

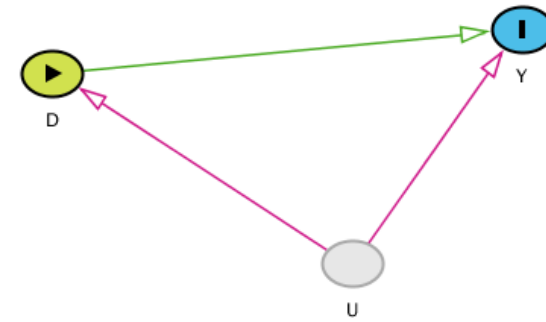
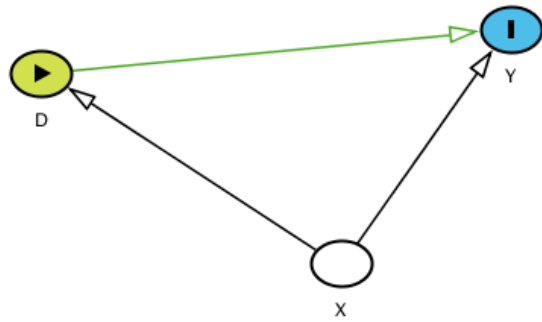
# Evaluación de Impacto: DAG y Matching

Francesco Bogliacino

# Direct Acyclical Graphs

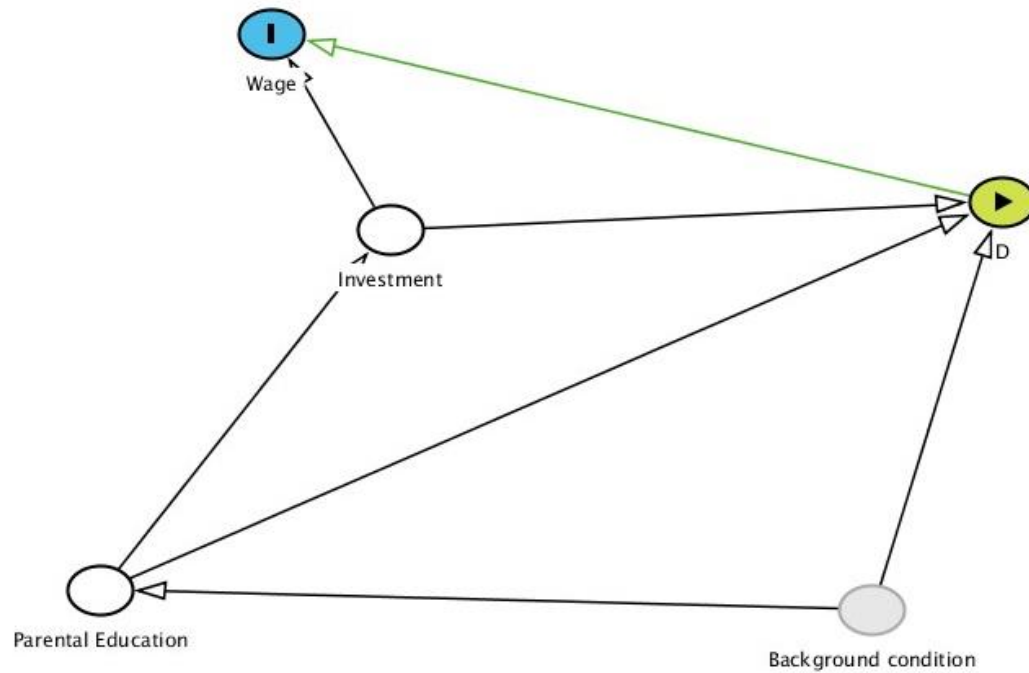
- Mientras que en economía usamos ecuaciones para describir el “modelo estructural” hay otro enfoque usado en otras disciplinas
- DAG postula relaciones causales a través de grafos con nodos y flechas
  - No sirven para relaciones simultaneas [flechas directas van solo por un lado]
- En general ignorados, son útiles para introducir los conceptos de
  - Confound
  - Collider
  - Backdoor path

# Using DAGITTY



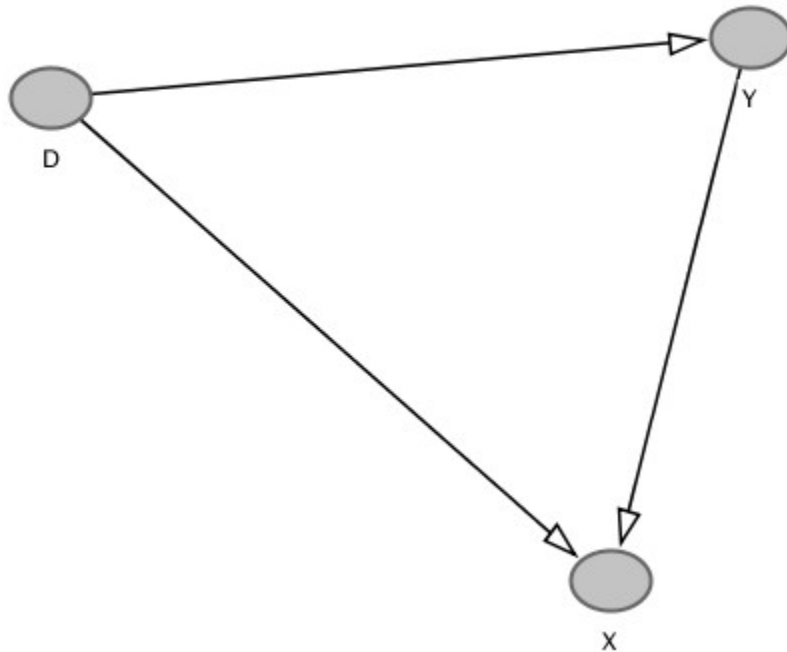
CONFOUNDER

# Backdoor paths



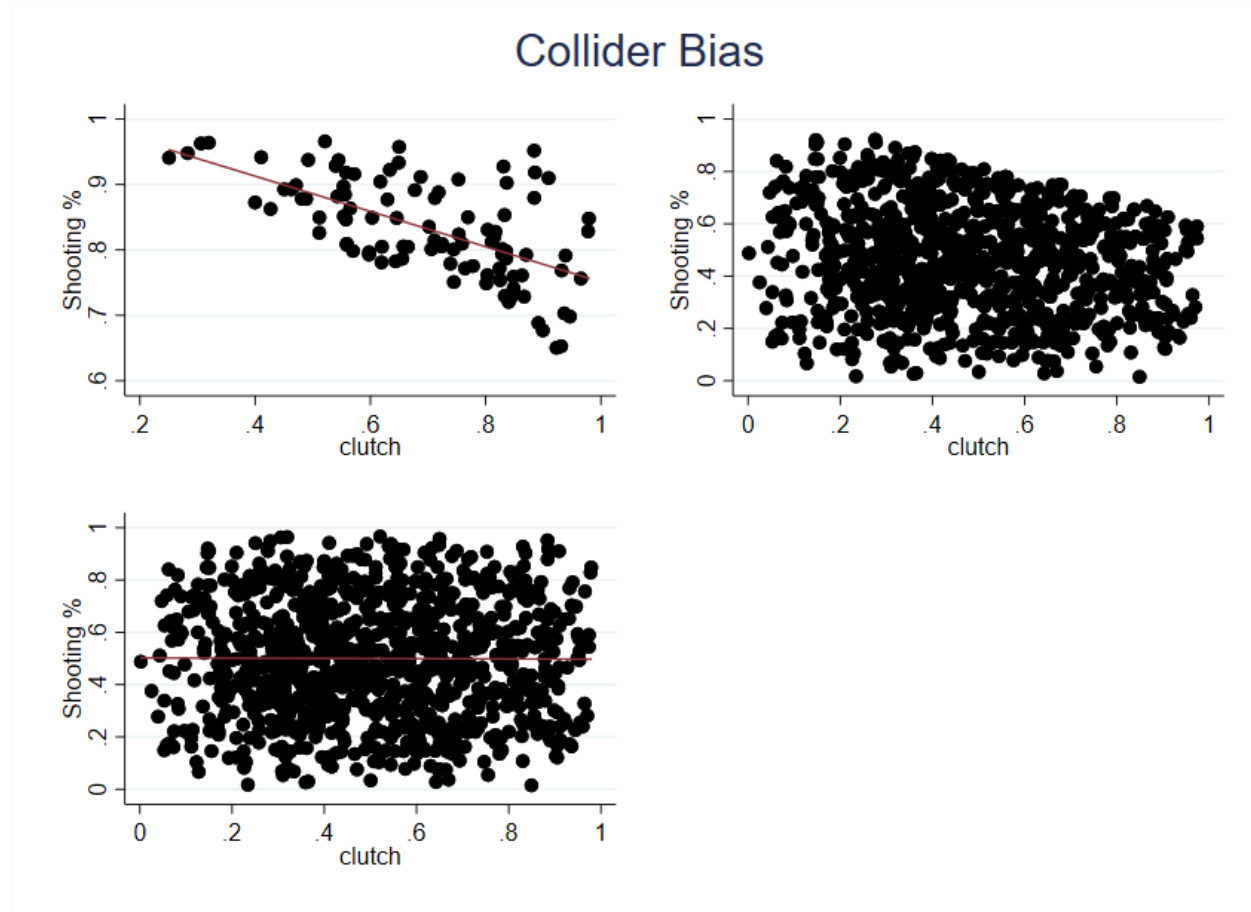
- $D \rightarrow Y$
- $D \leftarrow I \rightarrow Y$
- $D \leftarrow PE \rightarrow I \rightarrow Y$
- $D \leftarrow BC \rightarrow PE \rightarrow I \rightarrow Y$

