

Métodos estadísticos y de inferencia causal

Isaac Meza López¹

¹ITAM

Junio 2019

1 Diseño de experimentos

- Simulación de poder

2 Inferencia Causal

- RCT
- IV - CF
- Matching
- Diff-in-Diff
- SCM

3 Attrition

- Manski bounds
- Lee bounds

4 Clasificación

- Gráficas de heterogeneidad por efectos fijos
- Clustering & PCA

- [AI17a] S. Athey and G.W. Imbens, *Chapter 3 - the econometrics of randomized experiments*, Handbook of Field Experiments (Abhijit Vinayak Banerjee and Esther Duflo, eds.), Handbook of Economic Field Experiments, vol. 1, North-Holland, 2017, pp. 73 – 140.
- [AI17b] Susan Athey and Guido W. Imbens, *The state of applied econometrics: Causality and policy evaluation*, Journal of Economic Perspectives **31** (2017), no. 2, 3–32.
- [AP08] Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, December 2008.

Los materiales y ésta presentación la pueden encontrar [aquí](#).

Diseño de experimentos

Poder estadístico ($1 - \beta$) es la verosimilitud/probabilidad de detectar cierto efecto cuando hay un efecto que detectar.

$$H_0 : \theta = 0$$

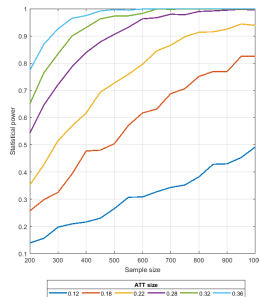
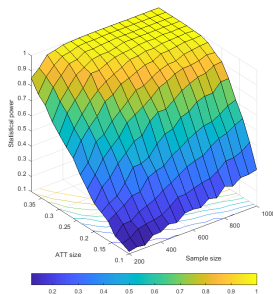
$$H_1 : \theta \neq 0$$

$$1 - \beta = Pr(\text{rechazar } H_0 \mid H_1 \text{ es verdadera})$$

Concluir		Realidad	
	H_0 TRUE	H_0 FALSE	
H_0 TRUE	$(1 - \alpha)$	Type I : β	
H_0 FALSE	Type II: α	$(1 - \beta)$	

El proceso de simulación de poder identifica 2 etapas:

- ④ DGP : $(X, Y) \sim F$
- ② Método de identificación : $\mathbb{E}[Y|X] = \theta^\top X$



Do files: `sim_AB_FS.1.do`, `simulation_iv.do`

Inferencia Causal

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

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

¿Cómo obtenemos efectos de tratamiento promedio - (ATE)?

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

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

¿Cómo obtenemos efectos de tratamiento promedio - (ATE)?

$$\begin{aligned} \mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i] &= \overbrace{\mathbb{E}[Y_{1i} \mid D_i = 1, X_i] - \mathbb{E}[Y_{0i} \mid D_i = 1, X_i]}^{\text{ATT}} \\ &\quad + \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1, X_i] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i]}_{\text{Bias}} \end{aligned}$$

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

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

¿Cómo obtenemos efectos de tratamiento promedio - (ATE)?

$$\begin{aligned} \mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i] &= \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1, X_i] - \mathbb{E}[Y_{0i} \mid D_i = 1, X_i]}_{\text{ATT}} \\ &+ \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1, X_i] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i]}_{\text{Bias}} \end{aligned}$$

Problema fundamental de la inferencia causal

$$\mathbb{E}[Y_{0i} \mid D_i = 1, X_i]$$

Table 1: Treatment Effects

	Months after treatment							
	Same day settlement				2 months	5 months	Long run	
	Phase 1		Phase 2		Phase 1/2			
	OLS				OLS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control (constant)	0.060*** (0.013)	0.034*** (0.011)	0.11*** (0.030)	0.10*** (0.030)	0.094*** (0.026)	0.15*** (0.043)	0.39*** (0.039)	0.45*** (0.049)
Calculator	0.051** (0.022)	0.019 (0.019)	0.047** (0.021)	0.0077 (0.019)	0.018 (0.014)	0.0035 (0.021)	-0.0069 (0.024)	-0.0025 (0.025)
Conciliator	0.054*** (0.019)	0.033* (0.018)			0.016 (0.019)	-0.0028 (0.023)	-0.030 (0.028)	-0.053 (0.036)
Emp present (EP)		0.14*** (0.050)		0.14* (0.072)	0.14*** (0.041)	0.11** (0.046)	0.094* (0.048)	0.070 (0.050)
Calculator#EP		0.16** (0.079)		0.16* (0.089)	0.16*** (0.056)	0.18*** (0.061)	0.16** (0.064)	0.14** (0.061)
Conciliator#EP		0.16** (0.074)			0.16** (0.071)	0.21*** (0.079)	0.27*** (0.075)	0.20** (0.078)
Observations	1074	1074	1092	1092	2166	2166	2166	2166
R-squared	0.0072	0.12	0.051	0.11	0.13	0.12	0.11	0.087
Court dummies	NO	NO	YES	YES	YES	YES	YES	YES
DepVarMean		0.095		0.20	0.15	0.19	0.32	0.43
InteractionVarMean		0.18				0.18		
Calc=Conc	0.88	0.53	-	-	0.94	0.82	0.79	0.40
Calc#EP=Conc#EP	-	0.98	-	-	1.00	0.58	0.68	0.085

