinir: mujer = 1 si mujer, 0 si hombre)

# INFORMACIÓN CUALITATIVA - BINARIAS

Gráfica de  $wage = \beta_0 + \delta_0 female + \beta_1 educ$  en la que  $\delta_0 < 0$ .



## ¿Cómo interpretamos el beta?

Es la diferencia en Y, ceteris paribus que compara al grupo con la dummy “activa” versus al grupo de referencia.

## INFORMACIÓN CUALITATIVA

### Variable Categórica

- No podemos simplemente transformarla en un número, ya que el beta representa el cambio.
- Podemos definir múltiples dummy's
- Es la idea de un “interruptor”
- Cuidado con la **multicolinealidad perfecta**

### INFORMACIÓN CUALITATIVA

Algunas veces, queremos saber como tener 2 cualidades a la vez afecta a un individuo. O como es que una variable cambia para dos grupos, en estos casos podemos usar **interacciones**.

Lo veremos con un par de ejemplos:

$$\log(wage) = \beta_0 + \gamma_0 female + \beta_1 educ + \gamma_1 female \times educ + u$$

---

¿ Y SI MI VARIABLE  
INDEPENDIENTE ES  
CUALITATIVA?

## MODELO DE PROBABILIDAD LINEAL

Variable dependiente toma dos valores (0 ó 1)

- Ejemplos: terminó educación secundaria, fue madre adolescente, trabaja o no, ha recibido capacitación laboral, etc.
- $\beta$ 's no tienen la misma interpretación: y solo cambia de 0 a 1
- Modelo de probabilidad lineal:

$$pr(y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

- Coeficientes:

$$\Delta pr(y = 1|X) = \beta_j \Delta X_j$$

## MODELO DE PROBABILIDAD LINEAL

Un ejemplo posible es estudiar la participación laboral de una mujer.

Definimos nuestra variable de interés:

$$participa = \begin{cases} 1 & \text{si 1 si trabajó o buscó trabajo al menos 1 mes durante los últimos 12 meses} \\ 0 & \text{si no} \end{cases}$$

Nuestra regresión es:

$$participa = \beta_0 + \beta_1 pareja + \beta_2 educ + \beta_3 edad + \beta_4 edad^2 + \beta_5 hijo_{0-5} + \beta_6 hijo_{6-13} + \beta_7 hijo_{14-17} + u$$

## MODELO DE PROBABILIDAD LINEAL

Variable	Obs	Mean	Std. Dev.	Min	Max
participa	6287	0.542	0.498	0.00	1.00
pareja	6287	0.579	0.494	0.00	1.00
educ_years	6287	10.18	4.02	0.00	22.00
edad	6287	46.21	11.48	25.00	69.00
edadsq	6287	2267.38	1082.18	625.00	4761.00
hijo_0_5	6287	0.133	0.381	0.00	3.00
hijo_6_13	6287	0.384	0.642	0.00	5.00
hijo_14_17	6287	0.320	0.569	0.00	4.00

*Antes de estimar una regresión, es importante conocer los datos...en Chile en el año 2009, 54.2% de las mujeres entre 25-69 años participó en el mercado laboral; 57.9% de ellas vivía con una pareja; tenían 10.2 años de educación y 46.2 años de edad en promedio; 13.3% tenía un hijo en edad 0-5 años, 38.4% tenía un hijo en edad 6-13 años, y 32% tenía un hijo en edad 14-17 años.*

# MODELO DE PROBABILIDAD LINEAL

Participación laboral de mujeres en Chile (25-69 años)	
VARIABLES	(1)
pareja	-0.163*** (0.0119)
educ_years	0.0314*** (0.0016)
edad	0.0276*** (0.0043)
edad_sq	-0.000389*** (0.0000)
hijo_0_5	-0.0846*** (0.0163)
hijo_6_13	-0.0561*** (0.0098)
hijo_14_17	-0.00641 (0.0107)
Constant	-0.0417 (0.0979)
Observations	6,287
R-squared	0.175
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

*La probabilidad de participar en el mercado laboral fue 0.163 más baja para mujeres que vivían con una pareja vs. mujeres sin pareja.*