# Collider

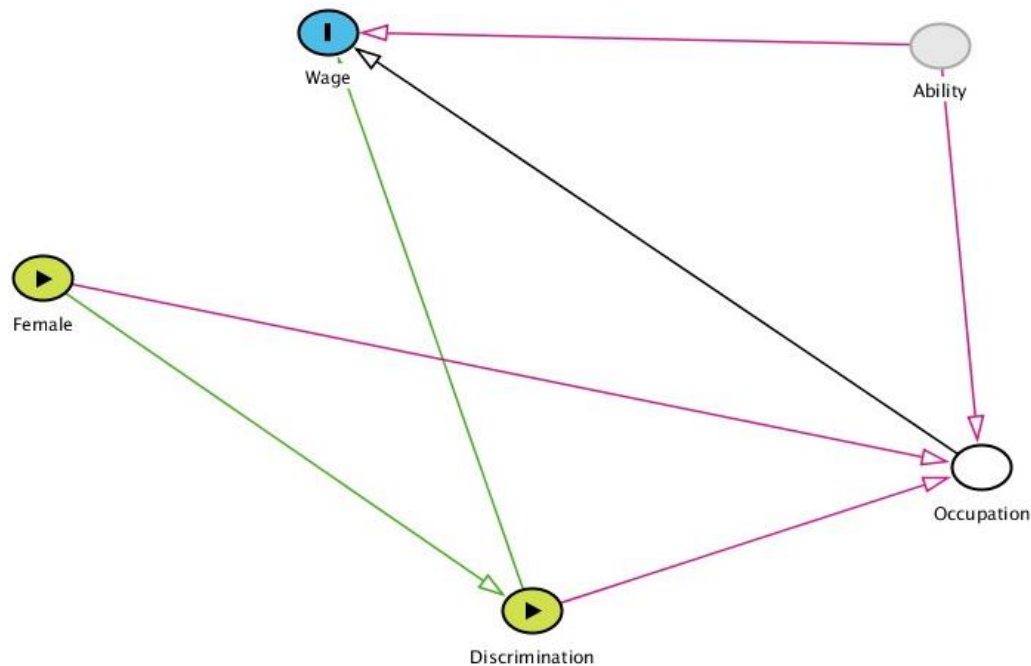


- En X chocan los efectos causales de D y de Y
- Es como si X absorbiera esos efectos **cerrando** ese camino
- Si bloqueo X, indirectamente creo correlación entre D y Y
- Es decir el collider lo tengo que dejar solo o abro un camino que estaba cerrado

# Collider via Sample Selection



# Discriminación en el mercado del trabajo



- $D \rightarrow Wage$
- $D \rightarrow o \rightarrow Wage$
- $D \leftarrow F \rightarrow o \rightarrow Wage$
- $D \rightarrow o \leftarrow A \rightarrow Wage$
- $D \leftarrow F \rightarrow o \leftarrow A \rightarrow Wage$

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

VARIABLES	(1) Unconditional	(2) Collider	(3) Identified
female	-2.99*** (0.09)	0.59*** (0.03)	-0.99*** (0.03)
occupation		1.80*** (0.01)	1.01*** (0.01)
ability			1.97*** (0.02)
Constant	1.97*** (0.06)	0.21*** (0.02)	0.99*** (0.02)
Observations	10,000	10,000	10,000
R-squared	0.11	0.92	0.95



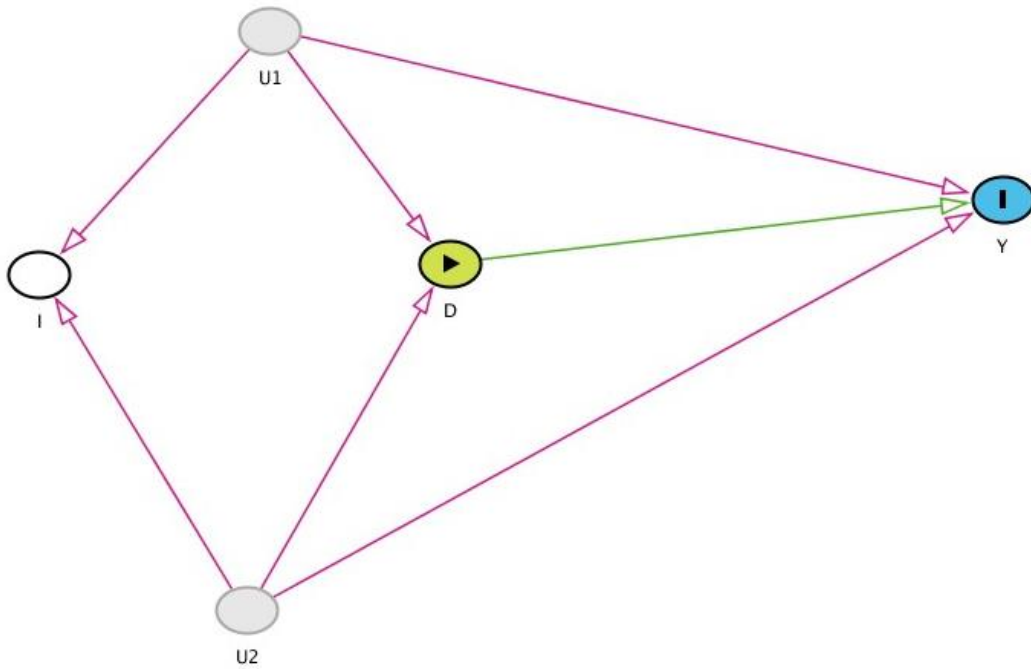
	(1)	(2)	(3)
VARIABLES	Identified	Collider	Biased
d	50.05***		0.33
	(0.04)		(0.97)
x		0.50***	0.50***
		(0.00)	(0.01)
Constant	99.99***	24.94***	25.43***
	(0.02)	(0.06)	(1.46)
Observations	2,500	2,500	2,500
R-squared	1.00	1.00	1.00