IV - corrigiendo la endogeneidad

Consideremos un modelo $y_i = X_i\beta + \epsilon_i$

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$$

IV - corrigiendo la endogeneidad

Consideremos un modelo $y_i = X_i\beta + \epsilon_i$

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$$

Variables instrumentales Z

- Primera etapa fuerte : $X_i = Z_i\gamma + \nu_i$
- Restricción de exclusión : $Z'e = 0$

IV - corrigiendo la endogeneidad

Consideremos un modelo $y_i = X_i\beta + \epsilon_i$

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon$$

Variables instrumentales Z

- Primera etapa fuerte : $X_i = Z_i\gamma + \nu_i$
- Restricción de exclusión : $Z'e = 0$

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y \rightarrow \beta$$

Estimación - 2SLS

(i) $X = Z\gamma + \nu$

$$\begin{aligned}\hat{\gamma} &= (Z'Z)^{-1}Z'X \\ \hat{X} &= X\hat{\gamma} = \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z}X\end{aligned}$$

(ii) $y = \hat{X}\beta + \epsilon$

$$\hat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_Zy = \hat{\beta}_{GMM}$$

- Lluvia

- Lluvia - no funcionó

- Lluvia - **no funcionó**
- Distancia a la junta

- Lluvia - no funcionó
- Distancia a la junta - no funcionó

Es un método de dos etapas *generalizado*, pues explota la estructura de la variable endógena - por lo que obtenemos residuos generalizados.

$y = X\beta + \epsilon$	Ec. estructural
$X = Z\gamma + \nu$	Primera etapa
$\mathbb{E}[Z'\nu] = 0$	Ortogonalidad
$\epsilon = \nu\rho + u$ $\mathbb{E}[\nu u] = 0$	Residuo generalizado

De la última ecuación: $\rho = \mathbb{E}[\nu\nu']^{-1}\mathbb{E}[\nu\epsilon]$ y notemos que

$$\{\epsilon, \nu\} \perp Z \implies u \perp Z$$

$$\therefore u \perp X$$

Entonces,

CF

$$y = X\beta + \nu\rho + u$$

- 1 En el modelo lineal básico con coeficientes constantes, donde las VEE aparecen linealmente, y donde uso formas reducidas lineales, CF es lo mismo que 2SLS. Pero el primero proporciona una prueba simple y robusta de la hipótesis nula de que X es exógena : $\rho = 0$
- 2 Cuando exploto características especiales de la VEE, por ejemplo, reconozco que es una variable binaria, CF es probablemente más eficiente que 2SLS pero, en términos de consistencia, el enfoque de CF suele ser menos robusto que el de IV.
- 3 En modelos con múltiples funciones no lineales de VEE, el enfoque CF maneja parsimoniosamente la endogeneidad y proporciona pruebas de exogeneidad simples.

Control function III

Months after treatment									
	Same day settlement				2 months	5 months	Long run	Same day	
	Phase 1		Phase 2		Phase 1/2				
	OLS				OLS				CF OLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Control (constant)	0.060*** (0.013)	0.034*** (0.011)	0.11*** (0.030)	0.10*** (0.030)	0.094*** (0.026)	0.15*** (0.043)	0.39*** (0.039)	0.45*** (0.049)	0.053 (0.040)
Calculator	0.051** (0.022)	0.019 (0.019)	0.047** (0.021)	0.0077 (0.019)	0.018 (0.014)	0.0035 (0.021)	-0.0069 (0.024)	-0.0025 (0.025)	0.0084 (0.014)
Conciliator	0.054*** (0.019)	0.033* (0.018)			0.016 (0.019)	-0.0028 (0.023)	-0.030 (0.028)	-0.053 (0.036)	0.023 (0.019)
Emp present (EP)		0.14*** (0.050)		0.14* (0.072)	0.14*** (0.041)	0.11** (0.046)	0.094* (0.048)	0.070 (0.050)	0.47** (0.21)
Calculator#EP		0.16** (0.079)		0.16* (0.089)	0.16*** (0.056)	0.18*** (0.061)	0.16** (0.064)	0.14** (0.061)	0.16*** (0.054)
Conciliator#EP		0.16** (0.074)			0.16** (0.071)	0.21*** (0.079)	0.27*** (0.075)	0.20** (0.078)	0.17** (0.074)
Control Function									-0.19 (0.12)
Observations	1074	1074	1092	1092	2166	2166	2166	2166	2166
R-squared	0.0072	0.12	0.051	0.11	0.13	0.12	0.11	0.087	0.135
Court dummies	NO	NO	YES	YES	YES	YES	YES	YES	YES
DepVarMean		0.095		0.20	0.15	0.19	0.32	0.43	0.15
InteractionVarMean		0.18					0.18		
Calc=Conc	0.88	0.53	-	-	0.94	0.82	0.79	0.40	0.47
Calc#EP=Conc#EP	-	0.98	-	-	1.00	0.58	0.68	0.085	0.91