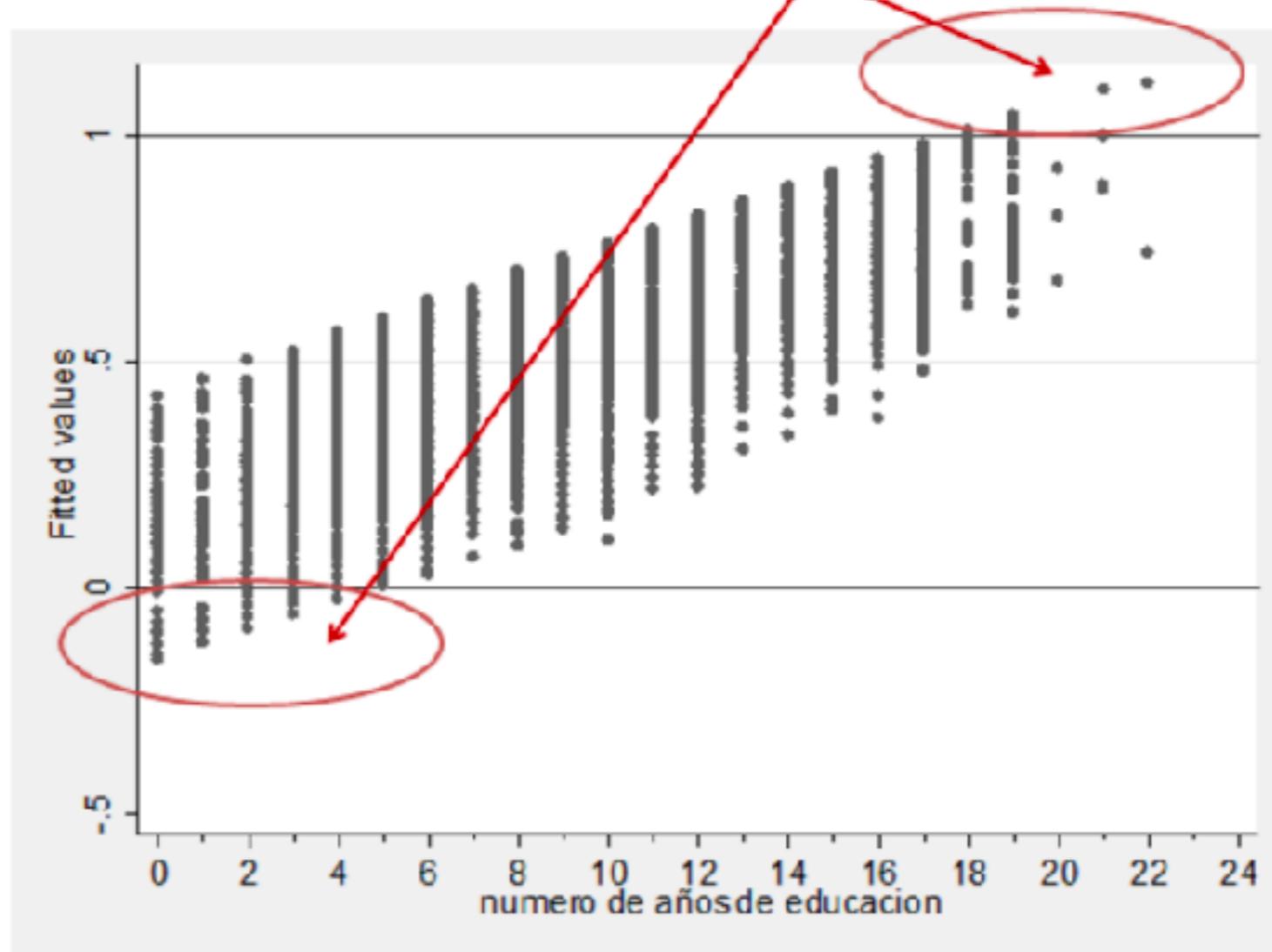
*¿Es importante el efecto? Estadísticamente sí. ¿Y económico? La probabilidad promedio de participación laboral es 0.542 (tabla anterior).*

*Una reducción en la probabilidad de participar de 0.163 equivale a 30% menor probabilidad de participar (30% = 0.163/0.542 \* 100), lo cual es muy importante.*

Fuente: Encuesta de Protección Social 2009.

# MODELO DE PROBABILIDAD LINEAL

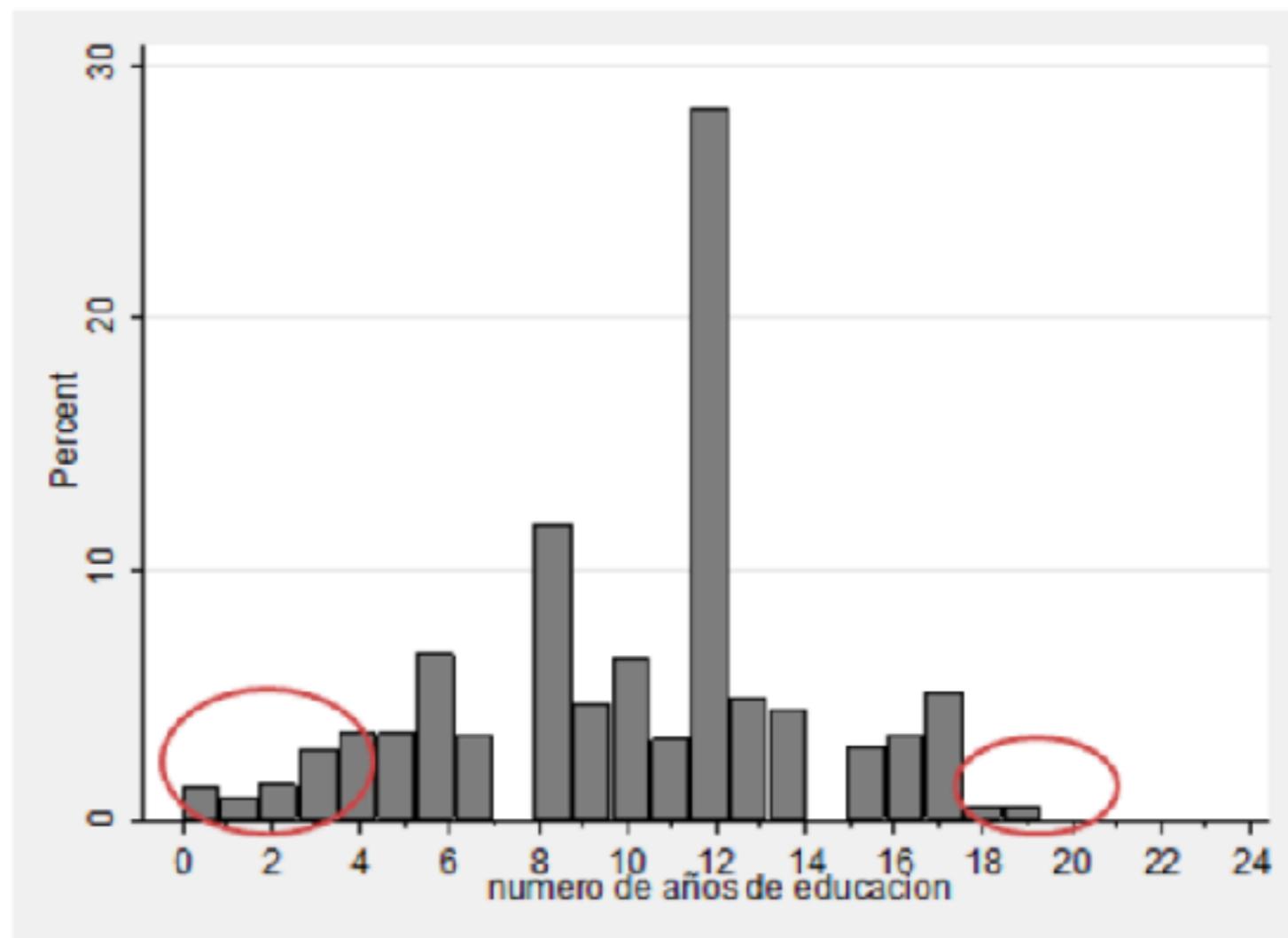
*El MPL tiene muchas ventajas: fácil de estimar, interpretación intuitiva  
Desventaja: el MPL puede predecir valores fuera del rango [0-1]*