- clear all
- set seed 541
- \* Creating collider bias
- \* Z -> D -> Y
- \* D -> X <- Y
- \* 2500 independent draws from standard normal distribution
- clear
- set obs 2500
- gen z = rnormal()
- gen k = rnormal(10,4)
- gen d = 0
- replace d =1 if k>=12
- \* Treatment effect = 50
- gen y = d\*50 + 100 + rnormal()
- gen x = d\*50 + y + rnormal(50,1)

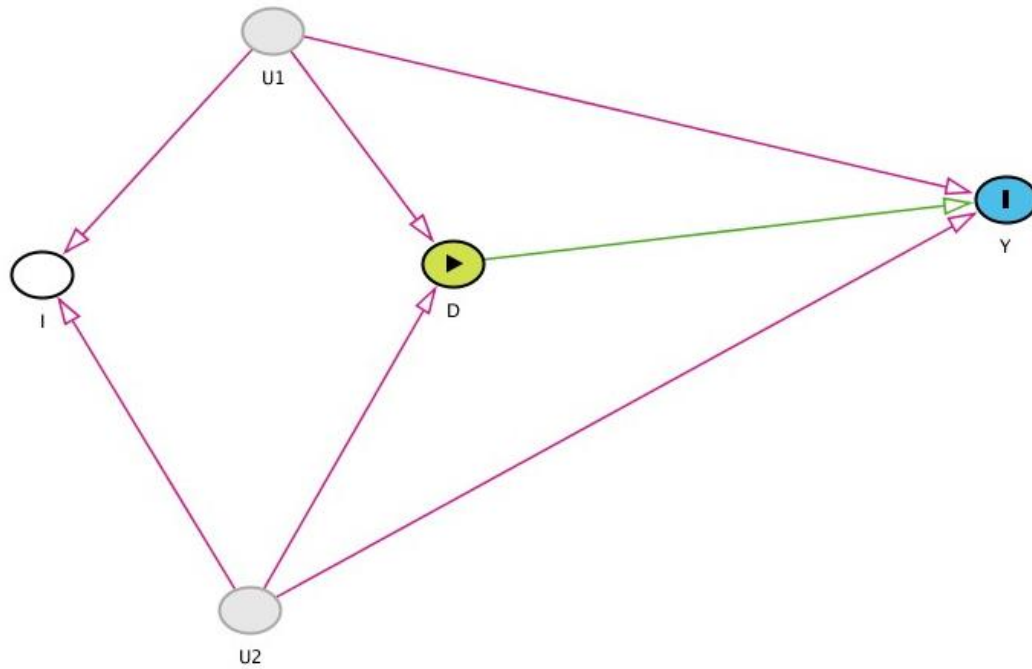
Source: Mixtape

# Un ejercicio todos juntos

- Me ayudan a escribir todos los backdoor paths?



# Un ejercicio todos juntos



- $D \leftarrow U1 \rightarrow Y$
- $D \leftarrow U2 \rightarrow Y$
- $D \leftarrow U2 \rightarrow I \leftarrow U1 \rightarrow Y$
- $D \leftarrow U1 \rightarrow I \leftarrow U2 \rightarrow Y$

# Backdoor criterion

“a set of variables  $X$  satisfies the backdoor criterion in a DAG if and only if  $X$  blocks every path between confounders that contain an arrow from  $D$  to  $Y$ ”

En la práctica cierro un camino trasero si:

- Condiciono sobre un no-collider
- Hay un collider y no estoy controlando por el



Matching

# Selection on observable

- Un **diseño de investigación** (que no es muy popular, pero existe) se basa en la idea de usar el “control” como la estrategia para estimar los contrafactuales. El control lo hago sobre las variables **observables**
- **Un estimador** muy popular basado en este supuesto de identificación es el emparejamiento o matching
- En práctica, el contrafactual lo voy a imputar usando las covariadas observables

# Covariadas

## Definición

Una variable  $X$  es predeterminada respecto a un tratamiento  $D$  si por cada individuo  $i$ ,  $X_i^0 = X_i^1$

Note:

- Esto no implica independencia
- Puede que las covariadas predeterminadas no varíen en el tiempo, pero no es una condición necesaria

# Un ejemplo creíble de identificación sobre observables

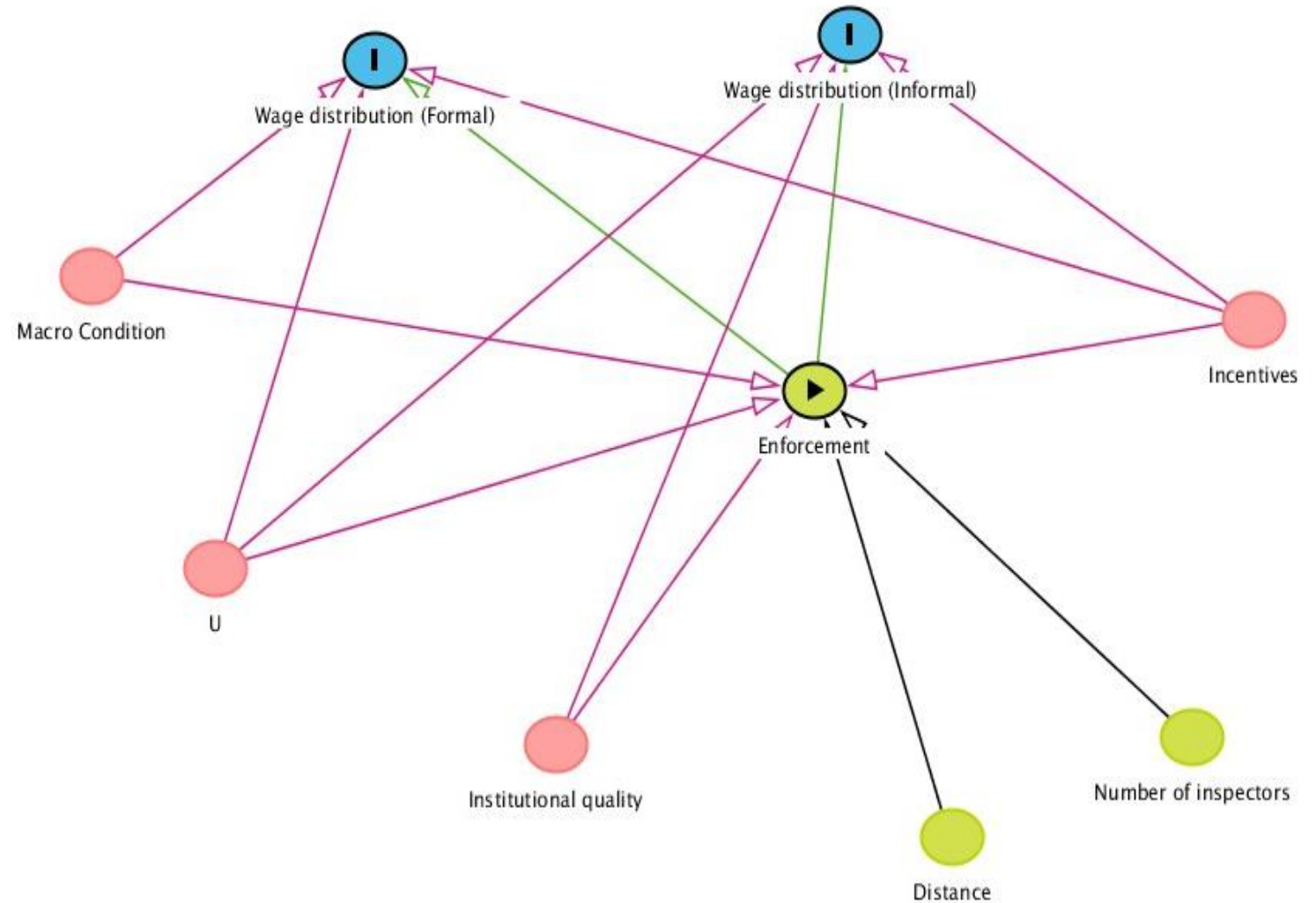
Almeida y Carneiro, AEJ AE 2012

- Trabajo formal e informal (Brasil)
  - Informal costoso para el trabajador a largo plazo
  - Formal costoso para la empresa
- Inspecciones para el cumplimiento:
  - Deberían golpear empresas informales
  - Golpean empresas formales (más fáciles a detectar, por incentivos)
- Consecuencias:
  - Ajustan distribución en el sector informal (aumenta salario)
  - Ajustan distribución en el sector formal (aumenta ocupación, reduce salario y desigualdad)
  - Aumenta desempleo



# Un ejemplo creíble de identificación sobre observables

- Controles no son random, tenemos problemas de identificación;
- Pregunta: ¿podemos usar datos a nivel de empresa?