Do files: `treatment_effects.do`, `treatment_effects_IV_CF.do`

Recordemos el problema fundamental de inferencia causal: ¿ $\mathbb{E}[Y_{0i} \mid D_i = 1, X_i]$?

Supongamos : *strong ignorability*

CIA : $(Y_i \mid X) \perp D_i$

Overlap : $0 < e(x) := \mathbb{E}[D_i \mid X_i = x] < 1$ para todo x en el soporte de X

Recordemos el problema fundamental de inferencia causal: $\mathbb{E}[Y_{0i} \mid D_i = 1, X_i]$?

Supongamos : *strong ignorability*

CIA : $(Y_i \mid X) \perp D_i$

Overlap : $0 < e(x) := \mathbb{E}[D_i \mid X_i = x] < 1$ para todo x en el soporte de X

Sea

$$m(i) = \operatorname{argmin}_{j: D_j \neq D_i} \|X_i - X_j\|$$

$$\begin{aligned}\hat{Y}_i(0) &= \begin{cases} Y_i^{obs} & \text{si } D_i = 0 \\ Y_{m(i)}^{obs} & \text{si } D_i = 1 \end{cases} & \hat{Y}_i(1) &= \begin{cases} Y_{m(i)}^{obs} & \text{si } D_i = 0 \\ Y_i^{obs} & \text{si } D_i = 1 \end{cases} \\ \hat{X}_i(0) &= \begin{cases} X_i^{obs} & \text{si } D_i = 0 \\ X_{m(i)}^{obs} & \text{si } D_i = 1 \end{cases} & \hat{X}_i(1) &= \begin{cases} X_{m(i)}^{obs} & \text{si } D_i = 0 \\ X_i^{obs} & \text{si } D_i = 1 \end{cases}\end{aligned}$$

El estimador de matching está dado por:

$$\hat{\tau} = \frac{1}{N} \sum_i \hat{Y}_i(1) - \hat{Y}_i(0)$$

Recordemos el problema fundamental de inferencia causal: ¿ $\mathbb{E}[Y_{0i} \mid D_i = 1, X_i]$?

Supongamos : *strong ignorability*

CIA : $(Y_i \mid X) \perp D_i$

Overlap : $0 < e(x) := \mathbb{E}[D_i \mid X_i = x] < 1$ para todo x en el soporte de X

Sea

$$m(i) = \operatorname{argmin}_{j: D_j \neq D_i} \|X_i - X_j\|$$

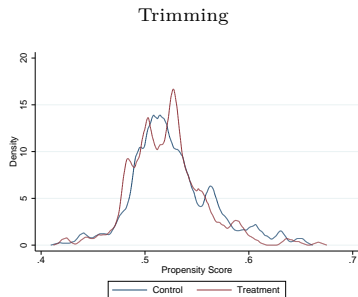
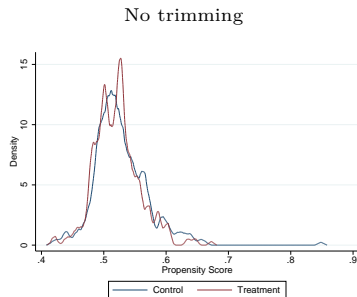
$$\begin{aligned}\hat{Y}_i(0) &= \begin{cases} Y_i^{obs} & \text{si } D_i = 0 \\ Y_{m(i)}^{obs} & \text{si } D_i = 1 \end{cases} & \hat{Y}_i(1) &= \begin{cases} Y_{m(i)}^{obs} & \text{si } D_i = 0 \\ Y_i^{obs} & \text{si } D_i = 1 \end{cases} \\ \hat{X}_i(0) &= \begin{cases} X_i^{obs} & \text{si } D_i = 0 \\ X_{m(i)}^{obs} & \text{si } D_i = 1 \end{cases} & \hat{X}_i(1) &= \begin{cases} X_{m(i)}^{obs} & \text{si } D_i = 0 \\ X_i^{obs} & \text{si } D_i = 1 \end{cases}\end{aligned}$$

El estimador de matching está dado por:

$$\hat{\tau} = \frac{1}{N} \sum_i \hat{Y}_i(1) - \hat{Y}_i(0)$$

Se puede mejorar el sesgo de este estimador usando regresión lineal para *ajustar el sesgo* asociado con diferencias entre $\hat{X}_i(0)$ y $\hat{X}_i(1)$.

