

Revisão de Probabilidade

Prof. Raphael Corbi
Monitor: Alan Leal

FEA/USP

Probabilidade: Revisão

- Um evento $A \in \Omega$ tem uma probabilidade $P(A)$ de ocorrência. Na abordagem axiomática da probabilidade, a função P tem três propriedades.
 - $P(\Omega) = 1$
 - $0 \leq P(A) \leq 1$
 - Para qualquer eventos A_i disjuntos (ou mutuamente exclusivos) contáveis, temos que: $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Processo Aleatório e Variável Aleatória

- Um processo aleatório é aquele cujos resultados dependem da realização de um experimento probabilístico.
- Por sua vez, uma variável aleatória é uma função $f : S \rightarrow R$, ou seja, é uma função que leva do espaço amostral, espaço onde os resultados do experimento probabilístico ocorrem, e leva à reta real.
- As variáveis aleatórias podem de forma simplificada serem discretas ou contínuas (ou mistas, que não iremos trabalhar aqui).
- Exemplo de variáveis aleatórias:
 - ① Discretas: distribuição de Bernoulli, Binomial, Poisson, dentre outras.
 - ② Contínuas: distribuição normal, qui-quadrado, t de Student, F de Snedecor, dentre outras.

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - ➊ Qual é a probabilidade da retirada de uma bola vermelha? $1/5$
 - ➋ Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - ➊ Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - ➋ Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - ➌ $P(X = 2|n = 6, p = 0.2) = \binom{6}{2} 0.2^2 (1 - 0.2)^{6-2} = 0.2478$
 - ➍ Isso te lembra alguma distribuição?

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - 1 Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - 2 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - 3 $P(X = 2|n = 6, p = 0.2) = \binom{6}{2}0.2^2(1 - 0.2)^{6-2} = 0.2478$
 - 4 Isso te lembra alguma distribuição?
 - 5 Binomial

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - 1 Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - 2 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - 3 $P(X = 2|n = 6, p = 0.2) = \binom{6}{2} 0.2^2 (1 - 0.2)^{6-2} = 0.2478$
 - 4 Isso te lembra alguma distribuição?
 - 5 Binomial
 - 6 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos sucesso na sexta retirada, sem reposição?

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - ① Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - ② Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - ③ $P(X = 2|n = 6, p = 0.2) = \binom{6}{2} 0.2^2 (1 - 0.2)^{6-2} = 0.2478$
 - ④ Isso te lembra alguma distribuição?
 - ⑤ Binomial
 - ⑥ Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos sucesso na sexta retirada, sem reposição?
 - ⑦ $P(X = 1|10, 2, 6) = \frac{\binom{2}{1} \binom{10-2}{6-1}}{\binom{10}{6}} = 0,53333$

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - 1 Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - 2 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - 3 $P(X = 2|n = 6, p = 0.2) = \binom{6}{2} 0.2^2 (1 - 0.2)^{6-2} = 0.2478$
 - 4 Isso te lembra alguma distribuição?
 - 5 Binomial
 - 6 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos sucesso na sexta retirada, sem reposição?
 - 7 $P(X = 1|10, 2, 6) = \frac{\binom{2}{1} \binom{10-2}{6-1}}{\binom{10}{6}} = 0,53333$
 - 8 Isso te lembra alguma distribuição?

Exemplos de variáveis aleatórias: exemplo de um experimento

- Há uma urna com 10 bolas, as quais 3 são pretas, 5 verdes e 2 vermelhas.
 - 1 Qual é a probabilidade da retirada de uma bola vermelha? 1/5
 - 2 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos dois sucessos em seis retiradas?
 - 3 $P(X = 2|n = 6, p = 0.2) = \binom{6}{2} 0.2^2 (1 - 0.2)^{6-2} = 0.2478$
 - 4 Isso te lembra alguma distribuição?
 - 5 Binomial
 - 6 Considerando como sucesso a retirada da bola vermelha, qual é a probabilidade de termos sucesso na sexta retirada, sem reposição?
 - 7 $P(X = 1|10, 2, 6) = \frac{\binom{2}{1} \binom{10-2}{6-1}}{\binom{10}{6}} = 0,53333$
 - 8 Isso te lembra alguma distribuição?
 - 9 Hipergeométrica

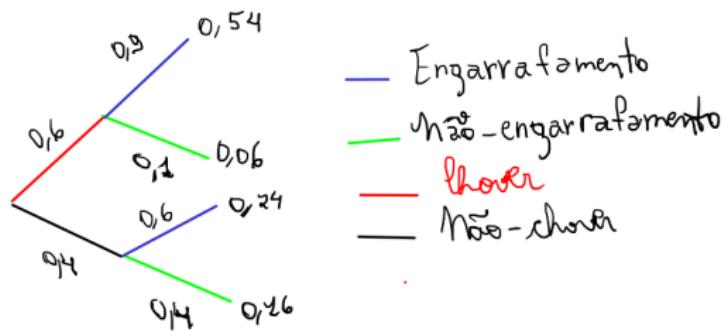
Independência Estatística

- O que é a independência estatística?
- Definição formal: $P(A|B) = P(A)$.
- Eventos mutuamente exclusivos são independentes?
- Se um evento não ocorre quando outro sempre ocorre, eles são independentes?

Árvores de probabilidade

- Representação de eventos encadeados de alguma forma (ou melhor dizendo são condicionados).
- Suponha que haja uma probabilidade de 0,6 de chuva e 0,4 que não chova. No caso de haver chuva, a probabilidade de engarrafamento é de 0,9. No caso de não haver chuva, a chance de engarrafamento é de 0,6. Qual é a probabilidade de não engarrafamento? 0,22

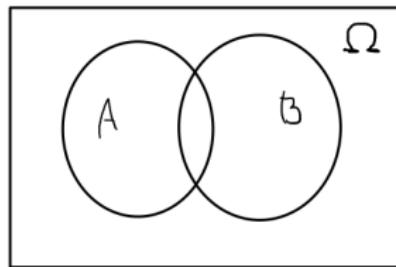
Figure: Árvore de probabilidade



Diagramas de Venn

- O que são os diagramas de Venn?

Figure: Diagrama de Venn



- União, intersecção, complementar de um conjunto e diferença de conjuntos.
- $x \in A \cup B$ significa $x \in A$ OU $x \in B$
- $x \in A \cap B$ significa $x \in A$ E $x \in B$
- $x \in \bar{A}$ significa que $x \notin A$

Tabelas de probabilidade

- O que são as tabelas de probabilidades.
- Exemplo e análise.

	Chuva	Não chuva	Total
Engarrafamento	0,54	0,24	0,78
Não-engarrafamento	0,06	0,16	0,22
Total	0,6	0,4	1

Operador somatório

- Definição: $\sum_{i=1}^N x_i = x_1 + x_2 + \cdots + x_N$
- Propriedades:
 - ① $\sum_{i=1}^N x_i \pm y_i = \sum_{i=1}^N x_i \pm \sum_{i=1}^N y_i =$
 - ② $\sum_{i=1}^N a * x_i = a * \sum_{i=1}^N x_i$

Valor esperado

- $E(x) = \int_D xf(x)dx$ ou $E(X) = \sum_D x * p(x)$
- Propriedade:
 - $E[aX+Y] = aE[X] + E[Y]$

Variância

- $V(x) = \int_D (x - E(X))^2 f(x) dx$ ou $V(X) = \sum_D [x - E(X)]^2 * p(x)$
- Propriedade:
 - $V(aX + Y + b) = a^2 V(X) + V(Y) + a * Cov(X, Y)$

Covariância

- $\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$
- Propriedade:
 - ① $\text{cov}(X, c) = 0$
 - ② $\text{cov}(X, X) = V(X)$
 - ③ $\text{cov}(X, Y) = \text{cov}(Y, X)$
 - ④ $\text{cov}(aX, bY) = ab * \text{cov}(X, Y)$

Modelo populacional da regressão linear

- $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$
- Hipóteses adjacentes:
- O modelo é linear nos parâmetros e está corretamente especificado
- Exogeneidade estrita (independência na média de ε): $E[\varepsilon|X] = 0$
- $\text{posto}(X)=k$
- Erros esféricos: $Var[\varepsilon|X] = \sigma^2 I$
- Erros normalmente distribuídos.

Principais distribuições e suas propriedades

- Discretas: Binomial, Geométrica, Hipergeométrica, dentre outras.
- Contínuas: Normal, t de Student, F de Snedecor, logística, dentre outras.

Distribuições Discretas: Hipergeométrica

- Exemplo: há uma urna com N bolas com M vermelhas e $N-M$ verdes. São retiradas simultaneamente K bolas (ou seja, sem reposição). Qual é a probabilidade de que x dessas K bolas sejam vermelhas?
- Temos $\binom{N}{K}$ grupos possíveis de grupos de K bolas retiradas de N . Isso nos deixa com $\binom{N-M}{K-x}$ grupos possíveis de $K-x$ bolas em $N-M$. Por fim, $\binom{M}{x}$ é o número de combinações possíveis de M bolas vermelhas em x sucessos.
- Logo, a probabilidade de que haja x sucessos nesses experimento é dado pelo seguinte cálculo:

$$P(X = x | N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, x = 0, 1, 2, \dots, K$$

- $E(X) = \frac{KM}{N}$
- $V(X) = \frac{KN}{M} \left(\frac{(N-M)(N-K)}{N(N-1)} \right)$

Distribuições Discretas: Binomial

- Qual é a probabilidade de ocorrer duas caras em dez lançamentos de uma moeda não viciada?
- A distribuição binomial é uma realização de n trials de Bernoulli, ou seja, um evento que produz apenas dois resultados, sucesso e fracasso.
- Além disso, na distribuição binomial, cada experimento de Bernoulli independe um do outro. Isso implica que a probabilidade de y sucessos ocorrer em n trials de Bernoulli, nos quais a probabilidade de sucesso é individualmente dada por p é de:

$$P(Y = y|n, p) = \binom{n}{y} p^y (1 - p)^{n-y}, y = 0, 1, 2, \dots, n$$

- $E(X) = np$
- $V(X) = np(1-p)$

Distribuições Discretas: Poisson

- Modelar o tempo (discreto) até a ocorrência de um fenômeno.
- É a distribuição subjacente à modelagem de dados de contagem
- Hipótese: para pequenos intervalos de tempo, a probabilidade de sucesso é proporcional ao tempo até sua ocorrência.
- Como parâmetro da distribuição de Poisson, tem-se λ que mede a intensidade. Assim, a probabilidade de que X assuma o valor x é dado por:

$$P(X = x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

- $E(X) = \lambda$
- $V(X) = \lambda$

Distribuições Discretas: Geométrica

- Probabilidade de que na tentativa x tenhamos o primeiro sucesso (caso específico da distribuição binomial negativa)
- No caso, a probabilidade de um evento ocorrer na x -ésima tentativa é dada por:

$$P(X = x|p) = p(1 - p)^{x-1}$$

- $E(X) = \frac{1}{p}$
- $V(X) = \frac{1-p}{p^2}$

Distribuições Contínuas: Uniforme

- Distribuição de massa uniformemente num segmento da reta $[a, b]$:

$$f(x|a, b) = \frac{1}{b - a}, x \in [a, b]$$

- $E(X) = \frac{b+a}{2}$
- $V(X) = \frac{(b-a)^2}{12}$

Distribuições Contínuas: Normal

- Distribuição com suporte real
- Sua função de densidade é definida por:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2}, -\infty < x < \infty$$

- $E(X) = \mu$
- $V(X) = \sigma^2$

Distribuições Contínuas: T de Student

Distribuição associada a retiradas da distribuição normal, ou seja, está relacionada diretamente ao processo de amostragem. Pelo CLT, dada a validade de algumas condições, ela converge assintoticamente para a Normal.