## La ecuación

$$\begin{aligned}(1) \quad Y_{cs} = & \alpha + \beta(D_{cs} \times I_s) + \psi_0 D_{cs} + \psi_1 (D_{cs})^2 + \phi(D_{cs} \times \mathbf{XState}_s) \\ & + \psi_2 DCapital_{cs} + \psi_3 (DCapital_{cs})^2 + \psi_4 (DCapital_{cs} \times I_s) \\ & + \tau(DCapital_{cs} \times \mathbf{XState}_s) + \psi_5 TCosts_{cs} + \psi_6 (TCosts_{cs})^2 \\ & + \psi_7 (TCosts_{cs} \times I_s) + \rho(TCosts_{cs} \times \mathbf{XState}_s) \\ & + \delta \mathbf{XCity}_{cs} + \sigma Y_{cs}^{1980} + \eta_s + u_{cs},\end{aligned}$$

# Los resultados

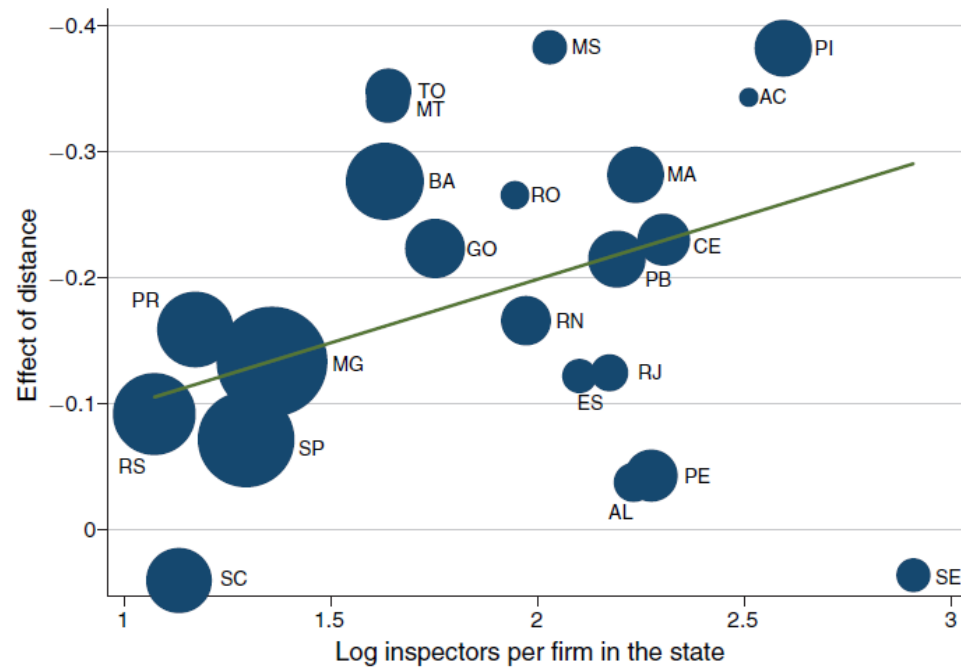


FIGURE 1. EFFECT OF DISTANCE ON INSPECTIONS PER FIRM IN THE CITY ACROSS BRAZILIAN STATES

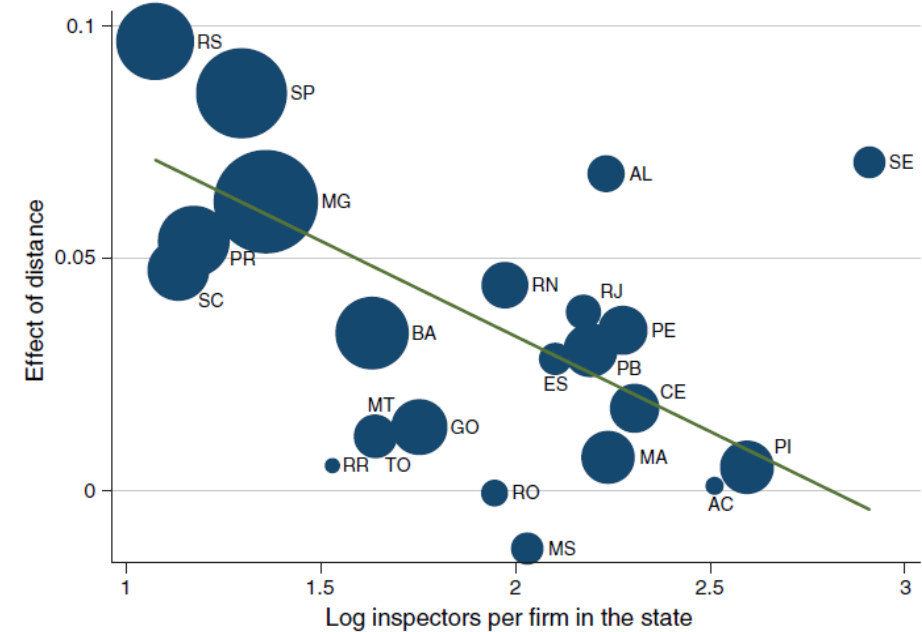


FIGURE 2. EFFECT OF DISTANCE ON THE SHARE OF INFORMAL WORKERS IN THE CITY ACROSS BRAZILIAN STATES

# Un ejemplo creíble de identificación sobre observables

Yanagitzawa-Drott, QJE 2014

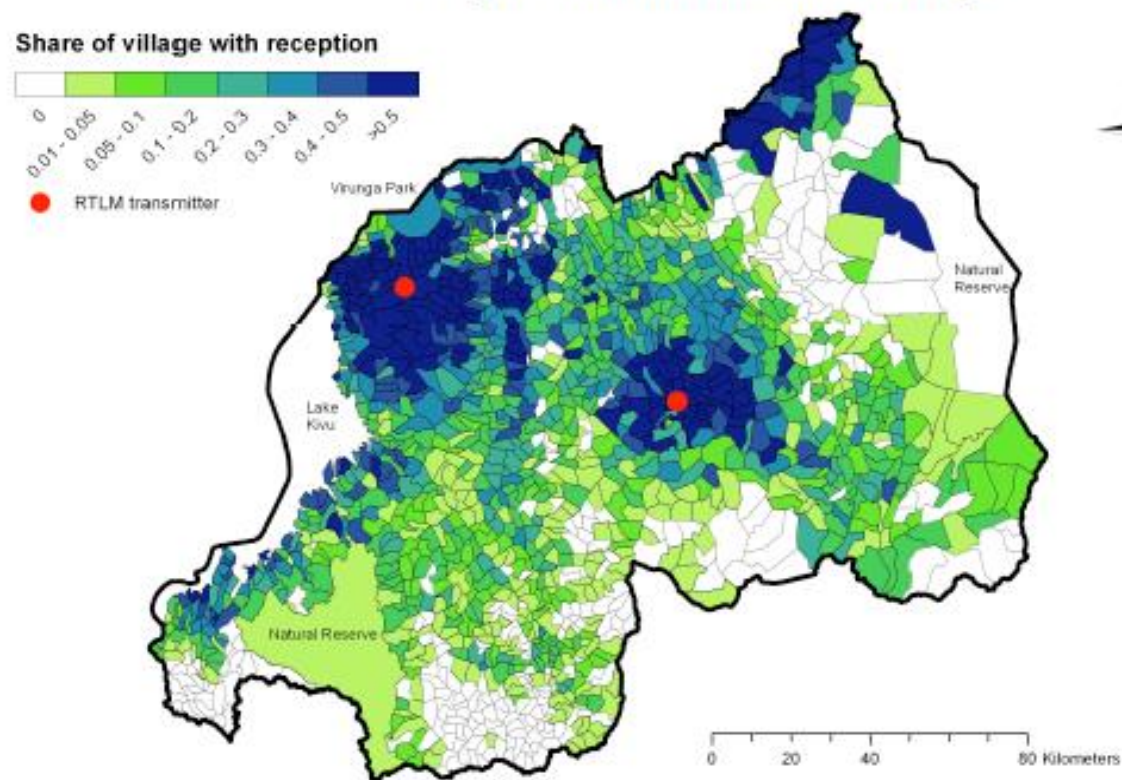
- Genocidio Rwandés
  - Estructura neoclánica destruida por la colonización
  - Dictadura a partido único sobre el modelo de Zaire
  - Formación de estructura paramilitares
- Arusha (1994) y el genocidio
  - El papel de RTLM
  - La justicia transicional (Gacaca) y la fuente de los datos
- El papel de los medios en explicar el comportamiento:
  - Teoría de la persuasión (política, publicidad, inversionistas)
  - Cambian creencias (rational or behavioral) y/o preferencias

# Un ejemplo creíble de identificación sobre observables

$$\begin{aligned} \log(h_{vci}) \\ = \beta_v r_{ci} + \mathbf{X}'_{ci} \pi + \gamma_i + \varepsilon_{ci} \end{aligned}$$

¿Por qué no efecto fijo de pueblo c?

Figure II. RTLM radio coverage



## Resultados

	Total Violence			Militia Violence			Individual Violence		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Radio Coverage in Village	0.507** (0.226)	0.526** (0.242)	0.484** (0.235)	0.582** (0.239)	0.559*** (0.216)	0.544*** (0.206)	0.450* (0.233)	0.465* (0.252)	0.418* (0.246)
Population in 1991, log			0.590*** (0.131)			0.589*** (0.171)			0.624*** (0.150)
Population Density in 1991, log			-0.014 (0.070)			0.004 (0.101)			-0.015 (0.069)
Distance to Major Town, log			0.068 (0.150)			-0.233 (0.149)			0.113 (0.152)
Distance to Major Road, log			-0.196** (0.076)			-0.245*** (0.090)			-0.193** (0.075)
Distance to the Border, log			0.171* (0.103)			0.030 (0.126)			0.186* (0.103)
East Sloping, dummy			0.017 (0.070)			0.098 (0.092)			0.014 (0.084)
North Sloping, dummy			0.065 (0.068)			0.041 (0.092)			0.079 (0.068)
South Sloping, dummy			-0.013 (0.074)			-0.028 (0.101)			-0.012 (0.077)
Observations	1065	1065	1065	1065	1065	1065	1065	1065	1065
R-squared	0.63	0.64	0.66	0.52	0.53	0.55	0.62	0.63	0.65
Commune FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Propagation Controls	N	Y	Y	N	Y	Y	N	Y	Y