(I) 'Verificar' overlap - Trimming procedure



$$\hat{e}(x) = \frac{\exp \beta' x}{1 + \exp \beta' x}$$

(II) Balance

Table 2: Balance

	Control	Treatment	p-value
Entitlement by law	60234.96 (3400.38)	57567.95 (4098.63)	0.62
Public lawyer	0.08 (0.01)	0.09 (0.01)	0.77
Woman	0.45 (0.02)	0.45 (0.03)	0.86
At will worker	0.07 (0.01)	0.06 (0.01)	0.44
Tenure	3.82 (0.25)	3.47 (0.22)	0.29
Daily wage	535.31 (32.76)	514.78 (40.85)	0.7
Weekly hours	58.5 (0.81)	56.79 (0.73)	0.12
Observations	416	377	

(III) ‘Evaluar’ CIA

Table 3: Pseudo-treatment effect

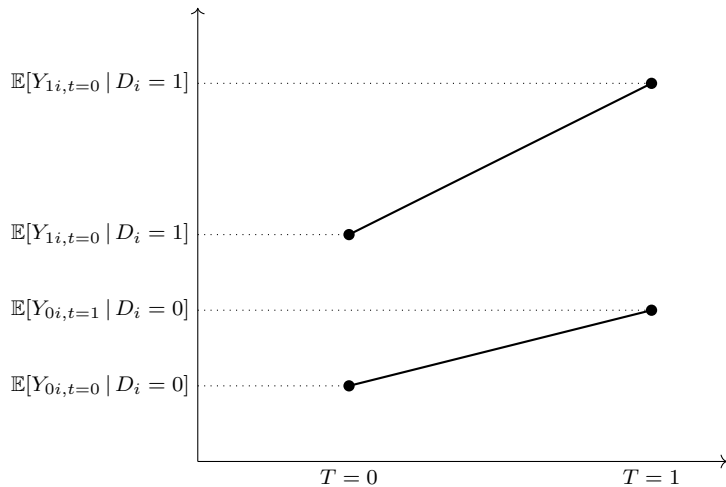
Pseudo treatment effect. Nearest-neighbor matching			
Phase 1/2			
	Entitlement	Daily wage	Tenure
	(1)	(2)	(3)
ATE	28.3 (1162.4)	-2.3 (11)	-0.2 (.2)
% ATE	0.05	-0.43	-5.22
Baseline mean	60342.9	536.2	3.8
Obs		377	
Obs HD		415	
Bias adjustment	YES	YES	YES
Matches	[1-3]	[1-3]	[1-3]