Retira-se n observações de uma distribuição normal padrão. Então, $\frac{\bar{X} - \mu}{S/\sqrt{n}}$, em que $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ segue uma distribuição t de Student com $n-1$ graus de liberdade.

Distribuições Contínuas: Qui-quadrado

Uma distribuição Y é qui-quadrado com k graus de liberdade se ela foi gerada como a soma de K variáveis aleatórias normais-padrão que são independentes.

Ela tem suporte na reta real positiva e está associada a testes sobre variâncias quando se conhece parâmetros populacionais.

Distribuições Contínuas: F de Snedecor

- A distribuição F de Snedecor surge intuitivamente como um quociente de distribuições qui-quadrados com graus de liberdade não necessariamente idênticos.
- Na prática, ela é usada em testes relacionados à variância e somas de variáveis normalmente distribuídas, na ausência de informações sobre os parâmetros verdadeiros da população.

References

-  Casella, George and Roger L Berger (2021). *Statistical inference*. Cengage Learning.
-  Cunningham, Scott (2021). *Causal inference*. Yale University Press.
-  Wooldridge, Jeffrey M (2015). *Introdução à econometria: uma abordagem moderna*. 6th. Ed. Cengage.

Álgebra Matricial

Prof. Raphael Corbi
Monitor: Alan Leal

FEA/USP

Álgebra Matricial: Definições Básicas

- Uma matriz é um arranjo retangular de números. Ela assume duas dimensões, linhas e colunas. Usualmente, identificamos cada célula pelo seu respectivo número de linha e número da coluna. Assim, por exemplo, uma matriz $m \times n$ é escrita como:

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

- Em que a_{ij} denota o elemento da i-ésima linha e j-ésima coluna da matriz A.
- Uma matriz cujos número de linhas e número de colunas são iguais é denominada matriz quadrada.
- Uma matriz na qual uma das dimensões é igual a 1 é denominada de vetor. Ele pode ser um vetor coluna ($n=1$) ou vetor linha ($m=1$). Uma matriz 1×1 é um escalar.

Operações: soma de matrizes

- Soma de Matrizes: Duas matrizes A e B da mesma dimensão $m \times n$ podem ser somadas elemento a elemento. Isto é:
 $C = A + B \iff C[c_{ij}] = A[a_{ij}] + B[b_{ij}]$
- De uma forma mais genérica, tem-se:

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & a_{m3} + b_{m3} & \dots & a_{mn} + b_{mn} \end{bmatrix}$$

Operações: multiplicação por escalar

- Multiplicação por escalar: Dado um número real γ e uma matriz A, então $\gamma A = A[\gamma a_{ij}]$.
- De uma forma mais genérica, tem-se:

$$\gamma A = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \gamma a_{13} & \dots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \gamma a_{23} & \dots & \gamma a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma a_{m1} & \gamma a_{m2} & \gamma a_{m3} & \dots & \gamma a_{mn} \end{bmatrix}$$

Operações: multiplicação de matrizes

- Duas matrizes A de dimensão mxn e B de dimensão pxq são ditas conformáveis à multiplicação se, e somente se, $n = p$. Neste caso, a multiplicação dessas matrizes é definida como:

$$AB = \left[\sum_{k=1}^n a_{ik} b_{kj} \right] \quad (1)$$

Operações: propriedades das operações de matrizes

- $(\alpha + \beta)A = \alpha A + \beta A$
- $\alpha(A + B) = \alpha A + \alpha B$
- $(\alpha\beta)A = \alpha(\beta A)$
- $\alpha(AB) = (\alpha A)B$
- $A + B = B + A$
- $(A + B) + C = A + (B + C)$
- $(AB)C = A(BC)$
- $A(B + C) = AB + AC$
- $(A + B)C = AC + BC$
- $IA = AI = A$
- $A + 0 = A$
- $A - A = 0$
- $A0 = 0A = 0$
- Não necessariamente $AB = BA$

Transposta de uma matriz

- A transposta de uma matriz A mxn é escrita como A' ou A^T e é definida como a troca de posição de linhas e colunas na matriz A . Assim, por exemplo, a transposta de uma matriz A mxn tem dimensões $n \times m$. Vejamos um exemplo:
- Seja A dada por:

$$A = \begin{bmatrix} 1 & -10 & 9 \\ 0 & 8 & 2/3 \end{bmatrix}$$

- A' nesse caso será dada por:

$$A' = \begin{bmatrix} 1 & 0 \\ -10 & 8 \\ 9 & 2/3 \end{bmatrix}$$

Propriedades da matriz transposta

- $(A')' = A$
- $(\alpha A)' = \alpha A'$
- $(A + B)' = A' + B'$
- $(AB)' = B'A'$, com A e B conformáveis à multiplicação
- $x'x = \sum_{i=1}^n x_i^2$, com x sendo um vetor coluna

Outras definições

- A é dita simétrica se $A=A'$
- Matrizes não-quadradas podem ser simétricas?
- A matriz identidade e matriz de zeros são simétricas?

Multiplicação de matrizes via particionamento

- Sejam A e B duas matrizes conformáveis à multiplicação, então uma outra forma de calcular AB se dá por definir A e B da seguinte forma:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

E

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

Em que A_{11} é $m_1 \times n_1$, A_{12} é $m_1 \times n_2$, A_{21} é $m_2 \times n_1$ e A_{22} é $m_2 \times n_2$.
Analogamente, B_{11} é $m_1 \times p_1$, B_{12} é $n_1 \times p_2$, B_{21} é $n_2 \times p_1$ e B_{22} é $n_2 \times p_2$,
com $m_1 + m_2 = m$, $n_1 + n_2 = n$ e $p_1 + p_2 = p$

O produto AB então será dado por:

$$AB = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}$$

Traço

- O traço de uma matriz A , escrito como $\text{tr}(A)$ é definido como:

$$\text{tr}(A) = \sum_{i=1}^n a_{ii}$$

- O traço tem as seguintes propriedades:

- $\text{tr}(I_n) = n$
- $\text{tr}(A') = \text{tr}(A)$
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(\alpha A) = \alpha \text{tr}(A)$
- $\text{tr}(AB) = \text{tr}(BA)$

Inversa

- A inversa de uma matriz quadrada A é escrita como A^{-1} e é definida como: $A^{-1}A = I_n$ ou $AA^{-1} = I_n$
- Caso A^{-1} exista, diz-se que A é invertível ou não-singular. Um resultado da álgebra linear é que A^{-1} existe se, e somente se, $\det(A) \neq 0$.
- Propriedades:
 - Se A^{-1} existe, então ela é única.
 - $(\alpha A)^{-1} = (1/\alpha)A^{-1}$, com $\alpha \neq 0$ e A invertível
 - $(AB)^{-1} = B^{-1}A^{-1}$, com A e B sendo matrizes quadradas de dimensão n e ambas individualmente invertíveis
 - $(A^{-1})^{-1} = A$
- Os softwares estatísticos automatizam o cálculo da inversa. Contudo, há algoritmos de inversão manual de matrizes, tais como o método de Gauss-Jordan.

Dependência Linear e Posto de uma matriz

- Um conjunto de vetores $nx1 \{x_1, x_2, \dots, x_r\}$ é dito linearmente independente se, e somente se, a solução da seguinte equação:

$$\alpha_1x_1 + \alpha_2x_2 + \cdots + \alpha_rx_r = 0$$

é única e dada por $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$, ou seja, esse sistema admite unicamente a solução trivial. Quando há soluções múltiplas (e nesse caso infinitas), diz-se que os vetores dados são linearmente dependentes.

Dependência Linear e Posto de uma matriz

- Seja A uma matriz $n \times m$. Então, o posto de A é definido como o número máximo de colunas linearmente independentes de A .
- Se $\text{posto}(A) = m$, então diz-se que A tem posto completo.
- Propriedades:
 - $\text{posto}(A) = \text{posto}(A')$
 - Se A é $n \times k$, então $\text{posto}(A) \leq \min\{n, k\}$
 - Se A é quadrada de dimensão k e $\text{posto}(A) = k$, então A é invertível.

Formas quadráticas e matrizes positivas definidas

- Seja A uma matriz quadrada de dimensão n, então para qualquer vetor x de dimensão $nx1$, a forma quadrática associada é dada por:

$$f(x) = x'Ax = \sum_{i=1}^n a_{ii}x_i^2 + 2 \sum_{i=1}^n \sum_{j>i} a_{ij}x_i x_j$$

- Uma matriz simétrica A é dita positiva definida se para qualquer vetor x $nx1$, exceto o vetor nulo, temos que:

$$x'Ax > 0$$

- Uma matriz simétrica A é dita positiva semi-definida se para qualquer vetor x $nx1$, exceto o vetor nulo, temos que:

$$x'Ax \geq 0$$

- Se uma matriz é positiva definida ou positiva semi-definida, assume-se automaticamente que ela também é simétrica.

Propriedades de matrizes positivas definidas e positivas semi-definidas

- Uma matriz positiva definida tem elementos diagonais estritamente positivos, enquanto uma matriz positiva semi-definida tem elementos diagonais não-negativos.
- Se A é positiva definida, então A^{-1} existe e é também positiva definida.
- Se X é $n \times k$, então $X'X$ e XX' são positivas semi-definidas.
- Se X é $n \times k$ e $\text{posto}(X) = k$, então $X'X$ é positiva definida (logo, não-singular).