*En resumen, el MPL se comporta adecuadamente para valores de las variables explicativas cercanas a la media; no se comporta bien para valores extremos....*

# MODELO DE PROBABILIDAD LINEAL

*....por suerte nuestra base de datos no tiene muchos valores extremos.*



## OTROS MODELOS: PROBIT, LOGIT

- ▶ El problema con el MPL es que  $p$  puede ser que  $p < 0$  ó  $p > 1$ . Podríamos restringir con kinks a que  $p \in [0, 1]$
- ▶ El modelo de probabilidad lineal con kinks es, por definición, la función de distribución acumulada de una distribución uniforme
- ▶ Sin embargo, hay otras funciones de distribución acumulada más continuas con forma de  $S$ :
  1. cuando  $F(\cdot)$  es la función de distribución acumulada de una distribución normal, hablamos de modelo probit
  2. cuando  $F(\cdot)$  es la función de distribución acumulada de una distribución logística, hablamos de modelo logit

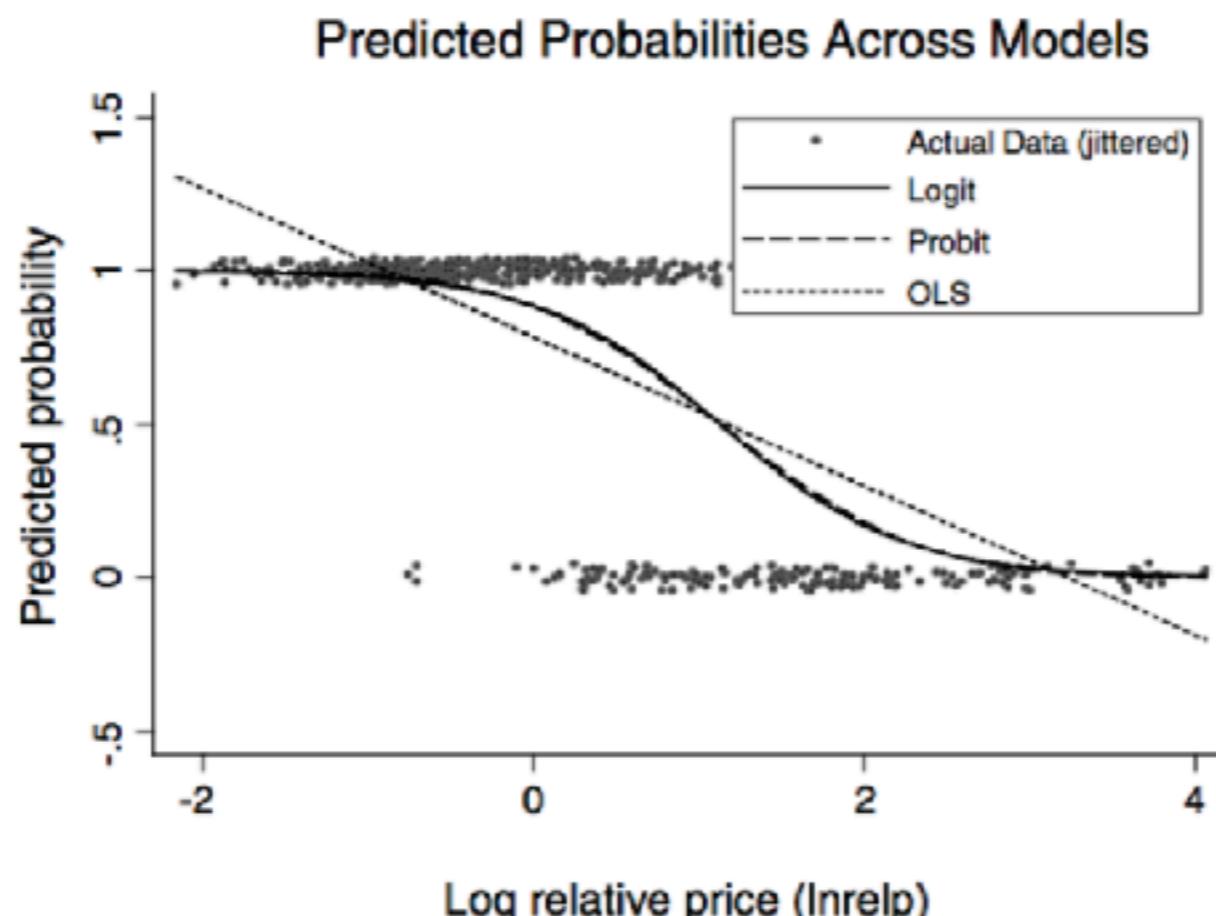
## OTROS MODELOS: PROBIT, LOGIT

**Table 14.2.** *Fishing Mode Choice: Logit and Probit Estimates<sup>a</sup>*

Regressor	Logit	Model Probit	OLS
Constant	2.053 (12.15)	1.194 (13.34)	0.784 (65.58)
ln relp	-1.823 (-12.61)	-1.056 (-13.87)	-0.243 (-28.15)
- ln L	-206.83	-204.41	-
Pseudo $R^2$	0.449	0.455	0.463

<sup>a</sup> Dependent variable  $y = 1$  if charter boat fishing and  $y = 0$  if pier fishing. Regressor  $x = \ln relp$ , the natural logarithm of the price of charter boat fishing relative to pier fishing. Intercept and slope parameter estimates with  $t$ -statistics in parentheses are from ML estimation of logit and probit models and from OLS estimation.