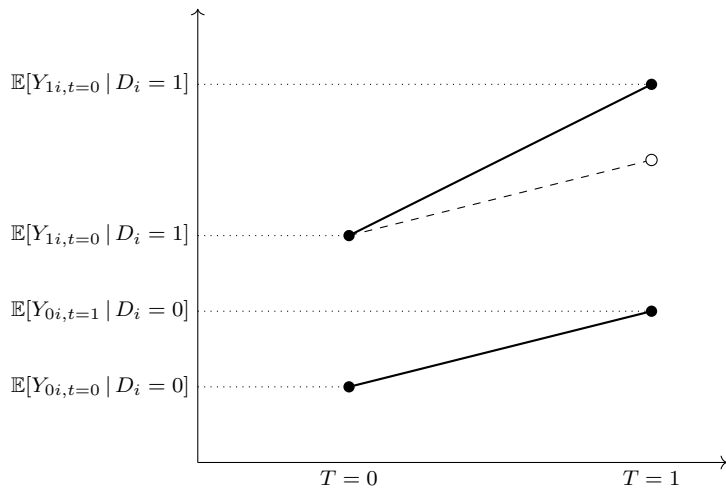
(III) Análisis

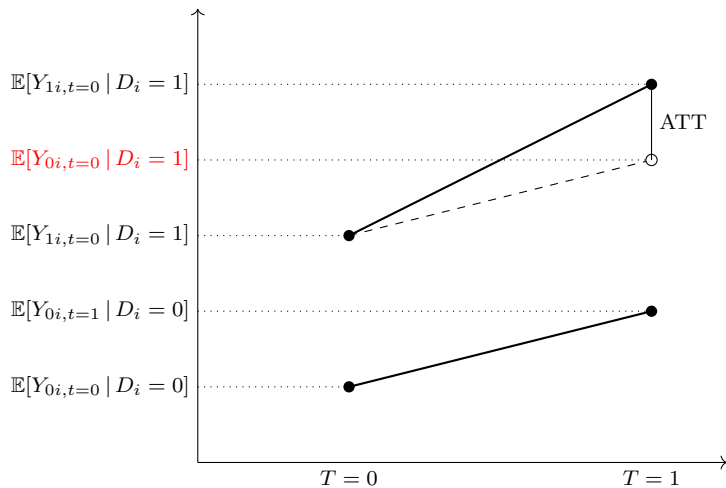
Table 4: Matching

Treatment effect. Nearest-neighbor matching						
Phase 1/2						
	Variable matching				PSM	
	(1)	(2)	(3)	(4)	(5)	(6)
ATE	2580 (1939)	3536** (1715)	3242* (1934)	3939** (1725)	3698* (1900)	3995** (1554)
% ATE	34	47	43	52	49	53
Baseline mean				7598		
Obs				377		
Obs HD				415		
Bias adjustment	NO	NO	YES	YES	-	-
Matches	[1-1]	[1-3]	[1-1]	[1-3]	[1-1]	[1-3]

Do files: `settlement_conciliator_matching.do`





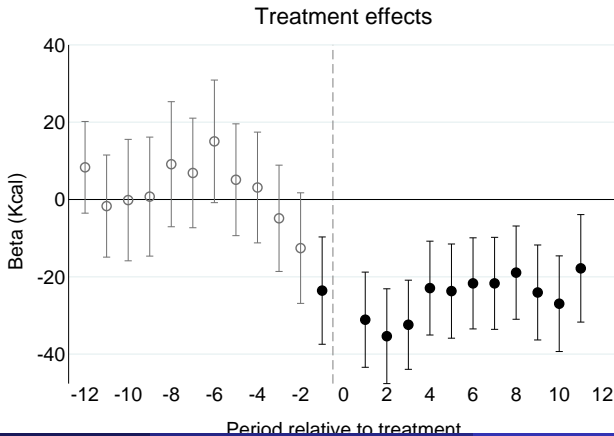


Impuesto a bebidas azucaradas

Regresión DiD con FE:

$$C_{it}^k = \alpha_i + \gamma_t + \sum_{j=-12}^{12} \beta_k T_i \times I(t = j) + \nu_{it}$$

donde $T_i = 1$ si i está en el grupo más expuesto



Synthetic Control Methods (SCM) I

Sean (y_{tn}^0, y_{tn}^1) las observaciones potenciales para la unidad n al tiempo t .

$$y_{tn} = D_{tn}y_{tn}^1 + (1 - D_{tn})y_{tn}^0, \quad D_{tn} = \begin{cases} 1 & \text{if } t \geq T_0, n = 0 \\ 0 & \text{otherwise} \end{cases}$$

El efecto de tratamiento es : $\tau_{tn} \equiv y_{tn}^1 - y_{tn}^0$

El supuesto clave en SCM es:

Existen pesos $\beta_n \in [0, 1]$ para $n = 1, \dots, N$ tales que

$$y_{t0}^0 = \sum_{n=1}^N \beta_n y_{tn}^0$$

para $t = 1 \dots, T$; y los pesos suman uno: $\sum_{n=1}^N \beta_n = 1$.

El estimador en $t = T_0, \dots, T$ está dado por:

$$\tau_t = y_{t0}^1 - \sum_{n=1}^N \beta_n y_{t0}^0$$

Denotando por $x_t \equiv (y_{t1}, \dots, y_{tN})^\top$ a el vector de las observaciones para los controles. Podemos considerar el modelo de regresión

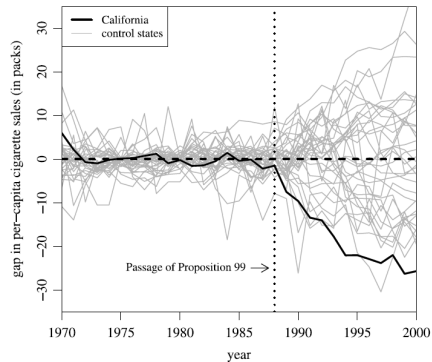
$$y_{t0} = \beta^\top x_t + u_{t0} \quad t = 1, \dots, T_0 \quad (1)$$

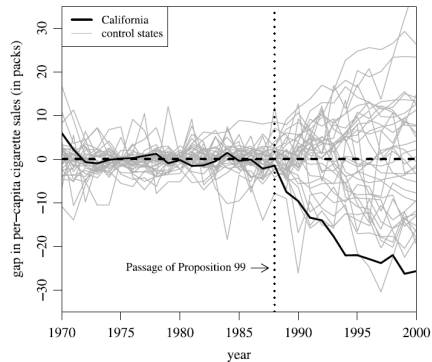
y la estimación por lo tanto está dada por:

$$\begin{aligned} \min \quad & \sum_{t=1}^{T_0} (y_{t0} - \beta^\top x_t)^2 \\ \text{s.t.} \quad & \|\beta\|_1 = 1 \\ & \beta_i \geq 0 \quad i = 1, \dots, n \end{aligned} \quad (2)$$

El principal problema con esta metodología es la dificultad para conducir **inferencia**, es decir, hay poco o ningún conocimiento sobre la distribución asintótica del estimador SCM, o su **intervalo de confianza**.