Notes: Militia Violence is the number of prosecuted person under category 1 crimes, which are prosecutions against organizers, leaders, army and militia. Individual Violence is crime category 2 prosecutions for homicides, attempted homicides and serious violence. Total Violence is the sum of both categories. Radio Coverage in Village is the share of the village area that has RTLM reception. The radio propagation controls are: latitude, longitude, and second-order polynomials in village mean altitude, village altitude variance, distance to the nearest RTLM transmitter. Standard errors in parentheses, adjusted for spatial correlation (Conley, 1999). Significance levels at \*10%, \*\*5%, \*\*\*1%.

# Supuestos de identificación

- CIA:

- $Y_i^1, Y_i^0 \perp D_i | X$

$$\begin{aligned} \rightarrow E[Y_i^1 - Y_i^0 | X] &= E[Y_i^1 | D_i = 1, X] - E[Y_i^0 | D_i = 1, X] = \\ &= E[Y_i^1 | D_i = 1, X] - E[Y_i^0 | D_i = 0, X] = \\ &= E[Y_i | D_i = 1, X] - E[Y_i | D_i = 0, X] \end{aligned}$$

- Common Support

- $1 > P(D_i = 1 | X) > 0$

este se puede imponer en práctica

¿Como se lee?

Es tan bueno como aleatoriamente asignado, una vez controlado por X

Recuerden, el SDI no se puede testear, pero tiene que ser plausible

	1	2	3	4
4	Promedio tratados=867000 Promedio no tratados= xxxxxxxx	Promedio tratados=yyyyyy Promedio no tratados= xxxxxxxx	Promedio tratados=yyyyyyy Promedio no tratados= xxxxxxxx	
5				
6				
7				

	1	2	3	4
4	Prob asignacio=80%			
5	Prob asignación 75%			
6		Prob( )=50%		<del>Prob asignación 0</del>
7				<del>Prob asignación 1</del>



# ¿Cómo puedo estimar el impacto?

- Puedo estimarlo como
  - $\widehat{\delta_{ATE}} = \sum_i [E[Y_i|D_i = 1, X] - E[Y_i|D_i = 0, X]]w_i$
  - Necesito poder estimar esos pesos, por eso necesito observar datos. Los pesos dependen de las “celdas” en la cual estoy calculando
- Si estoy feliz con el ATT
  - $\widehat{\delta_{ATT}} = \sum_i [E[Y_i|D_i = 1, X] - E[Y_i|D_i = 0, X]](w_i|D = 1)$
  - Para este segundo puedo permitir que haya muchos controles sin parejo, pero necesito que haya un match para cada tratado

# A simple Excel example

OUTCOME U TRATADAS	
	756
	89
	543
	259
	985
	656
	406
	344
	178
	74
	244
	797

# Matching

- Típicamente la distribución de las no covariadas en los no tratados está desbalanceada
- La maldición de la dimensionalidad
- En muchos casos tengo variables continuas entonces no puede usar matching exacto, tengo que usar aproximado
- Este necesariamente requiere de algo de programación porque es muy non lineal

# Matching Exacto

- Un simple estimador para emparejamiento exacto es  $\sum_i^{N_t} Y_i - \left(\frac{1}{M} \sum_{j(i)}^M Y_{j(i)}\right)$  para ATT
- Para ATE, sería  $\sum_i^N (2D_i - 1) \left[Y_i - \left(\frac{1}{M} \sum_{j(i)}^M Y_{j(i)}\right)\right]$