Matrizes idempotentes

- Uma matriz simétrica de dimensão $n \times n$ é dita idempotente se, e somente se, $AA=A$.
- Exemplo de uma matriz idempotente:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Propriedades:
 - $\text{posto}(A) = \text{tr}(A)$
 - A é positiva semi-definida
- Exemplos de duas matrizes idempotentes importantes na Econometria. Seja X uma matriz $n \times k$ com posto completo, então as matrizes P e M a seguir são idempotentes:
 - $P = X(X'X)^{-1}X'$
 - $M = I_n - X(X'X)^{-1}X' = I_n - P$

Diferenciação de formas quadráticas e lineares

- Há basicamente dois tipos de diferenciações de matrizes que são usadas na derivação do estimador de MQO para o modelo de regressão linear.
- Considere um vetor a de dimensão $nx1$ e defina uma função linear dada por $f(x) = a'x$, com x sendo um vetor coluna de tamanho n . Então, a derivada de f em relação a x é um vetor $1xn$ de derivadas parciais dadas por:

$$\frac{\partial f(x)}{\partial x} = a'$$

- Para uma matriz simétrica A de tamanho nxn , defina a forma quadrática como $g(x) = x'Ax$, então:

$$\frac{\partial g(x)}{\partial x} = 2x'A$$

Que é um vetor $1xn$.

Momentos de vetores aleatórios

- Um vetor aleatório é um vetor cujos elementos são variáveis aleatórias.
- Se y é um vetor $nx1$, o valor esperado de y , denotado por $E(y)$ é o vetor de valores esperados, isto é, $E(y) = [E(y_1), E(y_2), \dots, E(y_n)]'$
- Se Z é uma matriz nxm aleatória, então $E(Z)$ é uma matriz nxm de valores esperados: $E(Z) = [E(z_{ij})]$
- Se y é um vetor aleatório de tamanho $nx1$, sua matriz de variância-covariância é dada por:

$$Var(y) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{11} & \sigma_{m2} & \dots & \sigma_n \end{bmatrix}$$

Com $\sigma_i^2 = Var(y_i)$ e $\sigma_{ij} = Cov(y_i, y_j)$

Momentos de vetores aleatórios

- Propriedades:

- Se a é um vetor de dimensão $nx1$, então $Var(a'y) = a'[Var(y)]a \geq 0$
- Se $Var(a'y) > 0$ para todo $a \neq 0$, então é positiva definida.
- $Var(y) = E[(y - \mu)(y - \mu)']$, com $\mu = E(y)$
- Se todos os elementos de y são não-correlacionados, então $Var(y)$ é uma matriz diagonal.

References

-  Cunningham, Scott (2021). *Causal inference*. Yale University Press.
-  Davidson, Russell, James G MacKinnon, et al. (2004). *Econometric theory and methods*. Vol. 5. Oxford University Press New York.
-  Wooldridge, Jeffrey M (2015). *Introdução à econometria: uma abordagem moderna*. 6th. Ed. Cengage.

Ordinary Least Squares: Part I

Raphael Corbi

Universidade de São Paulo

April 2021

Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

Terminology

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor
LHS	RHS

The terms “explained” and “explanatory” are probably best, as they are the most descriptive and widely applicable. But “dependent” and “independent” are used often. (The “independence” here is not really statistical independence.)

We said we must confront three issues:

- ① How do we allow factors other than x to affect y ?
- ② What is the functional relationship between y and x ?
- ③ How can we be sure we are capturing a ceteris paribus relationship between y and x ?

We will argue that the simple regression model

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

addresses each of them.

Simple linear regression model

- The simple linear regression (SLR) model is a population model.
- When it comes to estimating β_1 (and β_0) using a random sample of data, we must restrict how u and x are related to each other.
- What we must do is restrict the way u and x relate to each other in the population.

The error term

We make a simplifying assumption (without loss of generality): the average, or expected, value of u is zero in the population:

$$E(u) = 0 \tag{2}$$

where $E(\cdot)$ is the expected value operator.

The intercept

The presence of β_0 in

$$y = \beta_0 + \beta_1 x + u \quad (3)$$

allows us to assume $E(u) = 0$. If the average of u is different from zero, say α_0 , we just adjust the intercept, leaving the slope the same:

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0) \quad (4)$$

where $\alpha_0 = E(u)$. The new error is $u - \alpha_0$ and the new intercept is $\beta_0 + \alpha_0$. The important point is that the slope, β_1 , has not changed.

Mean independence of the error term

An assumption that meshes well with our introductory treatment involves the mean of the error term for each “slice” of the population determined by values of x :

$$E(u|x) = E(u), \text{ all values } x \quad (5)$$

where $E(u|x)$ means “the expected value of u given x ”.
Then, we say u is **mean independent** of x .

Distribution of ability across education

- Suppose u is “ability” and x is years of education. We need, for example,

$$E(\text{ability}|x = 8) = E(\text{ability}|x = 12) = E(\text{ability}|x = 16)$$

so that the average ability is the same in the different portions of the population with an 8th grade education, a 12th grade education, and a four-year college education.

- Because people choose education levels partly based on ability, this assumption is almost certainly false.

Zero conditional mean assumption

Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives the **zero conditional mean assumption**.

$$E(u|x) = 0, \text{ all values } x \quad (6)$$

Population regression function

Because the conditional expected value is a linear operator,
 $E(u|x) = 0$ implies

$$E(y|x) = \beta_0 + \beta_1 x \quad (7)$$

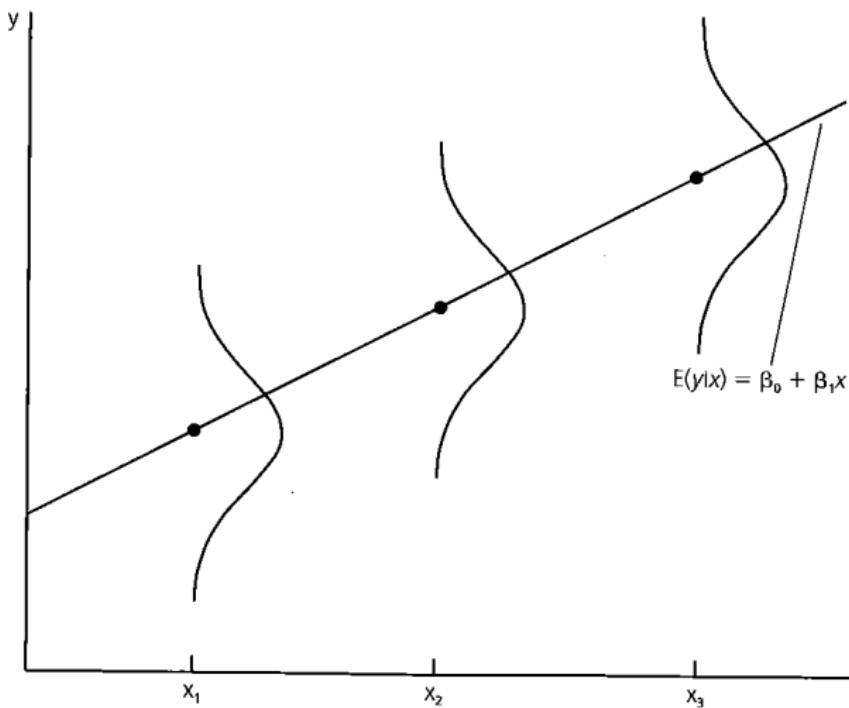
which shows the **population regression function** is a linear function of x .

- The straight line in the graph on the next page is what Wooldridge calls the **population regression function**, and what Angrist and Pischke call the **conditional expectation function**

$$E(y|x) = \beta_0 + \beta_1 x$$

- The conditional distribution of y at three different values of x are superimposed. for a given value of x , we see a range of y values: remember, $y = \beta_0 + \beta_1 x + u$, and u has a distribution in the population.

$E(y|x)$ as a linear function of x .



Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

Deriving the Ordinary Least Squares Estimates

- Given data on x and y , how can we estimate the population parameters, β_0 and β_1 ?
- Let $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ be a **random** sample of size n (the number of observations) from the population.
- Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (8)$$

where the i subscript indicates a particular observation.

- We observe y_i and x_i , but not u_i (but we know it is there).

We use the two population restrictions:

$$E(u) = 0$$

$$\text{Cov}(x, u) = 0$$

to obtain estimating equations for β_0 and β_1 . We talked about the first condition. The second condition means that x and u are uncorrelated. Both conditions are implied by $E(u|x) = 0$

With $E(u) = 0$, $\text{Cov}(x, u) = 0$ is the same as $E(xu) = 0$. Next we plug in for u :

$$\begin{aligned}E(y - \beta_0 - \beta_1 x) &= 0 \\E[x(y - \beta_0 - \beta_1 x)] &= 0\end{aligned}$$

These are the two conditions in the **population** that effectively determine β_0 and β_1 .

So we use their sample counterparts (which is a method of moments approach to estimation):

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$n^{-1} \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the data.

These are two linear equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$.

Pass the summation operator through the first equation:

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \quad (9)$$

$$= n^{-1} \sum_{i=1}^n y_i - n^{-1} \sum_{i=1}^n \hat{\beta}_0 - n^{-1} \sum_{i=1}^n \hat{\beta}_1 x_i \quad (10)$$

$$= n^{-1} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 \left(n^{-1} \sum_{i=1}^n x_i \right) \quad (11)$$

$$= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} \quad (12)$$

We use the standard notation $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ for the average of the n numbers $\{y_i : i = 1, 2, \dots, n\}$. For emphasis, we call \bar{y} a **sample average**.

We have shown that the first equation,

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (13)$$

implies

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (14)$$

Now, use this equation to write the intercept in terms of the slope:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (15)$$

Plug this into the second equation (but where we take away the division by n):

$$\sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (16)$$

so

$$\sum_{i=1}^n x_i[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0 \quad (17)$$

Simple algebra gives

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \hat{\beta}_1 \left[\sum_{i=1}^n x_i(x_i - \bar{x}) \right] \quad (18)$$

So, the equation to solve is

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \quad (19)$$

If $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$, we can write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)} \quad (20)$$

OLS

- The previous formula for $\hat{\beta}_1$ is important. It shows us how to take the data we have and compute the slope estimate.
- $\hat{\beta}_1$ is called the **ordinary least squares (OLS)** slope estimate.
- It can be computed whenever the sample variance of the x_i is not zero, which only rules out the case where each x_i has the same value.
- The intuition is that the variation in x is what permits us to identify its impact on y .