- (i) *Enfoques de población finita* : Supuesto de que las unidades de tratamiento se asignan aleatoriamente y usan placebos, pruebas de permutación o alguna variante que explota la estructura de datos del panel, para realizar inferencia, que son llamados
- (ii) *Enfoques asintóticos* : Los supuestos clave hacen que el número de individuos o períodos de tiempo tienden a infinito.





Theorem

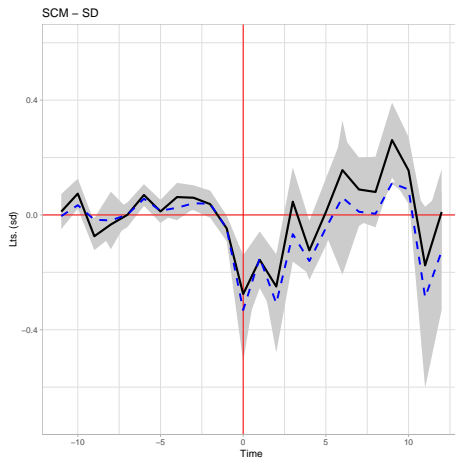
La región de confianza para el estimador SCM está dada por:

$$B\left(\hat{\tau}_t, \mathbf{r} ||x_t|| + \sigma z_{(1-\alpha/2)}\right)$$

$$\text{donde } \mathbf{r} = \frac{||\nabla R_{T_0}(\hat{\beta})|| + \sqrt{||\nabla R_{T_0}(\hat{\beta})||^2 - 2|\Omega_{\hat{\beta}}| \left(R_{T_0}(\hat{\beta}) - T_0^{-1} \chi_{1-\alpha}\right)}}{|\Omega_{\hat{\beta}}|}$$

Explotamos la estructura de panel y las múltiples unidades tratadas.

Explotamos la estructura de panel y las múltiples unidades tratadas.



Rscript: SCM_synth.R

Attrition

- Ningún supuesto sobre el mecanismo de selección
- La variable dependiente necesita estar acotada
- NA imputados de acuerdo al mínimo y máximo valor posible
- Cotas no informativas :(

Sea S_i una dummy identificando a los ‘no-attriters’.

$$¿E[Y_{1i} | S_i = 0, D_i = 1] \quad E[Y_{0i} | S_i = 0, D_i = 0]?$$

Worst-case scenario : $E[Y_{1i} | S_i = 0, D_i = 1] = 0$ y $E[Y_{0i} | S_i = 0, D_i = 0] = 1$

$$\begin{aligned} MB^L &= P(S_i = 1 | D_i = 1)E(Y_i | D_i = 1, S_i = 1) \\ &\quad - [P(S_i = 1 | D_i = 0)E(Y_i | D_i = 0, S_i = 1) + P(S_i = 0 | D_i = 0)] \end{aligned}$$

análogamente:

$$\begin{aligned} MB^U &= P(S_i = 1 | D_i = 1)E(Y_i | D_i = 1, S_i = 1) + P(S_i = 0 | D_i = 1) \\ &\quad - P(S_i = 1 | D_i = 0)E(Y_i | D_i = 0, S_i = 1) \end{aligned}$$

- RAT : $(Y_i, S_i) \perp D_i$
- Monotonicidad : $Pr(S_1 \geq S_0) = 1$
Asignación de tratamiento sólo afecta *attrition* en una dirección solamente.

El objetivo es encontrar cotas

$$LB^L \leq \mathbb{E}[Y_1 - Y_0 \mid S_0 = 1, S_1 = 1] \leq LB^U$$

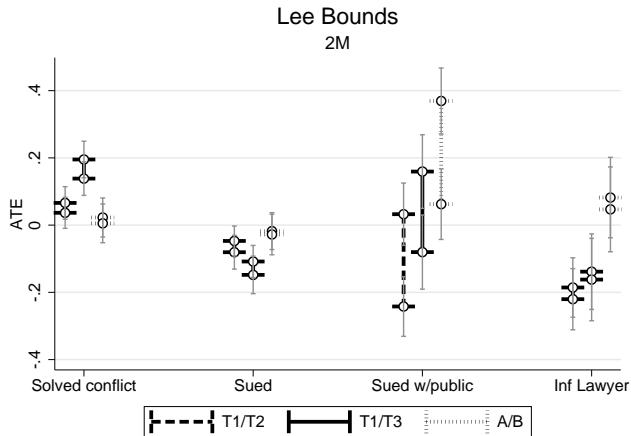
Theorem

Bajo RAT, monotonicidad y $PR(S = 1 \mid D = 0) \neq 0$

$$LB^L = \mathbb{E}[Y \mid D = 1, S = 1, Y \leq F^{-1}(1 - p)] - \mathbb{E}[Y \mid D = 0, S = 1]$$