# OTROS MODELOS: PROBIT, LOGIT



**Figure 14.1:** Charter boat fishing: predicted probability from logit and probit models and OLS prediction when the single regressor is the natural logarithm of relative price. Actual outcomes of 1 or 0 are also plotted after jittering for readability. Data for 620 individuals.

## OTROS MODELOS: PROBIT, LOGIT

TABLE 2—ESTIMATES OF LOAN DENIALS AND INTEREST RATE CHARGED ON APPROVED LOANS, 1998 AND 2003

Variables	Loan denial, probit regression			Interest rate, OLS regression	
	New loans		Loan renewals	New loans	
	1998 (1)	2003 (2)	2003 (3)	1998 (4)	2003 (5)
Blacks	0.220*** (0.000)	0.369*** (0.000)	0.182*** (0.000)	0.002 (0.999)	1.055 (0.465)
Hispanics	0.278*** (0.000)	-0.015 (0.489)	0.170*** (0.000)	0.228 (0.570)	2.447** (0.028)
ANP	0.053*** (0.000)	0.081*** (0.003)	-0.070*** (0.000)	0.881 (0.140)	-0.279 (0.654)
White females	0.008 (0.398)	0.001 (0.943)	-0.006 (0.516)	-0.707** (0.023)	0.552 (0.181)
Number of firms	879	995	1,246	764	873

*Notes:* The control variables include only firm's credit history, firm characteristics, owner characteristics, region fixed effects, and the number of institutions that the firm used for all financial services. The coefficient estimates reported are the average partial effects. Robust *p*-values are in parentheses.

\*\*\* Significant at the 1 percent level.

\*\* Significant at the 5 percent level.

\* Significant at the 10 percent level.

## OTROS MODELOS: ELECCIÓN CON ALTERNATIVAS

1. **Conditional Logit**: utilizado cuando  $Y$  no tiene orden, regresores varían entre alternativas (ej. precio modo de viaje) y hay más de dos alternativas
2. **Multinomial Logit**: utilizado cuando  $Y$  no tiene orden, regresores no varían entre alternativas (ej. ingreso) y hay más de dos alternativas en el conjunto de opciones
3. **Ordered probit**: utilizado cuando  $Y$  es discreta en intervalo finito y con orden lógico. Ejemplo: ¿cuánto le gusta el curso? 1: nada, 2: poco, 3: algo, 4: bastante, 5: mucho