# Approximate matching

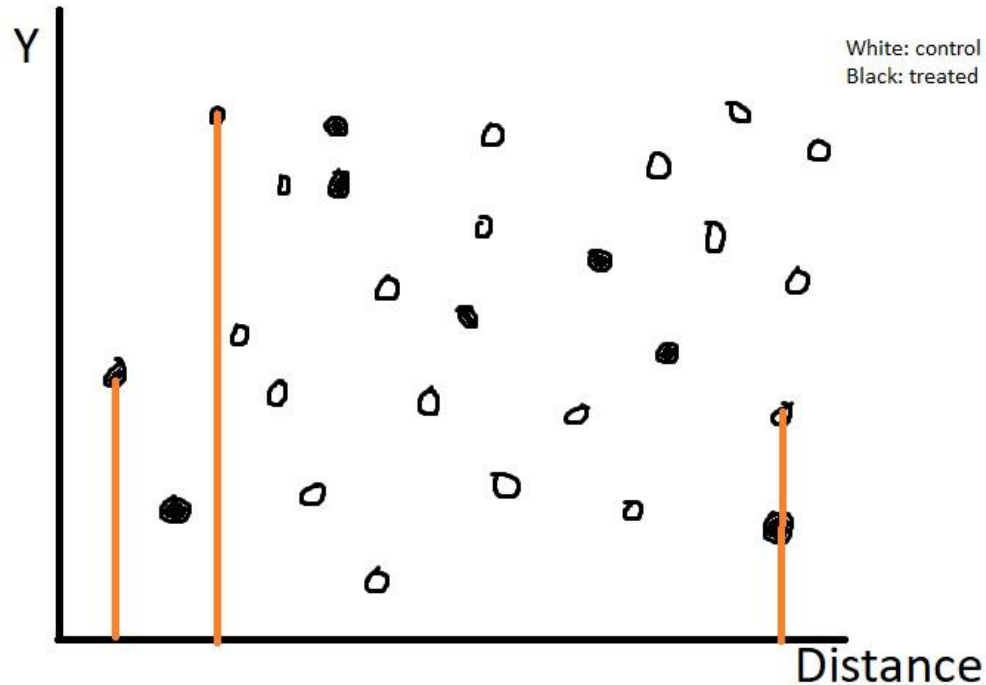
- Si tengo N variables que me sirven para cumplir con la CIA, voy a tener un problema en término de lograr observaciones para emparejar;
- Una solución es medir la *distancia* en el espacio de las covariadas:

- Euclídea:  $\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{\sum_{n=1}^k (x_{ni} - x_{nj})^2}$

- Euclídea normalizada:  $\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \widehat{V}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$

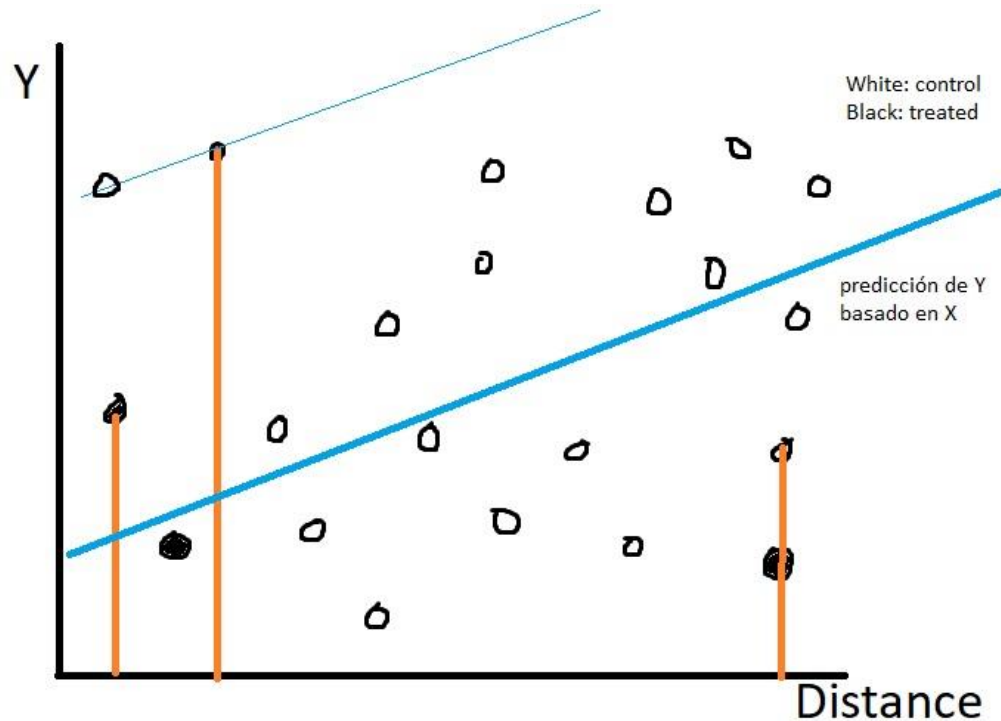
- Mahalanobis:  $\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \widehat{\Sigma}_X^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$

# El papel de la distancia



- Obviamente hay un sesgo introducido por el hecho que estoy usando emparejamiento aproximado
- Para corregir por este sesgo debería imputar un valor que tenga en cuenta las diferencias en las covariadas:
  - Cuál es la relación entre X y Y
  - Qué tanto debería ajustarse ese contrafactual teniendo en cuenta que estoy usando alguien “cercano”

# El papel de la distancia



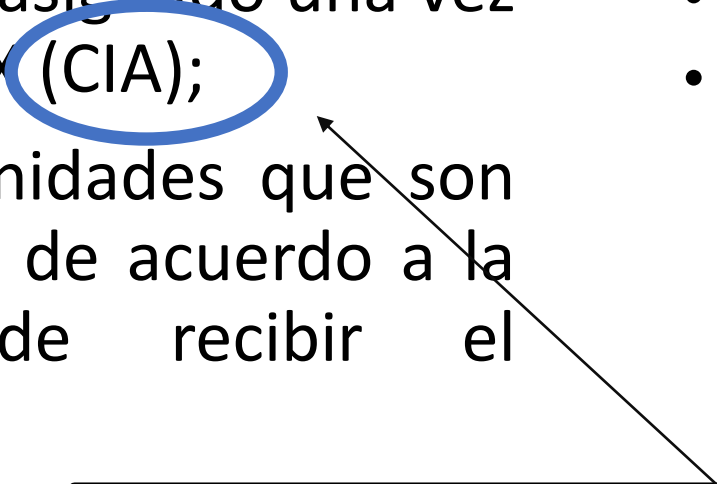
- Obviamente necesitamos un buen estimador de la CEF

$$\mu^0(X_i)$$

- Una opción puede ser OLS
- Puedo hacer cosas más sofisticadas tipo polinomios o otras estimaciones no paramétricas (pero hacerlas si vale la pena...)
- EXCEL

# Propensity Score Matching

- El tratamiento no es random, pero es tan bueno como aleatoriamente asignado una vez controlado por  $X$  (CIA);
- Comparamos unidades que son intercambiables de acuerdo a la probabilidad de recibir el treatment
- En práctica:
  - Estimo el PS
  - Hago el match (emparejo)
  - Estimo los errores estándar



Ojo, tiene que ser creíble

- Detalles institucionales
- Razones teóricas
- ...