$$LB^U = \mathbb{E}[Y \mid D = 1, S = 1, Y \geq F^{-1}(p)] - \mathbb{E}[Y \mid D = 0, S = 1]$$

$$\text{donde } p = \frac{Pr(S=1 \mid D=1) - Pr(S=1 \mid D=0)}{Pr(S=1 \mid D=1)}.$$



Do files: `plot_lee_bounds.do`

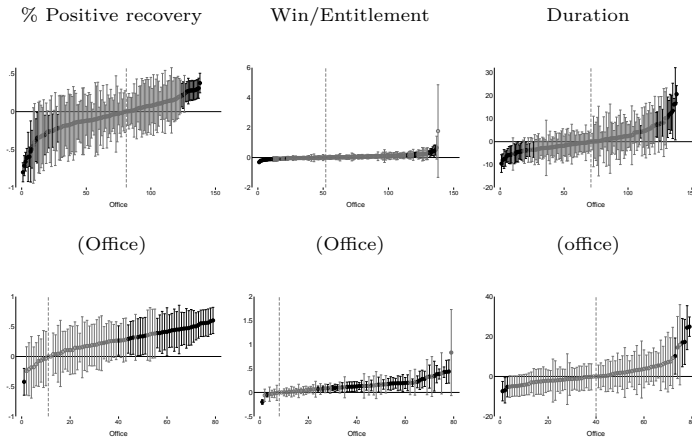
Clasificación

Table 5: Predicciones

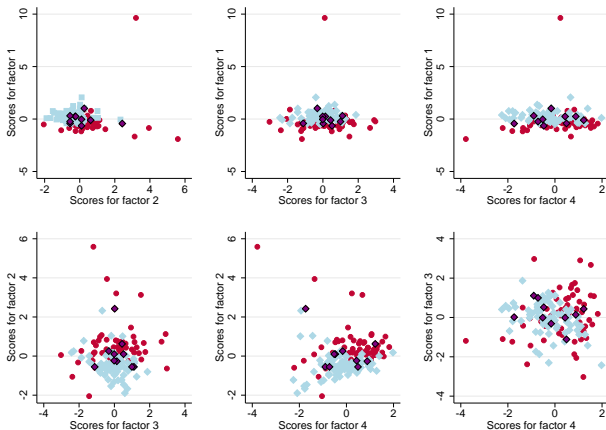
Outcome		Judge ruling prediction		End mode prediction		Payment prediction		Total	
Lawyer		Lawyer	Calc	Lawyer	Calc	Lawyer	Calc	Lawyer	Calc
1	12.8%	44.4%	27.8%	61.1%	55.6%	19.5%	19.5%	34.5%	34.3%
2	11.2%	62.5%	68.8%	87.5%	62.5%	0%	12.8%	40.3%	48%
3	15.2%	56.3%	62.5%	62.5%	62.5%	22.7%	17.3%	39.2%	47.4%
4	35.2%	45.5%	50%	36.4%	77.3%	15.3%	18.8%	33.1%	48.7%
5	5.6%	33.3%	75%	33.3%	62.5%	9.9%	19.8%	20.5%	52.4%
6	9.6%	60%	70%	70%	65%	0%	18.6%	34.9%	51.2%
7	16%	50%	55.6%	55.6%	72.2%	13.9%	13.9%	33.9%	47.2%
8	13.6%	54.5%	77.3%	86.4%	72.7%	21.4%	21.4%	44%	57.1%

Do files: lawyer_dataset.do, lawyer_dataset_predicc.do, cleaning_base_calidad.do, calif_abogados_pagos.do

Figure 2: Heterogeneous outcome graphs



Do files: `fe_betas_pres.do`, `fe_betas_desp.do`



Do files: lawyer_dataset_collapsed.do, cluster_analysis.do

Gracias :)

- (1) Install [Node.js](#) 8.11.3 LTS (or current)
- (2) Extract the file `osrm_Release.zip` (link from my [Dropbox](#))
- (3) Copy the file [mexico-latest.osm.pbf](#) in the `osrm_Release` folder.
- (4) Open Node.js command prompt
 - ❶ Change directory to the `osrm_Release` folder
`cd ~/osrm_Release`
 - ❷ Extracting the road network
`osrm-extract mexico-latest.osm.pbf`
`osrm-contract mexico-latest.osm.pbf`

This previous steps only need to be done once. This sets the road network for Mexico.