# REGRESIÓN DISCONTINUA

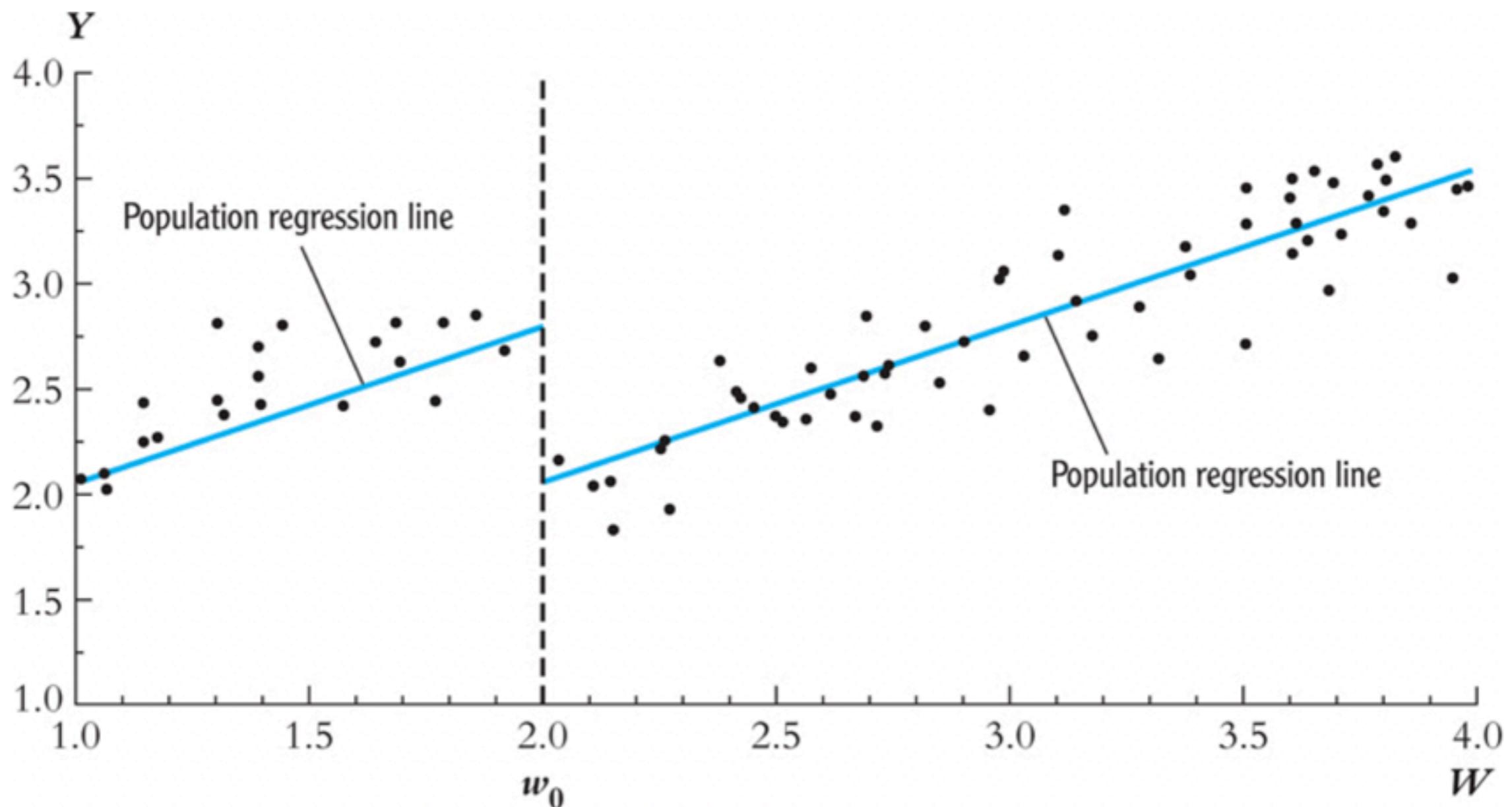
---

OTROS  
MÉTODOS

# REGRESIÓN DISCONTINUA

- ▶ Si el tratamiento ocurre cuando una variable continua  $W$  cruza un nivel  $W_o$ , entonces es posible estimar el efecto de tratamiento comparando a los que están justo antes de ese nivel y a los que están justo después.
- ▶ Si el efecto directo en  $Y$  de  $W$  es continuo, entonces el efecto de tratamiento se observa como un salto en el valor de  $Y$ .
- ▶ La magnitud del salto muestra el efecto de tratamiento.
- ▶ Hay de dos tipos:
  - ▶ Sharp
  - ▶ Fuzzy o difusa

## REGRESIÓN DISCONTINUA



## REGRESIÓN DISCONTINUA SHARP

Todos con  $W < w_0$  son tratados, entonces

$$D_i = 1 \text{ if } W_i < w_0 \text{ and } D_i = 0 \text{ otherwise.}$$

El efecto de tratamiento  $\beta_1$ , se estima por MCO:

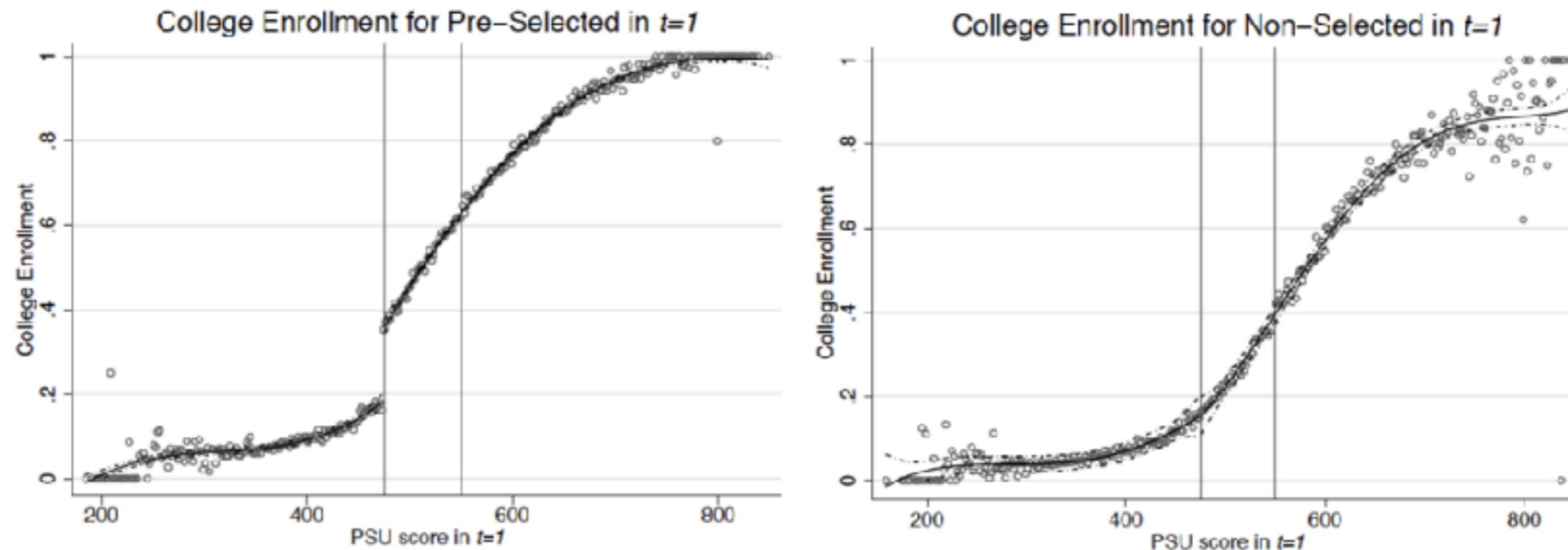
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Si cruzar el umbral afecta a  $Y_i$  solo mediante el tratamiento, entonces  $E(u_i|X_i, W_i) = E(u_i|W_i)$  y el estimador  $\hat{\beta}_1$  es insesgado.

# REGRESIÓN DISCONTINUA - EJEMPLO

Figura: Acceso al CAE y matrícula en alguna carrera universitaria

Figure 1: RD for Immediate College Enrollment.



- ▶ ¿Por qué graficamos primero? (1) forma funcional visual, (2) tamaño discontinuidad, (3) check saltos inesperados

## REGRESIÓN DISCONTINUA DIFUSA

Sea:

$D_i$  = tratamiento binario.

$Z_i = 1$  if  $W < w_0$  and  $Z_i = 0$  en otro caso.

If crossing the threshold has no direct effect on  $Y_i$ , so only affects  $Y_i$  by influencing the probability of treatment, then  $E(u_i | Z_i, W_i) = 0$ . Thus  $Z_i$  is an exogenous instrument for  $X_i$ .

Example:

Matsudaira, Jordan D. (2008). "Mandatory Summer School and Student Achievement." *Journal of Econometrics* 142: 829-850. This paper studies the effect of mandatory summer school by comparing subsequent perormance of students who fell just below, and just above, the grade cutoff at which summer school was required.