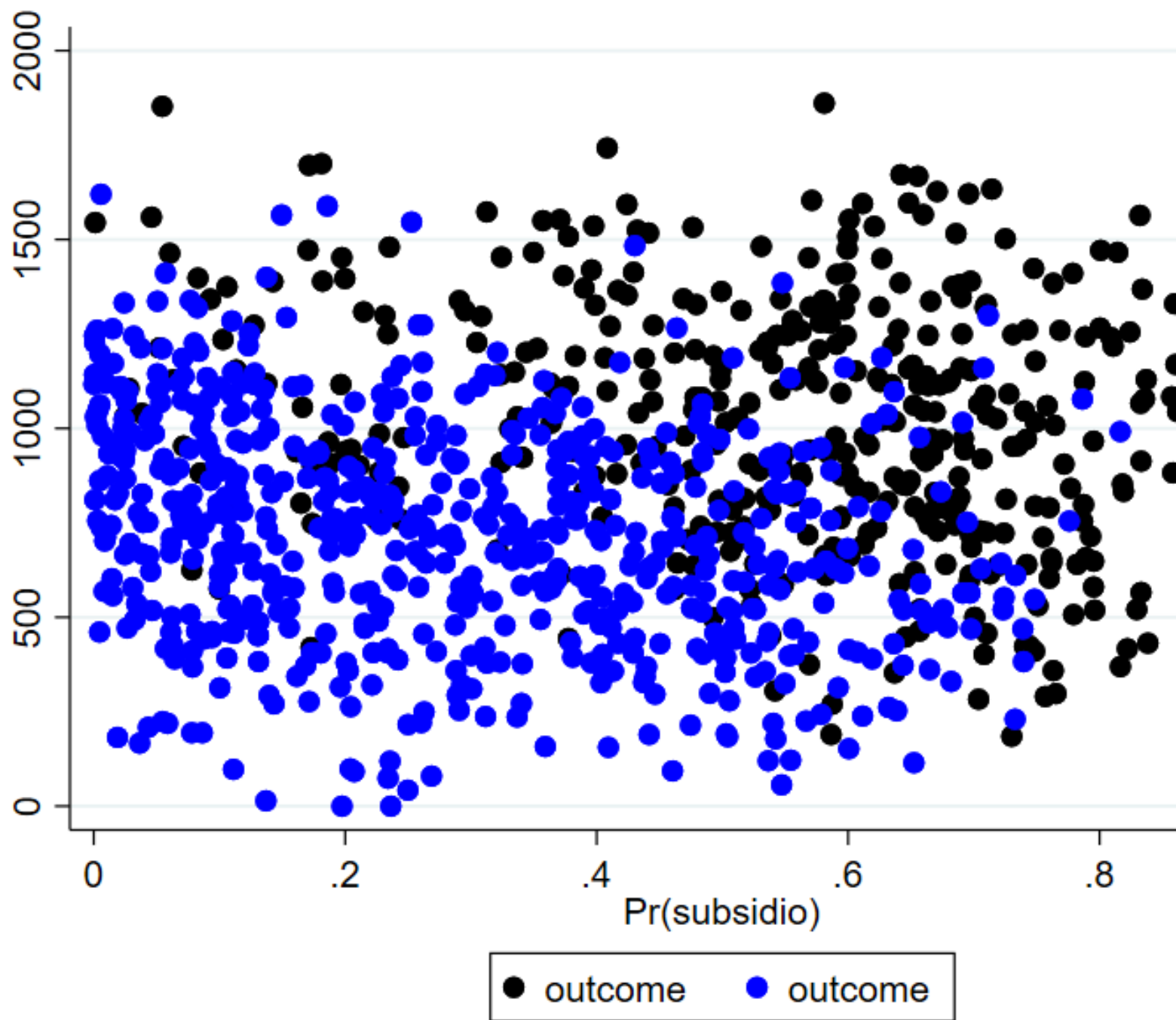


# Un ejemplo

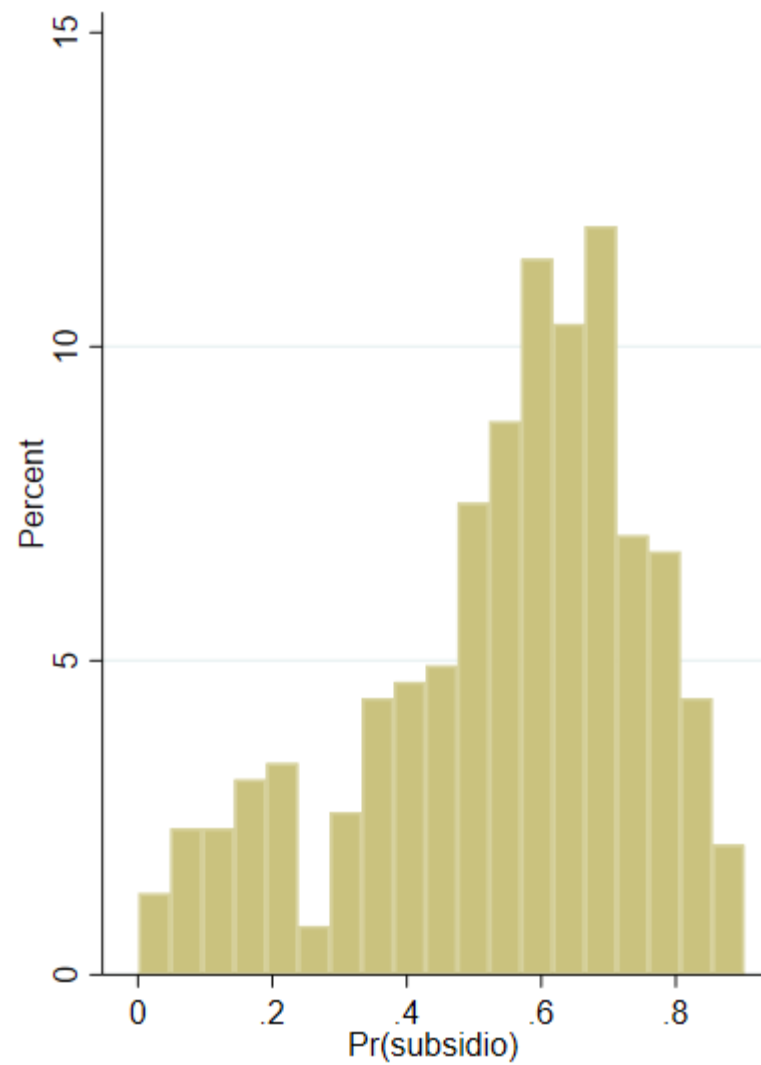
- Se paga un subsidio (en especie) a las familias para salud y alimentación de los hijos
- Queremos ver el impacto sobre el ahorro familiar porque esto puede generar mejores decisiones y puede aliviar la carga sobre las mujeres permitiendo trabajar

	mean t	sd t	mean c	sd c	t stat	p value
outcome	1016.99	334.26	714.73	299.43	-14.85	0.00
female	0.51	0.50	0.50	0.50	-0.19	0.85
puntaje	110.67	73.37	204.74	94.07	16.71	0.00
estrato	3.47	1.77	3.52	1.71	0.43	0.67
formality	0.50	0.50	0.49	0.50	-0.45	0.65
householdsize	6.10	1.95	5.91	2.01	-1.45	0.15