Solving for $\hat{\beta}$

- Once we have $\hat{\beta}_1$, we compute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. This is the OLS intercept estimate.
- These days, we let the computer do the calculations, which are tedious even if n is small.

Predicting y

- For any candidates $\hat{\beta}_0$ and $\hat{\beta}_1$, define a **fitted value** for each i as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (21)$$

We have n of these.

- \hat{y}_i is the value we predict for y_i given that $x = x_i$ and $\beta = \hat{\beta}$.

The residual

- The “mistake” from our *prediction* is called the **residual**:

$$\begin{aligned}\hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\end{aligned}$$

- Suppose we measure the size of the mistake, for each i , by squaring it. Then we add them all up to get the **sum of squared residuals**

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to *minimize* the sum of squared residuals which gives us the same solutions we obtained before.

Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

Algebraic Properties of OLS Statistics

Remembering how the **first moment** condition allows us to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, we have:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (22)$$

Notice the logic here: this means the OLS residuals *always* add up to zero, by *construction*,

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (23)$$

Because $y_i = \hat{y}_i + \hat{u}_i$ by definition,

$$n^{-1} \sum_{i=1}^n y_i = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^n \hat{u}_i \quad (24)$$

and so $\bar{y} = \bar{\hat{y}}$.

Second moment

Similarly the way we obtained our estimates,

$$n^{-1} \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (25)$$

The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero:

$$n^{-1} \sum_{i=1}^n x_i \hat{u}_i = 0 \quad (26)$$

Bringing things together

Because the \hat{y}_i are linear functions of the x_i , the fitted values and residuals are uncorrelated, too:

$$n^{-1} \sum_{i=1}^n \hat{y}_i \hat{u}_i = 0 \quad (27)$$

Averages

A third property is that the point (\bar{x}, \bar{y}) is always on the OLS regression line. That is, if we plug in the average for x , we predict the sample average for y :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{28}$$

Again, we chose the estimates to make this true.

Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

Expected Value of OLS

- Mathematical statistics: How do our estimators behave across different samples of data? On average, would we get the right answer if we could repeatedly sample?
- We need to find the expected value of the OLS estimators – in effect, the average outcome across all possible random samples – and determine if we are right on average.
- Leads to the notion of **unbiasedness**, which is a “desirable” characteristic for estimators.

$$E(\hat{\beta}) = \beta \tag{29}$$

Don't forget why we're here

- Plato's allegory of the cave - reality is outside the cave, the reflections on the wall are our estimates of that reality.
- The **population** parameter that describes the relationship between y and x is β_1
- For this class, β_1 is a causal parameter, and our sole objective is to estimate β_1 with a sample of data
- But never forget that $\hat{\beta}_1$ is an **estimator** of that causal parameter obtained with a *specific* sample from the population.

Uncertainty and sampling variance

- Different samples will generate different estimates ($\hat{\beta}_1$) for the “true” β_1 which makes $\hat{\beta}_1$ a random variable.
- Unbiasedness is the idea that if we could take as many random samples on Y as we want from the population, and compute an estimate each time, the average of these estimates would be equal to β_1 .
- But, this also implies that $\hat{\beta}_1$ has spread and therefore variance

Assumptions

Assumption SLR.1 (Linear in Parameters)

- The population model can be written as

$$y = \beta_0 + \beta_1 x + u \quad (30)$$

where β_0 and β_1 are the (unknown) population parameters.

- We view x and u as outcomes of random variables; thus, y is random.
- Stating this assumption formally shows that our goal is to estimate β_0 and β_1 .

Assumption SLR.2 (Random Sampling)

- We have a random sample of size n , $\{(x_i, y_i) : i = 1, \dots, n\}$, following the population model.
- We know how to use this data to estimate β_0 and β_1 by OLS.
- Because each i is a draw from the population, we can write, for each i ,

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (31)$$

- Notice that u_i here is the unobserved error for observation i . It is not the residual that we compute from the data!

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

- The sample outcomes on x_i are not all the same value.
- This is the same as saying the sample variance of $\{x_i : i = 1, \dots, n\}$ is not zero.
- In practice, this is no assumption at all. If the x_i are all the same value, we cannot learn how x affects y in the population.

Assumption SLR.4 (Zero Conditional Mean)

- In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = E(u) = 0. \quad (32)$$

- This is the key assumption for showing that OLS is unbiased, with the zero value not being important once we assume $E(u|x)$ does not change with x .
- Note that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model.

Showing OLS is unbiased

How do we show $\hat{\beta}_1$ is unbiased for β_1 ? What we need to show is

$$E(\hat{\beta}_1) = \beta_1 \tag{33}$$

where the expected value means averaging across random samples.

Step 1: Write down a formula for $\hat{\beta}_1$. It is convenient to use

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{34}$$

which is one of several equivalent forms.

It is convenient to define $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2$, to total variation in the x_i , and write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{SST_x} \quad (35)$$

Remember, SST_x is just some positive number. The existence of $\hat{\beta}_1$ is guaranteed by SLR.3.

Step 2: Replace each y_i with $y_i = \beta_0 + \beta_1 x_i + u_i$ (which uses SLR.1 and the fact that we have data from SLR.2).

The numerator becomes

$$\sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) \quad (36)$$

$$= \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i + \sum_{i=1}^n (x_i - \bar{x})u_i \quad (37)$$

$$= 0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})u_i \quad (38)$$

$$= \beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x})u_i \quad (39)$$

We used $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $\sum_{i=1}^n (x_i - \bar{x})x_i = \sum_{i=1}^n (x_i - \bar{x})^2$.

We have shown

$$\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x} \quad (40)$$

Note how the last piece is the slope coefficient from the OLS regression of u_i on x_i , $i = 1, \dots, n$. We cannot do this regression because the u_i are not observed.

Now define

$$w_i = \frac{(x_i - \bar{x})}{SST_x} \quad (41)$$

so we have

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i \quad (42)$$

- $\hat{\beta}_1$ is a linear function of the unobserved errors, u_i . The w_i are all functions of $\{x_1, x_2, \dots, x_n\}$.
- The (random) difference between $\hat{\beta}_1$ and β_1 is due to this linear function of the unobservables.

Step 3: Find $E(\hat{\beta}_1)$.

- Under Assumptions SLR.2 and SLR.4, $E(u_i|x_1, x_2, \dots, x_n) = 0$.
That means, *conditional* on $\{x_1, x_2, \dots, x_n\}$,

$$E(w_i u_i | x_1, x_2, \dots, x_n) = w_i E(u_i | x_1, x_2, \dots, x_n) = 0$$

because w_i is a function of $\{x_1, x_2, \dots, x_n\}$. (In the next slides I omit the conditioning in the expectations)

- This would not be true if, in the population, u and x are correlated.

Now we can complete the proof: conditional on $\{x_1, x_2, \dots, x_n\}$,

$$E(\hat{\beta}_1) = E\left(\beta_1 + \sum_{i=1}^n w_i u_i\right) \quad (43)$$

$$= \beta_1 + \sum_{i=1}^n E(w_i u_i) = \beta_1 + \sum_{i=1}^n w_i E(u_i) \quad (44)$$

$$= \beta_1 \quad (45)$$

Remember, β_1 is the fixed constant in the population. The estimator, $\hat{\beta}_1$, varies across samples and is the random outcome: before we collect our data, we do not know what $\hat{\beta}_1$ will be.

THEOREM (Unbiasedness of OLS)

Under Assumptions SLR.1 through SLR.4

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1. \quad (46)$$

- Omit the proof for $\hat{\beta}_0$.

- Each sample leads to a different estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$. Some will be very close to the true values $\beta_0 = 3$ and $\beta_1 = 2$. Nevertheless, some could be very far from those values.
- If we repeat the experiment again and again, and average the estimates, we would get very close to 2.
- The problem is, we do not know which kind of sample we have. We can never know whether we are close to the population value.
- We hope that our sample is "typical" and produces a slope estimate close to β_1 but we can never know.

Omitted Variable Bias

- A typical problem is when a key variable is omitted. Assume schooling causes earnings to rise:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i$$

Y_i = log of earnings

S_i = schooling measured in years

A_i = individual ability

- Typically the econometrician cannot observe A_i ; for instance, the Current Population Survey doesn't present adult respondents' family background, intelligence, or motivation.

Shorter regression

- What are the consequences of leaving ability out of the regression? Suppose you estimated this shorter regression instead:

$$Y_i = \beta_0 + \beta_1 S_i + \eta_i$$

where $\eta_i = \beta_2 A_i + u_i$; β_0 , β_1 , and β_2 are population regression coefficients; S_i is correlated with η_i through A_i only; and u_i is a regression residual uncorrelated with all regressors by definition.

Derivation of Ability Bias

- Suppressing the i subscripts, the OLS estimator for β_1 is:

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, S)}{\text{Var}(S)} = \frac{E[YS] - E[Y]E[S]}{\text{Var}(S)}$$

- Plugging in the true model for Y , we get:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{Cov}[(\beta_0 + \beta_1 S + \beta_2 A + u), S]}{\text{Var}(S)} \\ &= \frac{E[(\beta_0 S + \beta_1 S^2 + \beta_2 SA + uS)] - E(S)E[\beta_0 + \beta_1 S + \beta_2 A + u]}{\text{Var}(S)} \\ &= \frac{\beta_1 E(S^2) - \beta_1 E(S)^2 + \beta_2 E(SA) - \beta_2 E(S)E(A) + E(uS) - E(S)E(u)}{\text{Var}(S)} \\ &= \beta_1 + \beta_2 \frac{\text{Cov}(A, S)}{\text{Var}(S)}\end{aligned}$$

- If $\beta_2 > 0$ and $\text{Cov}(A, S) > 0$ the coefficient on schooling in the shortened regression (without controlling for A) would be upward biased

Summary

- When $\text{Cov}(A, S) > 0$ then ability and schooling are correlated.
- When ability is unobserved, then not even multiple regression will identify the causal effect of schooling on wages.
- Here we see one of the main justifications for this workshop – what will we do when the treatment variable is endogenous?
- We will need an *identification strategy* to recover the causal effect

Ordinary Least Squares: Part II

Raphael Corbi

Universidade de São Paulo

April 2021

Reminder

- **Errors** are the vertical distances between observations and the **unknown** Conditional Expectation Function. Therefore, they are unknown.
- **Residuals** are the vertical distances between observations and the **estimated** regression function. Therefore, they are known.