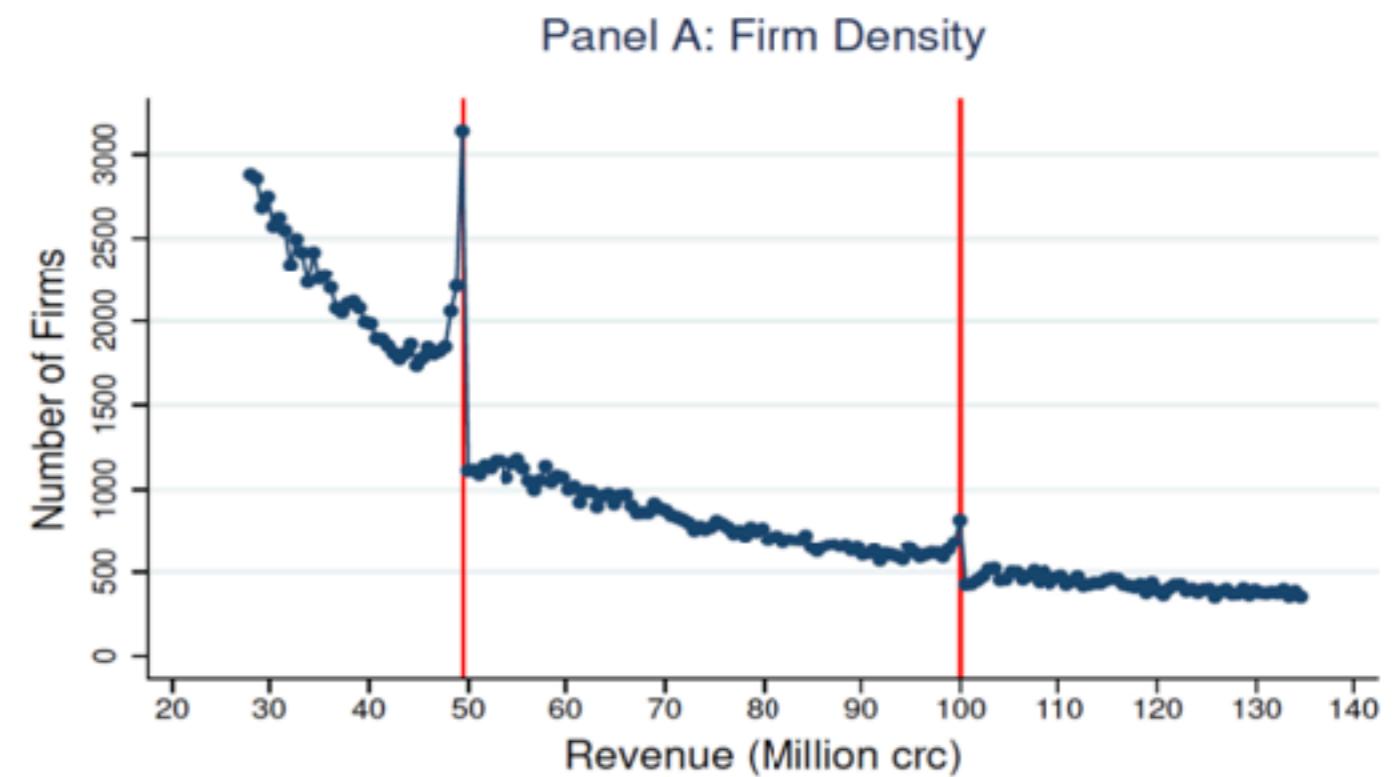
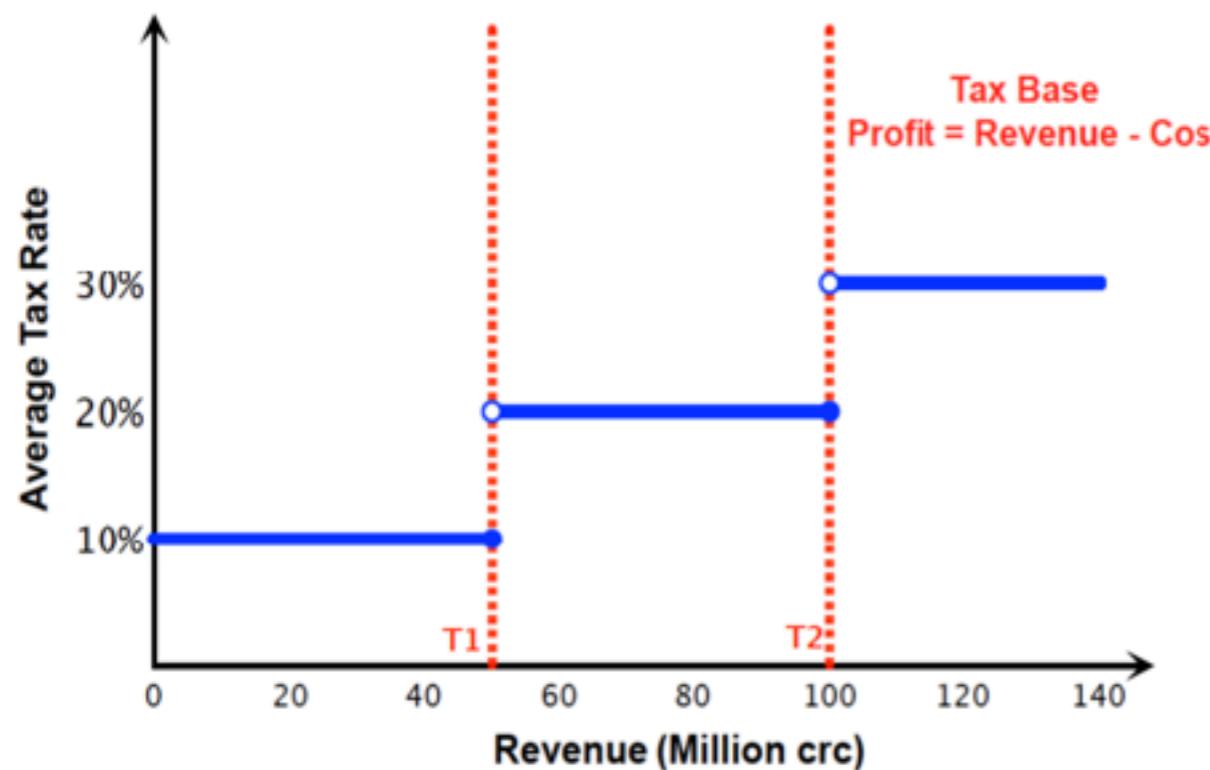
## CONSIDERACIONES EN RD