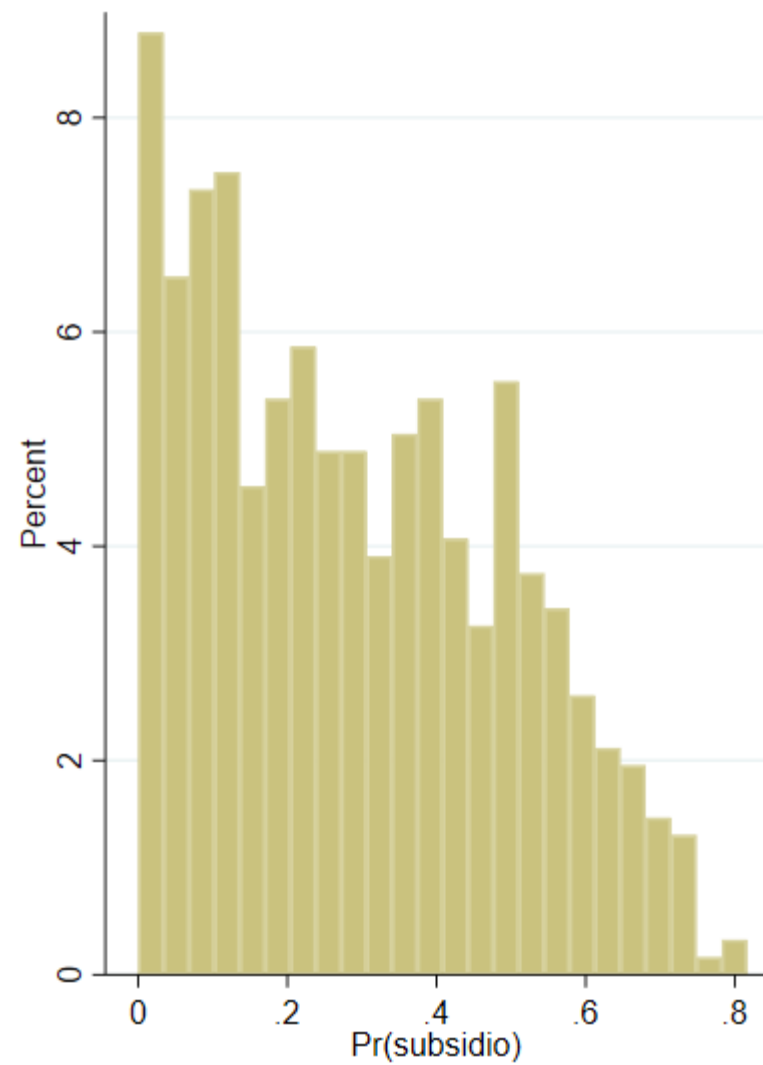
Un ejemplo



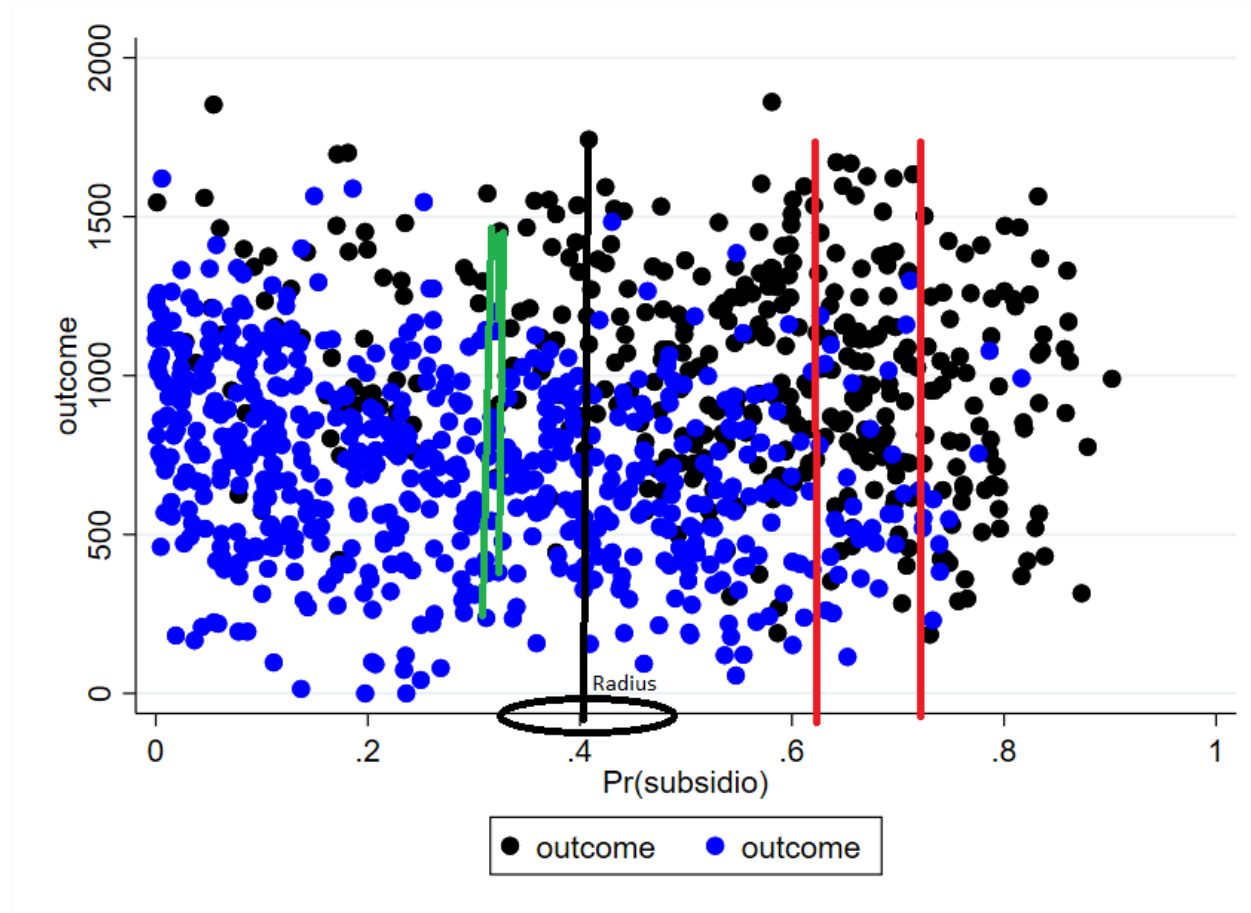
Distribution treated



Distribution untreated



# Matching



Let's go to Stata

# The Propensity Score theorem

If CIA:  $(Y^1, Y^0) \perp D|X$  then  $(Y^1, Y^0) \perp D|p(X)$ ,

donde  $p(X)$  es el propensity score ( $P(D = 1|X)$ )

$$\begin{aligned} & E[D|p(X), Y^1, Y^0] \\ &= 1P[D = 1|p(X), Y^1, Y^0] + 0P[D = 0|p(X), Y^1, Y^0] = \\ &= P[D = 1|p(X), Y^1, Y^0] = \\ &= E[E[D|X, p(X), Y^1, Y^0]|p(X), Y^1, Y^0] = \\ &= E[E[D|X, Y^1, Y^0]|p(X), Y^1, Y^0] = \\ &E[E[D|X]|p(X), Y^1, Y^0] = E[p(X)|p(X), Y^1, Y^0] = p(X) \end{aligned}$$

# Ponderación

$$\widehat{\delta_{ATE}} = \frac{1}{N} \sum_1^N Y_i \left( \frac{D_i - \widehat{p_i(X_i)}}{\widehat{p_i(X_i)}(1 - \widehat{p_i(X_i)})} \right)$$
$$\widehat{\delta_{ATT}} = \frac{1}{Nt} \sum_1^{Nt} Y_i \left( \frac{D_i - \widehat{p_i(X_i)}}{1 - \widehat{p_i(X_i)}} \right)$$

Ojo con los datos muy cerca de 0 y 1

# Proof

$$\begin{aligned} & E \left[ Y_i \left( \frac{D_i - p(X)}{p(X)(1 - p(X))} \right) \middle| X \right] \\ &= E \left[ Y_i \left( \frac{1 - p(X)}{p(X)(1 - p(X))} \right) \middle| X, D = 1 \right] p(X) \\ &+ E \left[ Y_i \left( \frac{-p(X)}{p(X)(1 - p(X))} \right) \middle| X, D = 0 \right] (1 - p(X)) = \\ &\quad = E(Y_i | X, D = 1) - E(Y_i | X, D = 0) \end{aligned}$$

# Notice something

- Matching ponderando por el propensity score

$$\bullet \widehat{\delta_{ATE}} = \frac{1}{N} \sum_1^N Y_i \left( \frac{D_i - \widehat{p_i(X_i)}}{\widehat{p_i(X_i)}(1 - \widehat{p_i(X_i)})} \right)$$

- Regresión OLS con modelo saturado de covariadas

$$\bullet \widehat{\delta_{ATE}} = \frac{\frac{1}{N} \sum_1^N Y_i (D_i - \widehat{p_i(X_i)})}{\frac{1}{N} \sum_1^N (D_i - \widehat{p_i(X_i)})^2} = \frac{\frac{1}{N} \sum_1^N Y_i (D_i - \widehat{p_i(X_i)})}{(1 - E[\widehat{p_i(X_i)}])E[\widehat{p_i(X_i)}]} = \frac{1}{N} \sum_1^N Y_i \left( \frac{D_i - \widehat{p_i(X_i)}}{(1 - E[\widehat{p_i(X_i)}])E[\widehat{p_i(X_i)}]} \right)$$



# Final thoughts

- Standard errors
- De vuelta a Lalonde
- AI y matching