In order to launch OSRM as a local instance the following steps are needed.

- ❶ Change directory to the osrm_Release folder in Node.js command prompt
`cd ~/osrm_Release`
- ❷ Set the local OSRM instance
`osrm-routed --max-table-size=1500 mexico-latest.osrm`
- ❸ In R, run the script:

```
1 # PACKAGES
2 library(tidyverse) # data wrangling
3 library(stringr)  # work with strings
4 library(sf)       # geospatial
5 library(geosphere) # calculate distances
6 library(haven)    # read/write .dta (Stata)
7 library(here)     # use relative filepaths
8 library(osrm)     # calculate road distances
9 library(assertthat)
10
11
12 # old spatial packages since osrm works with those
13 library(sp) #spatial objects; used by rgdal
14
15
16 # Read in data set
17 dataset = read_dta(here::here("folder", "dta.dta"))
18
19
20 # Convert to SpatialPointsDataFrame for use with osrmTable
21 dataset_spdf <- dataset %>% as('Spatial')
```

```
2
3
4 # Use OSRM's server
5 options(osrm.server = "http://router.project-osrm.org/")
6 # NOTE: above uses OSRM's server. If we try one with more than 10,000 queries:
7
8 # We get the error:
9 # OSRM returned an error:
10 # Error: The public OSRM API does not allow results with a number of durations
11 # higher than 10000. Ask for fewer durations or use your own server and set
12 #   its
13 # --max-table-size option.
14
15 # Now try it on a local server to avoid the 10,000 query limit:
16 # First follow the instructions here to install and build the OSRM server:
17 # https://datawookie.netlify.com/blog/2017/09/building-a-local-osrm-instance/
18 # or:
19 # https://github.com/Project-OSRM/osrm-backend/wiki/Running-OSRM
20
21 # Once the local server is running:
22 options(osrm.server = "http://localhost:5000/")
23 #Compute distances
24 distances <- osrmTable(src = src, dst = dst)
25 #Check the function osrmViaroute for pairwise comparisons.
```

./Rscripts/geocode.R


```
C:\Users\xps-seira\Dropbox\repos\osrm\osrm-backend>osrm-routed
osrm-routed <base.osrm> [<options>]:

Options:
  -v [ --version ]           Show version
  -h [ --help ]             Show this help message
  -l [ --verbosity ] arg (=INFO) Log verbosity level: NONE, ERROR,
                              WARNING, INFO, DEBUG
  --trial [=arg(=1)]        Quit after initialization

Configuration:
  -i [ --ip ] arg (=0.0.0.0) IP address
  -p [ --port ] arg (=5000)  TCP/IP port
  -t [ --threads ] arg (=4)  Number of threads to use
  -s [ --shared-memory ] [=arg(=1)] (=0) Load data from shared memory
  -a [ --algorithm ] arg (=CH) Algorithm to use for the data. Can be
                              CH, CoreCH, MLD.
  --max-viaroute-size arg (=500) Max. locations supported in viaroute
                              query
  --max-trip-size arg (=100)   Max. locations supported in trip query
  --max-table-size arg (=100)  Max. locations supported in distance
                              table query
  --max-matching-size arg (=100) Max. locations supported in map
                              matching query
  --max-nearest-size arg (=100) Max. results supported in nearest query
  --max-alternatives arg (=3)  Max. number of alternatives supported
                              in the MLD route query
  --max-matching-radius arg (=5) Max. radius size supported in map
                              matching query
```

De regreso a los [instrumentos](#).

Definición de grupo expuesto

Definir un grupo de tratamiento / control puro. Lo haremos eligiendo una partición óptima de la distribución total del gasto sujeto a impuestos, ya que es probable que los grandes gastadores sean más sensibles a un cambio de precio en SD, por lo que los definiremos como el grupo de tratamiento.

$$\begin{aligned} \min_{H,L} \quad & \sum_{t=-12}^{-2} |\beta_t| \\ \text{s.t} \quad & (\beta_t)_{-12 \leq t \leq 12} = \operatorname{argmin} \left\{ \left(y_{it} - \sum_{k=-12}^{12} \alpha_k \mathbb{1}(t=k) - \sum_{k=-12}^{12} \beta_k \mathbb{1}(i=T, k=t) + \gamma \mathbb{1}(i=T) - \lambda_i \right)^2 \right\} \\ & T = \mathbb{1}(x_i \geq H) \\ & C = \mathbb{1}(x_i \leq L) \\ & \min(x_i) \leq L \leq H \leq \max(x_i) \end{aligned}$$

De regreso a [DiD](#).