SE and the data

The correct SE estimation procedure is given by the underlying structure of the data

- It is very unlikely that all observations in a dataset are unrelated, but drawn from identical distributions (**homoskedasticity**)
- For instance, the variance of income is often greater in families belonging to top deciles than among poorer families (**heteroskedasticity**)
- Some phenomena do not affect observations individually, but they do affect groups of observations uniformly within each group (**clustered data**)

Variance of the OLS Estimators

- Under SLR.1 to SLR.4, the OLS estimators are unbiased. This tells us that, on average, the estimates will equal the population values.
- But we need a measure of dispersion (spread) in the sampling distribution of the estimators. We use the variance (and, ultimately, the standard deviation).
- We could characterize the variance of the OLS estimators under SLR.1 to SLR.4 (and we will later). For now, it is easiest to introduce an assumption that simplifies the calculations.

Assumption SLR.5 (Homoskedasticity, or Constant Variance)

The error has the same variance given any value of the explanatory variable x :

$$\text{Var}(u|x) = \sigma^2 > 0 \quad (47)$$

where σ^2 is (virtually always) unknown.

Because we assume SLR.4, that is, $E(u|x) = 0$ whenever we assume SLR.5, we can also write

$$E(u^2|x) = \sigma^2 = E(u^2) \quad (48)$$

Under the population Assumptions SLR.1 ($y = \beta_0 + \beta_1x + u$),
SRL.4 ($E(u|x) = 0$) and SLR.5 ($Var(u|x) = \sigma^2$),

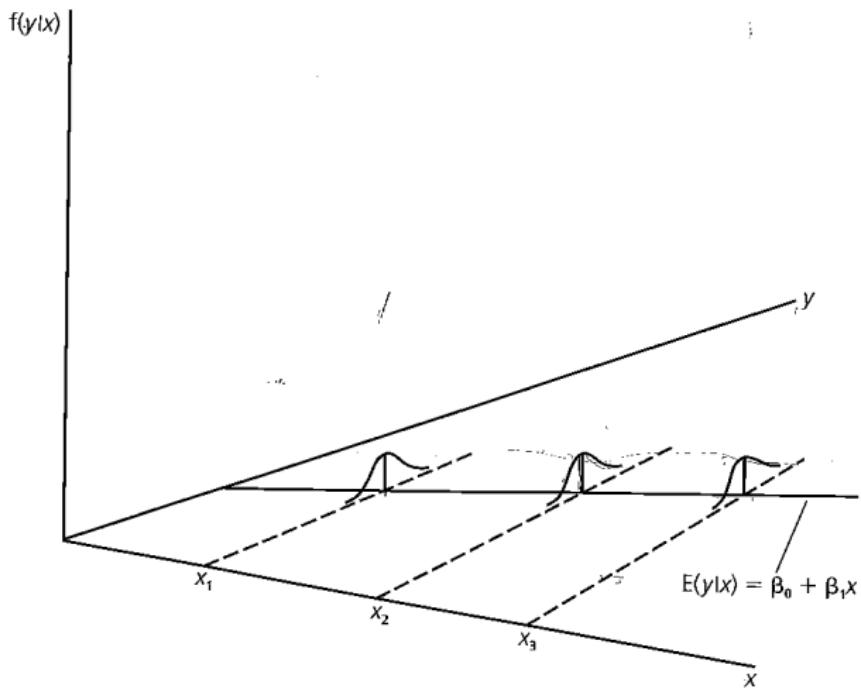
$$E(y|x) = \beta_0 + \beta_1x$$

$$Var(y|x) = \sigma^2$$

So the average or expected value of y is allowed to change with x –
in fact, this is what interests us – but the variance does not change
with x . (See Graphs on next two slides)

Figure 2.8

The simple regression model under homoskedasticity.



THEOREM (Sampling Variances of OLS)

Under Assumptions SLR.1 to SLR.2,

$$Var(\hat{\beta}_1|x) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\hat{\beta}_0|x) = \frac{\sigma^2 (n^{-1} \sum_{i=1}^n x_i^2)}{SST_x}$$

(conditional on the outcomes $\{x_1, x_2, \dots, x_n\}$).

To show this, write, as before,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i \quad (49)$$

where $w_i = (x_i - \bar{x})/SST_x$. We are treating this as nonrandom in the derivation. Because β_1 is a constant, it does not affect $Var(\hat{\beta}_1)$. Now, we need to use the fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances.

The $\{u_i : i = 1, 2, \dots, n\}$ are actually independent across i , and so they are uncorrelated. So (remember that if we know x , we know w)

$$\begin{aligned}Var(\hat{\beta}_1|x) &= Var\left(\sum_{i=1}^n w_i u_i | x\right) \\&= \sum_{i=1}^n Var(w_i u_i | x) = \sum_{i=1}^n w_i^2 Var(u_i | x) \\&= \sum_{i=1}^n w_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n w_i^2\end{aligned}$$

where the second-to-last equality uses Assumption SLR.5, so that the variance of u_i does not depend on x_i .

Now we have

$$\begin{aligned}\sum_{i=1}^n w_i^2 &= \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(SST_x)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(SST_x)^2} \\ &= \frac{SST_x}{(SST_x)^2} = \frac{1}{SST_x}\end{aligned}$$

We have shown

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \tag{50}$$

Usually we are interested in β_1 . We can easily study the two factors that affect its variance.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (51)$$

- ① As the error variance increases, i.e., as σ^2 increases, so does $Var(\hat{\beta}_1)$. The more “noise” in the relationship between y and x – that is, the larger variability in u – the harder it is to learn about β_1 .
- ② By contrast, more variation in $\{x_i\}$ is a *good* thing:

$$SST_x \uparrow \text{ implies } Var(\hat{\beta}_1) \downarrow \quad (52)$$

Notice that SST_x/n is the sample variance in x . We can think of this as getting close to the population variance of x , σ_x^2 , as n gets large. This means

$$SST_x \approx n\sigma_x^2 \tag{53}$$

which means, as n grows, $Var(\hat{\beta}_1)$ shrinks at the rate $1/n$. This is why more data is a good thing: it shrinks the sampling variance of our estimators.

The standard deviation of $\hat{\beta}_1$ is the square root of the variance. So

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}} \quad (54)$$

This turns out to be the measure of variation that appears in confidence intervals and test statistics.

Estimating the Error Variance

In the formula

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (55)$$

we can compute SST_x from $\{x_i : i = 1, \dots, n\}$. But we need to estimate σ^2 .

Recall that

$$\sigma^2 = E(u^2). \quad (56)$$

Therefore, if we could observe a sample on the errors,
 $\{u_i : i = 1, 2, \dots, n\}$, an unbiased estimator of σ^2 would be the sample average

$$n^{-1} \sum_{i=1}^n u_i^2 \quad (57)$$

But this is not an estimator because we cannot compute it from the data we observe, since u_i are unobserved.

How about replacing each u_i with its “estimate”, the OLS residual \hat{u}_i ?

$$\begin{aligned} u_i &= y_i - \beta_0 - \beta_1 x_i \\ \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \end{aligned}$$

\hat{u}_i can be computed from the data because it depends on the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Except by fluke,

$$\hat{u}_i \neq u_i \quad (58)$$

for any i .

$$\begin{aligned}\hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i\end{aligned}$$

$E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$, but the estimators almost always differ from the population values in a sample.

Now, what about this as an estimator of σ^2 ?

$$n^{-1} \sum_{i=1}^n \hat{u}_i^2 = SSR/n \quad (59)$$

It is a true estimator and easily computed from the data after OLS. As it turns out, this estimator is slightly biased: its expected value is a little less than σ^2 .

The estimator does not account for the two restrictions on the residuals, used to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\sum_{i=1}^n \hat{u}_i = 0$$

$$\sum_{i=1}^n x_i \hat{u}_i = 0$$

There is no such restriction on the unobserved errors.

The unbiased estimator of σ^2 uses a **degrees-of-freedom** adjustment. The residuals have only $n - 2$ degrees-of-freedom, not n .

$$\hat{\sigma}^2 = \frac{SSR}{(n - 2)} \quad (60)$$

THEOREM: Unbiased Estimator of σ^2
Under Assumptions SLR.1 to SLR.5,

$$E(\hat{\sigma}^2) = \sigma^2 \quad (61)$$

In regression output, it is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSR}{(n - 2)}} \quad (62)$$

that is usually reported. This is an estimator of $sd(u)$, the standard deviation of the population error. And $SSR = \sum_{i=1}^n \hat{u}^2$.

- $\hat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression. Stata calls it the **root mean squared error**.
- Given $\hat{\sigma}$, we can now estimate $sd(\hat{\beta}_1)$ and $sd(\hat{\beta}_0)$. The estimates of these are called the **standard errors** of the $\hat{\beta}_j$.

- We just plug $\hat{\sigma}$ in for σ :

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} \quad (63)$$

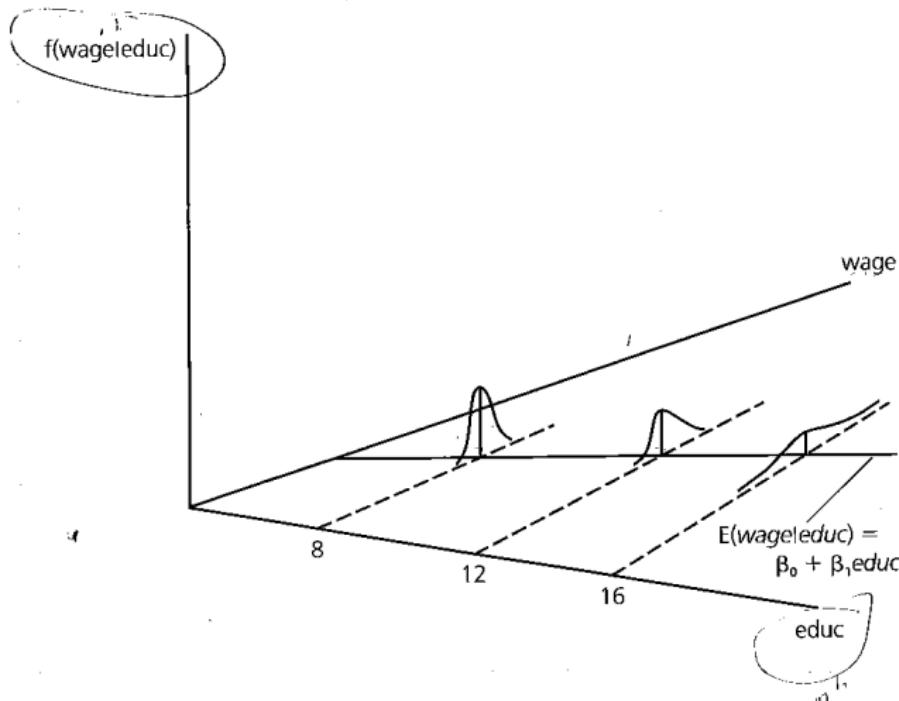
where both the numerator and denominator are computed from the data.