1. Diseños RD pueden ser inválidos si individuos pueden manipular la variable de asignación
2. Si individuos no pueden manipular de manera precisa, entonces esto implica que la variación alrededor del umbral es como un experimento aleatorio
3. Los diseños RD pueden ser analizados y evaluados como un experimento aleatorio (e.g. balance en observables)

## CONSIDERACIONES EN RD

Figura: Impuestos y manipulación de ganancias (bunching)

Figure 1: Costa Rica's Corporate Tax Schedule



## CONSIDERACIONES EN RD

Figura: Número de alumnos por sala de clase en Chile

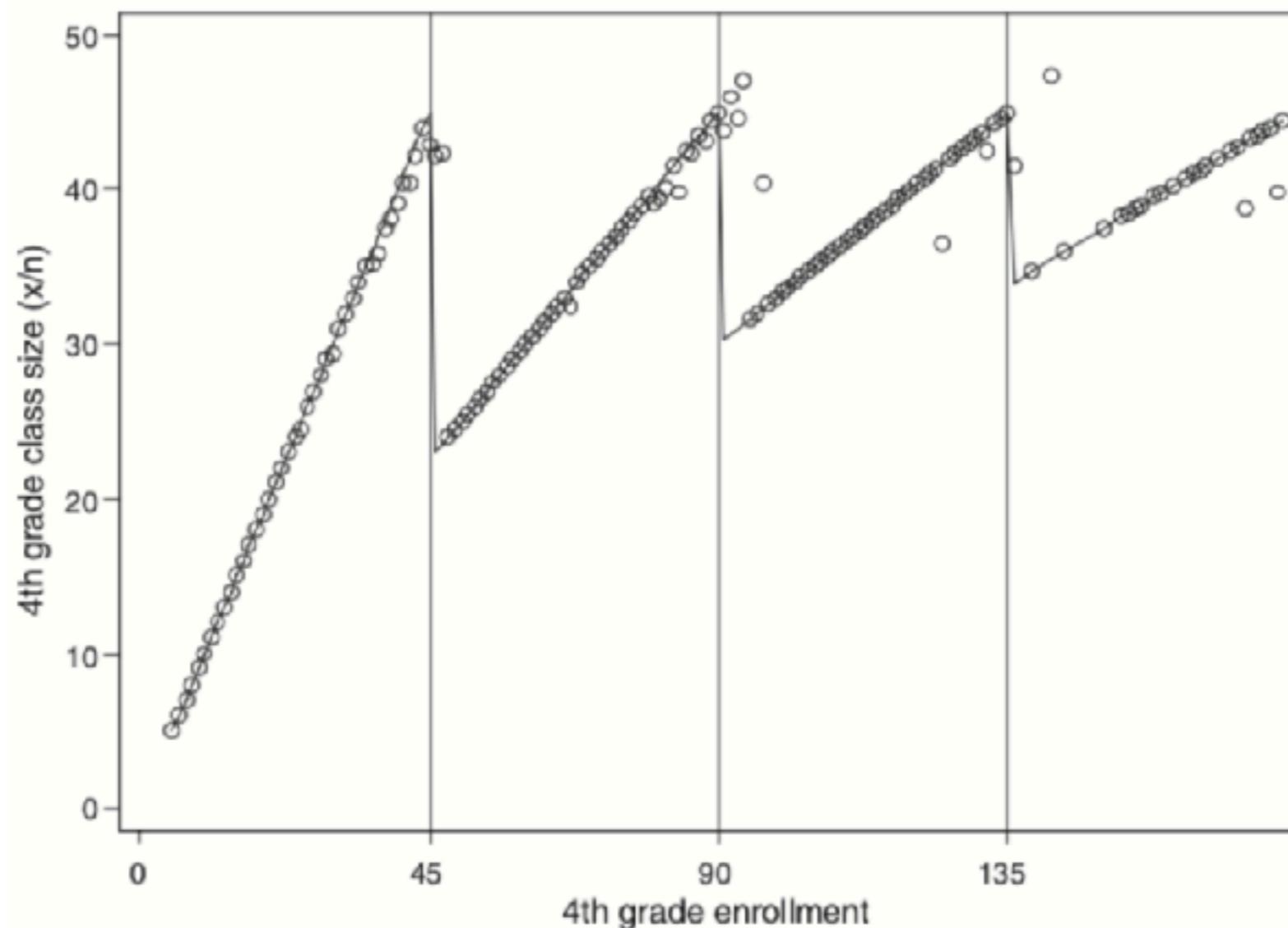
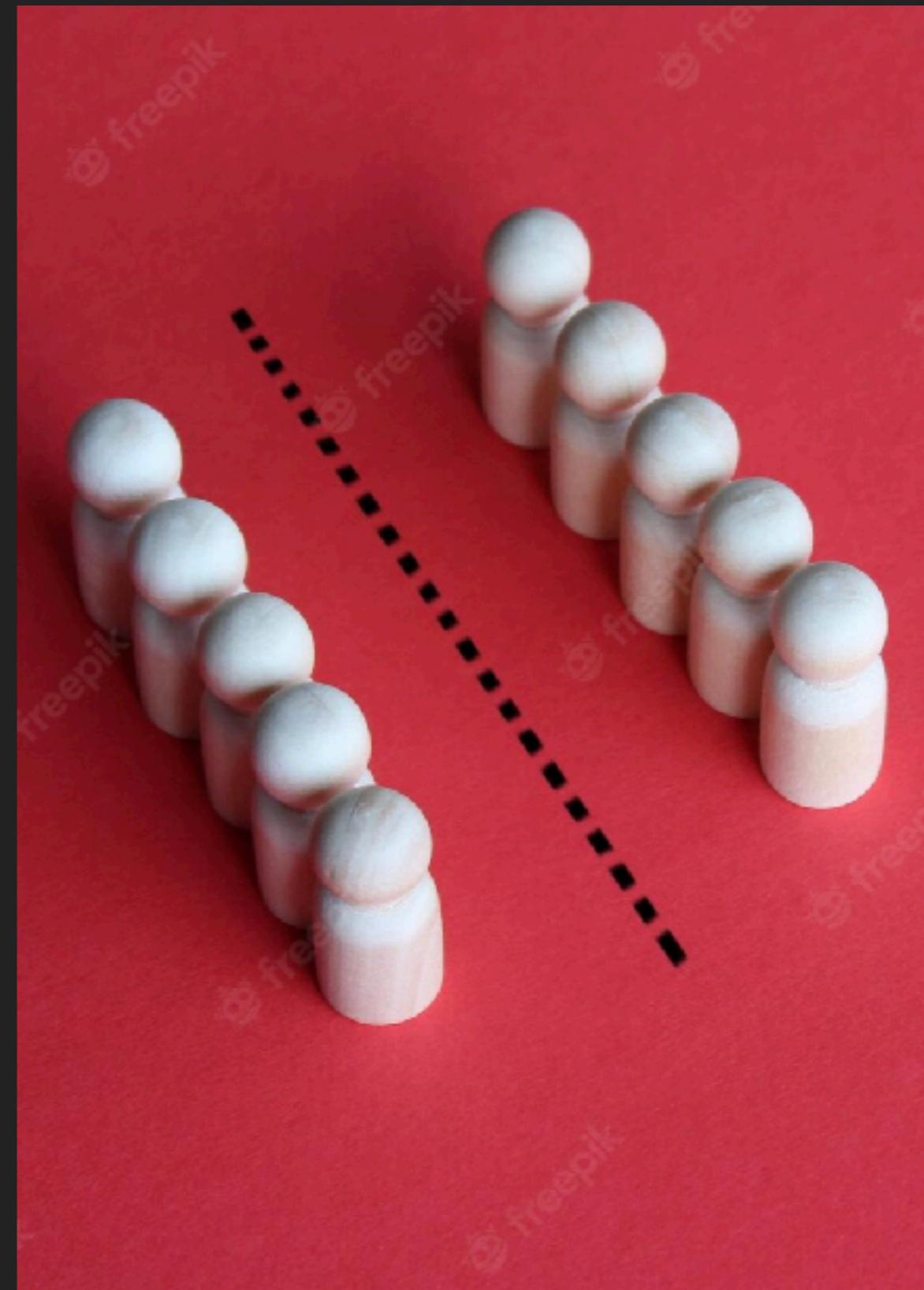