- For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

- OLS inference is generally faulty in the presence of heteroskedasticity

Figure 2.9

Var (wage|educ) increasing with educ.



- Fortunately, OLS is still useful
- Assume SLR.1-4 hold, but not SLR.5. Therefore

$$Var(u_i|x_i) = \sigma_i^2$$

- The variance of our estimator, $\hat{\beta}_1$ equals:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

- When $\sigma_i^2 = \sigma^2$ for all i , this formula reduces to the usual form,

$$\frac{\sigma^2}{SST_x^2}$$

- A valid estimator of $\text{Var}(\hat{\beta}_1)$ for heteroskedasticity of any form (including homoskedasticity) is

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$$

which is easily computed from the data after the OLS regression

- As a rule, you should always use the , robust command in STATA.

Clustered data

- But what if errors are not iid?
- For instance, maybe observations between units in a group are related to each other
 - You want to regress kids' grades on class size to determine the effect of class size on grades
 - The **unobservables** of kids belonging to the same classroom will be correlated (e.g., teacher quality, recess routines) while will not be correlated with kids in far away classrooms
- Then i.i.d. is violated. But maybe i.i.d. holds across clusters, just not within clusters

Simulations

- Let's first try to understand what's going on with a few simulations
- We will begin with a baseline of non-clustered data
- We'll show the distribution of estimates in Monte Carlo simulation for 1000 draws and iid errors
- We'll then show the number of times you reject the null incorrectly at $\alpha = 0.05$.

Least squares estimates of non-clustered data

Monte Carlo simulation of the slope

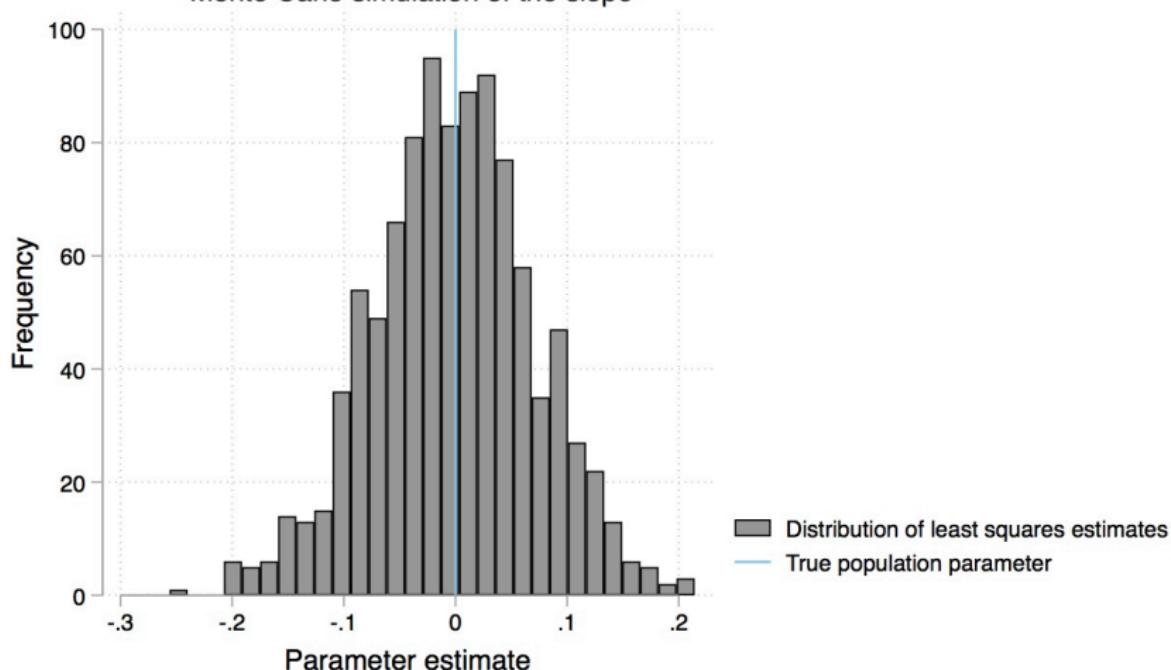


Figure: Distribution of the least squares estimator over 1,000 random draws.

Clustered data and heteroskedastic robust

- Now let's look at clustered data
- But this time we will estimate the model using heteroskedastic robust standard errors
- Earlier we saw mass all the way to -2.5 to 2; what do we get when we incorrectly estimate the standard errors?

Least squares estimates of clustered Data Monte Carlo simulation of the slope

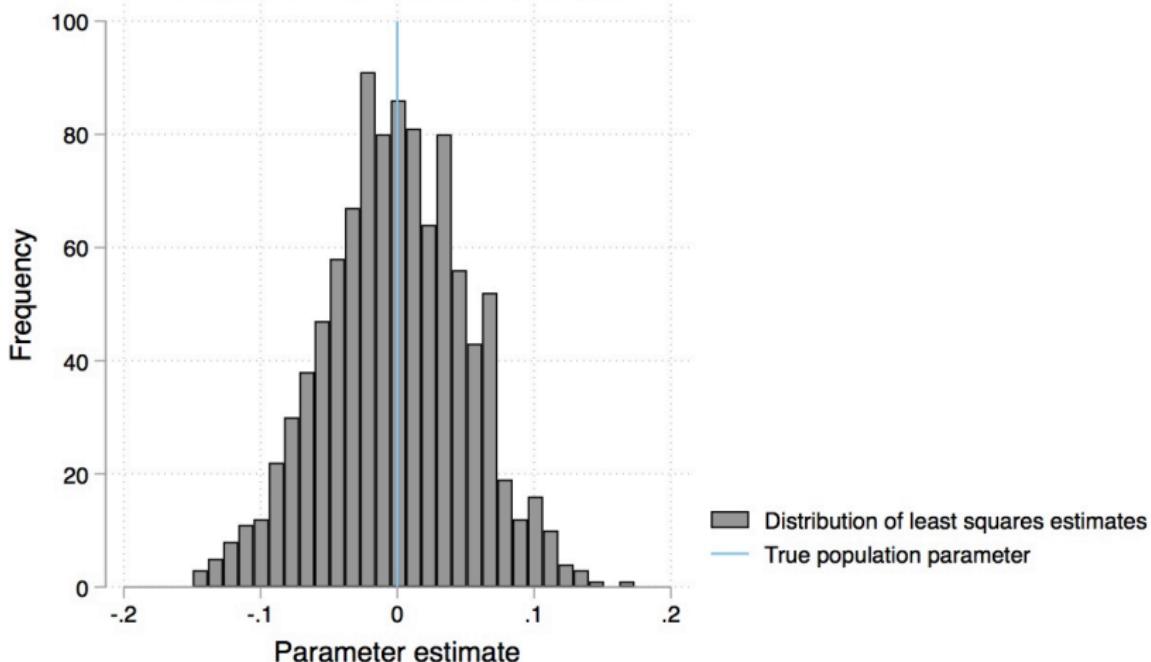


Figure: Distribution of the least squares estimator over 1,000 random draws. Clustered data without correcting for clustering

Over-rejecting the null

- Those 95 percent confidence intervals are based on an $\alpha = 0.05$.
- Look how many parameter estimates are different from zero; that's what we mean by "over-rejecting the null"
- You saw signs of it though in the variance of the estimated effect, bc the spread only went from -.15 to .15 (whereas earlier it had gone from -.25 to .2)
- Now let's correct for arbitrary within group correlations using the cluster robust option in Stata/R

Cluster robust standard errors

- Better. We don't have the same over-rejection problem as before. If anything it's more conservative.
- The formula for estimating standard errors changes when allowing for arbitrary serial correlation within group.
- Instead of summing over each individual, we first sum over groups
- I'll use matrix notation as it's easier for me to explain by stacking the data.

Variance-Covariance Matrix

Homokedasticity

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

Heterokedasticity

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

In matrix language:

$$\text{Var}(\hat{\beta}_1) = (X'X)^{-1} X' \sum X(X'X)^{-1}$$

where \sum is the variance-covariance matrix

Variance-Covariance Matrix

$$\Sigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 & \sigma_{02}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{20}^2 & \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$

Variance-Covariance Matrix

In other words, the variance-covariance matrix changes depending on the assumptions made about your error u :

1. classic homoskedastic standard errors assume that Σ is diagonal with identical elements σ^2 , which simplifies the expression for $\text{Var}(\hat{\beta}_1) = \sigma^2(X'X)^{-1}$
2. Huber-White standard errors assume that Σ is diagonal, but that the diagonal value varies σ_i ;
3. Clustered standard errors assume that Σ is block-diagonal according to the clusters in the sample, with unrestricted values in each block but zeros elsewhere.

The importance of knowing your data

- In real world you should never go with the “independent and identically distributed” (i.e., homoskedasticity) case. Life is not that simple.
- You need to know your data in order to choose the correct error structure and then infer the required SE calculation
- If you have aggregate variables, like class size, clustering at that level is *required*

Clustered data

- Let's stack the observations by cluster

$$y_g = x_g\beta + u_g$$

- The OLS estimator of β is:

$$\widehat{\beta} = [X'X]^{-1}X'y$$

- The variance is given by:

$$Var(\beta) = E[[X'X]^{-1}X'\Omega X[X'X]^{-1}]$$

Clustered data

With this in mind, we can now write the variance-covariance matrix for clustered data

$$Var(\hat{\beta}) = [X'X]^{-1} \left[\sum_{i=1}^G x_g' \hat{u}_g \hat{u}_g' x_g \right] [X'X]^{-1}$$

where \hat{u}_g are residuals from the stacked regression

- In STATA: `vce(cluster clustervar)`. Where `clustervar` is a variable that identifies the groups in which unobservables are allowed to correlate

Robust least squares estimates of clustered data 95% Confidence interval of the slope

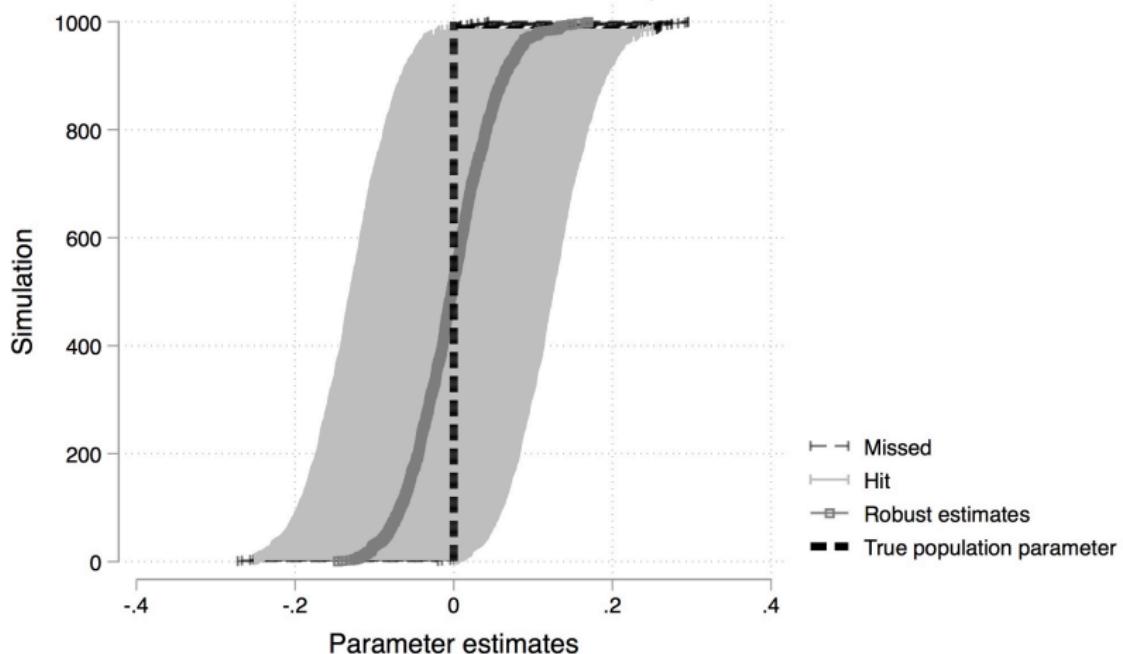


Figure: Distribution of 1,000 95% confidence intervals from a cluster robust least squares regression with dashed region representing those estimates that incorrectly reject the null.

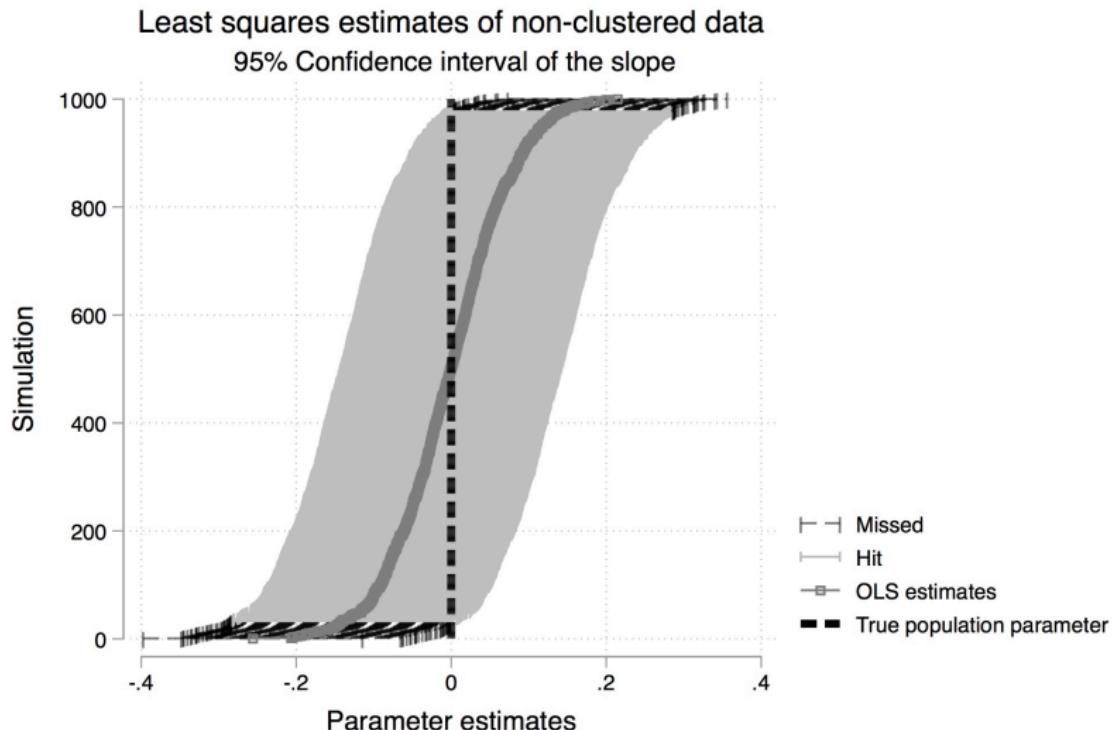


Figure: Distribution of the 95% confidence intervals with coloring showing those which are incorrectly rejecting the null.

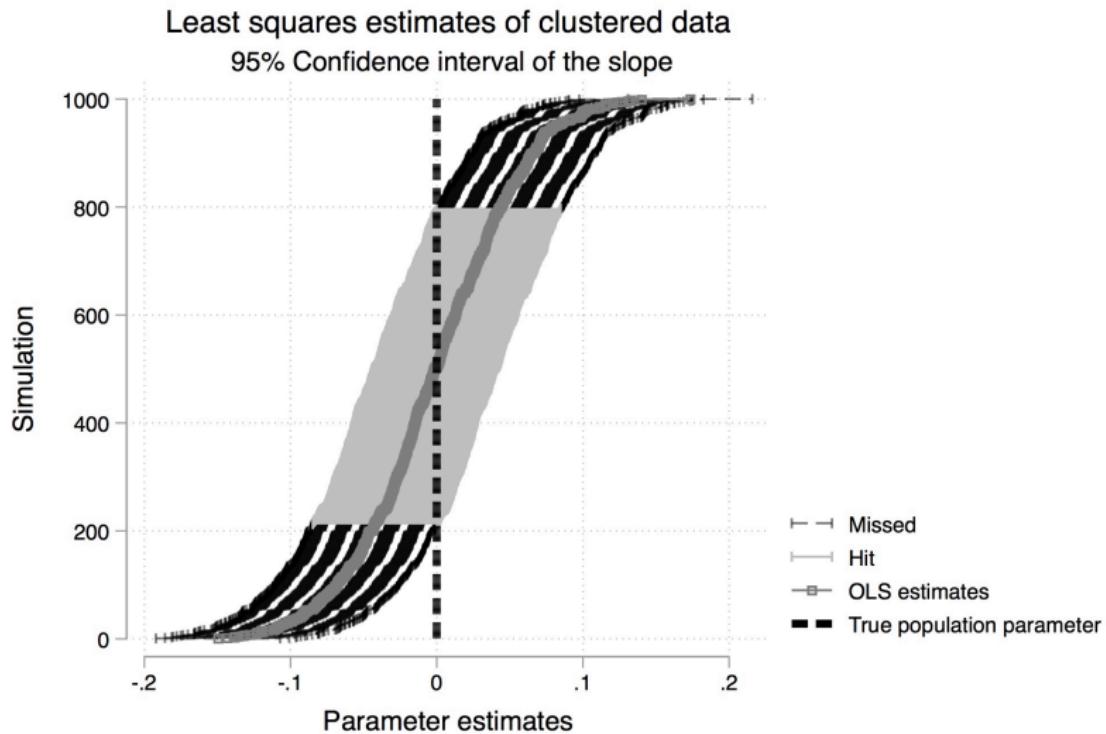


Figure: Distribution of 1,000 95% confidence intervals with dashed region representing those estimates that incorrectly reject the null.

Regressão e Causalidade

Raphael Corbi

Universidade de São Paulo

Agosto 2021

Quatro Perguntas Básicas do Projeto de Pesquisa

1. Qual a relação causal de interesse?

- ▶ útil para prever as consequências de políticas públicas
- ▶ conceito de *contrafactual*
- ▶ *relação de interesse teórico (parâmetro do modelo)*
- ▶ ex: *impacto de educação em salário (mudança no custo da universidade)*
- ▶ *unidade de análise: indivíduo, firmas, cidades, países*
- ▶ ex: *instituições democráticas e crescimento econômico (Acemoglu, Johnson, Robinson, 2001)*

Quatro Perguntas Básicas do Projeto de Pesquisa

2. Qual o experimento ideal para capturar a relação causal?

- ▶ experimentos ideais são geralmente hipotéticos
- ▶ zero restrição orçamentária e ética
- ▶ escolha da tópicos promissores e na formulação precisa da pergunta
- ▶ ex: alocação aleatória de instituições no dia da independência
- ▶ algumas perguntas não possuem experimento ideal
- ▶ ex: efeito de idade ao início da escola sobre aprendizado *vs* maturidade

Quatro Perguntas Básicas do Projeto de Pesquisa

3. Qual o estratégia de identificação?

- ▶ experimentos ideais quase nunca existem
- ▶ dados observados (aleatórios) aproximam experimento natural
- ▶ parte da informação observada é aleatoria
- ▶ ex: mês de nascimento e lei de escolaridade compulsória (Angrist Krueger, 2001)
- ▶ **first-best**: experimento ideal (alocação aleatória)
- ▶ **second-best**: experimento natural
 - ▶ IV, diff-in-diff, RDD, propensity score matching
 - ▶ validação interna: *instrumento afeta somente tratamento?*
 - ▶ *resolvem o problema de seleção*

Quatro Perguntas Básicas do Projeto de Pesquisa

4. Qual o modo da inferência estatística?

- ▶ depende da população estudada, amostragem
- ▶ hipóteses na construção do erros-padrão
- ▶ ex: microdados do censo vs dados agrupados
 - ▶ clustering (classroom), serial-correlation (DID)
 - ▶ nickel bias (dynamic pane with finite T)

*"T-stat looks too good.
Use robust standard errors
significance gone."*

— Keisuke Hirano

Problema de Seleção

- ▶ Pergunta: **Pronto-Socorro torna pacientes mais saudáveis?**
- ▶ hospitais oferecem tratamento, mas também contato com doentes
- ▶ dados da National Health Interview Survey
- ▶ 2 perguntas: foi ao pronto-socorro? condição de saúde 1 a 5?