FIGURE 5. FOURTH GRADE ENROLLMENT AND CLASS SIZE IN URBAN PRIVATE VOUCHER SCHOOLS, 2002

# SIGAMOS EL EJEMPLO

- ▶ Piense ahora que, al inicio del año los estudiantes se les administró una prueba de selección para entrar al curso de verano.
- ▶ La prueba coincidía en armar una figura con legos y el tiempo que se demoraban era el resultado. Las personas que se demorarán menos de un minuto fueron seleccionados y los que se demoraron más o igual a 1 minuto, quedaron fuera.
- ▶ ¿Es posible estimar un modelo RD con esta información?
- ▶ Realicemos la estimación.



# TAREA DE AYUDANTÍA

### Objetivo

- El objetivo de este taller es llevar a una aplicación práctica los contenidos de la sesión.
- Trabaje de manera individual o en grupos de hasta 3 personas.
- Entregue en un notebook, indicando claramente como título Taller 2 - Ayudantía - Métodos en Ciencias Sociales Computacionales y los integrantes del grupo.

### Preguntas

1. Seleccione una de las bases de datos disponibles en la carpeta de ayudantía.
2. Elabore una pregunta de investigación o hipótesis que sea posible responder con la base de datos seleccionada.
3. Explicite la relación causal que desea investigar y su operalización en variables observadas.
4. Comente: ¿Cuál sería un experimento ideal para identificar la relación causal de interés?
5. Comente: ¿Qué sesgos puede estar evidencian si estima directamente un modelo 'ingenuo' con los datos observacionales disponibles?
6. Proponga una estrategia de identificación factible. Comente sus ventajas y debilidades.
7. Realice la estimación e interprete sus resultados.

## TAREA DE AYUDANTÍA 2

### Objetivo

- El objetivo de este taller es llevar a un caso real algunos de los temas tratados en la sesión.
- Trabaje de manera individual o en grupos de hasta 3 personas.
- Entregue en un notebook, indicando claramente como título Taller 2 - Ayudantía - Métodos en Ciencias Sociales Computacionales y los integrantes del grupo.

### Instrucciones

Ingrese al GitHub del proyecto

### Preguntas

1.

## TAREA DE AYUDANTÍA 2

Ingrese al GitHub del proyecto

En esta ayudantía vamos a replicar el estudio [Dynamics of cross-platform attention to retracted papers] (<https://www.pnas.org/doi/10.1073/pnas.2119086119>).

Los autores pusieron a disposición de la comunidad un [repositorio en github]([https://github.com/haoopeng/retraction\\_attention](https://github.com/haoopeng/retraction_attention)) con los datos procesados y los algoritmos usados para obtener los resultados de la investigación.

## TAREA DE AYUDANTÍA 2

Actividades:

- 1. Replicar el match por nuestra cuenta**
- 2. Estimar los resultados con nuestro match**
- 3. Replicar los resultados con el match de los autores**
- 4. Comparar los resultados de ambas estimaciones**