	Sample size	Mean Health Status	Std Error
No Hospital	90,049	3.93	0,003
Hospital	7,774	3.21	0.014
Difference	—	0.72	0.000

Problema de Seleção - formalizando...

- ▶ dummy de tratamento $D_i = 0, 1$
- ▶ Resultado Potencial $Y_i = 0, 1$

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

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

- ▶ porém não observamos ambos estados para o mesmo indivíduo

Problema de Seleção - formalizando...

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{diferença observada em saúde média}} = \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{efeito tratamento médio nos tratados}} + \underbrace{E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]}_{\text{viés de seleção}}$$

- ▶ viés de seleção mascara o efeito causal do tratamento
- ▶ grande objetivo da pesquisa empírica é superar este viés

Benchmark: Experimento Ideal

- ▶ desenho de pesquisa mais confiável e influente
- ▶ idéia: **tratamento e controle são aleatórios**
- ▶ resolve o **problema de seleção!**

- ▶ e.g. *Perry treatment group, 1962-1997*
- ▶ *123 jovens negros escolhidos aleatoriamente*
- ▶ *moldes para o programa pré-escolar Head Start*
 - ▶ *reanálise dos dados: efeito somente sobre meninas (Anderson, 2008)*

Alocação Aleatória e o Problema de Seleção

- ▶ alocação aleatória do tratamento resolve o problema da seleção
- ▶ torna D_i independente de Y_i

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0] \\ &= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1] \\ &= E[Y_{1,i}] - E[Y_{0,i}] \end{aligned}$$

- ▶ efeito causal é igual a simples diferença de médias condicionais
- ▶ aleatorização resolve o grande problema da seleção
- ▶ porém há outros problemas...

Tennessee STAR Experiment

- ▶ qual o efeito do tamanho da classe sobre aprendizado?
- ▶ literatura não-experimental não acha efeito
- ▶ qual o problema de seleção aqui?

STAR experiment: 1985-1986 com 11,600 crianças na pré-escola

- ▶ acompanha os alunos por 4 anos (até 3^a série)
- ▶ controle: 22-25 alunos com assistente tempo parcial
- ▶ tratamento 1: 13-17 alunos
- ▶ tratamento 2: 22-25 alunos com assistente tempo integral
- ▶ escolas com 3+ classes por série podiam participar

Tennessee STAR Experiment

Aleatorizacao foi alcançada (pelo menos nas observáveis)?

Table 2.2.1: Comparison of treatment and control characteristics in the Tennessee STAR experiment

Variable	Students who entered STAR in kindergarten			Joint P-value
	Small	Regular	Regular/Aide	
1. Free lunch	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate	.49	.52	.53	.02
5. Class size in kindergarten	15.10	22.40	22.80	.00
6. Percentile score in kindergarten	54.70	48.90	50.00	.00

Análise de Regressão de Experimentos

- dados observados reescritos como regressão:

$$Y_i = \underbrace{\alpha}_{E[Y_{0i}]} + \underbrace{\rho}_{Y_{1i}-Y_{0i}} D_i + \underbrace{\eta_i}_{Y_{0i}-E[Y_{0i}]}$$

- tirando a média condicional:

$$E[Y_i|D_i = 1] = \alpha + \rho + E[\eta_i|D_i = 1]$$

$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0]$$

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{\rho}_{\text{treatment}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{selection bias}}$$

- viés de seleção corresponde a correlação entre erro η_i e tratamento D_i

Tennessee STAR Experiment

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	–	–	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	–	–	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	–	–	-13.15 (.77)	-13.07 (.77)
White teacher	–	–	–	-.57 (2.10)
Teacher experience	–	–	–	.26 (.10)
Master's degree	–	–	–	-0.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R ²	.01	.25	.31	.31

Propriedades Mecânicas da Regressão

- ▶ regresões são úteis para estimar efeito tratamento de um experimento (com ou sem variáveis de controle)
- ▶ sem tratamento alertório, a interpretação causal não está garantida
- ▶ essa é a questão chave do curso!
- ▶ mas antes... vamos rever algumas propriedades da regressão que valem independente da interpretação dos coeficientes
 - ▶ conexão entre regressão populacional e esperança condicional
 - ▶ distribuição amostral das estimativas

Relações Econômicas e Esperança Condicional

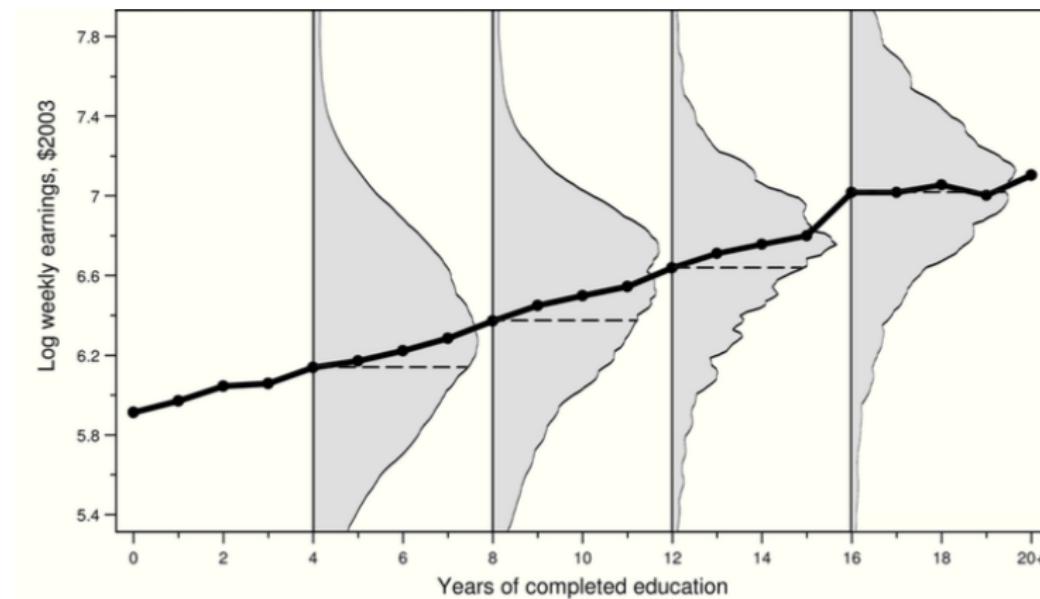
- ▶ diferenças no desempenho econômico dos indivíduos é aleatória
- ▶ regressões sistematizam aleatoriedade
 - ▶ relação entre educação e escolaridade
 - ▶ fato: pessoas mais educadas ganham mais!
 - ▶ independente de ser ou não uma relação causal
- ▶ educação estatisticamente prevê salário
- ▶ esse poder de previsão fica claro ao vermos que:
 - ▶ regressão \approx esperança condicional

Relações Econômicas e Esperança Condisional

$$\begin{aligned} E[Y_i|X_i = x] &= \int t f_y(t|X_i = x) dt \\ E[Y_i|X_i = x] &= \sum_t t P(Y_i = t|X_i = x) \end{aligned}$$

- ▶ Note: esperança é um conceito populacional
- ▶ por enquanto, não fazemos a distinção
- ▶ inferimos a EC populacional da EC amostral

Relações Econômicas e Esperança Condisional



Lei das Expectativas Iteradas

A LEI mostra que uma esperança incondicional pode ser escrita como uma média incondicional das esperanças condicionais:

$$E[Y_i] = E_X\{E[Y_i|X_i]\}$$

Prova: seja (X_i, Y_i) contínuos, $f_{xy}(u, t)$ distribuição conjunta, $f_y(t|X_i = u)$ distribuição condicional de Y_i dado $X_i = u$ e $g_y(t)$ e $g_x(u)$ as densidades marginais

Lei das Expectativas Iteradas - prova

$$E[Y_i] = E\{E[Y_i|X_i]\}$$

$$\begin{aligned} E\{E[Y_i|X_i]\} &= \int E[Y_i|X_i = u]g_x(u)du \\ &= \int [\int tf_y(t|X_i = u)dt]g_x(u)du \\ &= \int \int tf_y(t|X_i = u)g_x(u)dudt \\ &= \int t[\int f_y(t|X_i = u)g_x(u)du]dt \\ &= \int t[\int f_{xy}(u, t)du]dt \\ &= \int tg_y(t)dt = E[Y_i] \end{aligned}$$

Propriedade da decomposição da FEC

A LEI nos permite separar uma variável aleatória em dois pedaços, a FEC (*parte explicada por X*) e um resíduo ortogonal a qualquer função de X_i :

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

- (i) ϵ_i é mean-independent de X_i ou seja $E[\epsilon_i|X_i] = 0$,
 - $E[\epsilon_i|X_i] = E[Y_i - E[Y_i|X_i]|X_i] = E[Y_i|X_i] - E[Y_i|X_i] = 0$
- (ii) ϵ_i é não-correlacionado com qualquer função de X_i .
 - $E[h(X_i)\epsilon_i] = E\{h(X_i)E[\epsilon_i|X_i]\}$, e dado (i), $E[\epsilon_i|X_i]=0$

Outras Propriedades da FEC

A FEC é a melhor previsão de Y_i dado X_i :

- ▶ intuição: FEC resolve um problema de MMSE
- ▶ $E[Y_i|X_i] = \operatorname{argmin}_{m(X_i)} E[(Y_i - m(X_i))^2]$

A decomposição da ANOVA (analysis of variance):

- ▶ $\operatorname{Var}(Y_i) = \operatorname{Var}(E[Y_i|X_i]) + E[V(Y_i|X_i)]$
- ▶ intuição: $\operatorname{Var}(Y_i)$ é decomposto no que pode ser explicado pelas observadas, e pelo resíduo

Estas propriedades valem tanto para população quanto para amostra

- ▶ nada até aqui assumiu linearidade da FEC

Regressão Linear e FEC

Seja β um vetor $k \times 1$ definido como: $\beta = \operatorname{argmin}_b E[(Y_i - X'_i b)^2]$

Pela condição de primeira ordem: $E[X'_i (\underbrace{Y_i - X'_i b}_{\epsilon_i})] = 0$

- ▶ a solução é $\beta = E[X_i X'_i]^{-1} E[X'_i Y_i]$
- ▶ note ϵ_i que não é observado, ele é uma função de β
- ▶ na regressão linear, os resíduos são “forçados” a respeitarem a ortogonalidade em relação a X_i
- ▶ o que acontece se a regressão estiver mal especificada?

3 Links entre Regressão e FEC

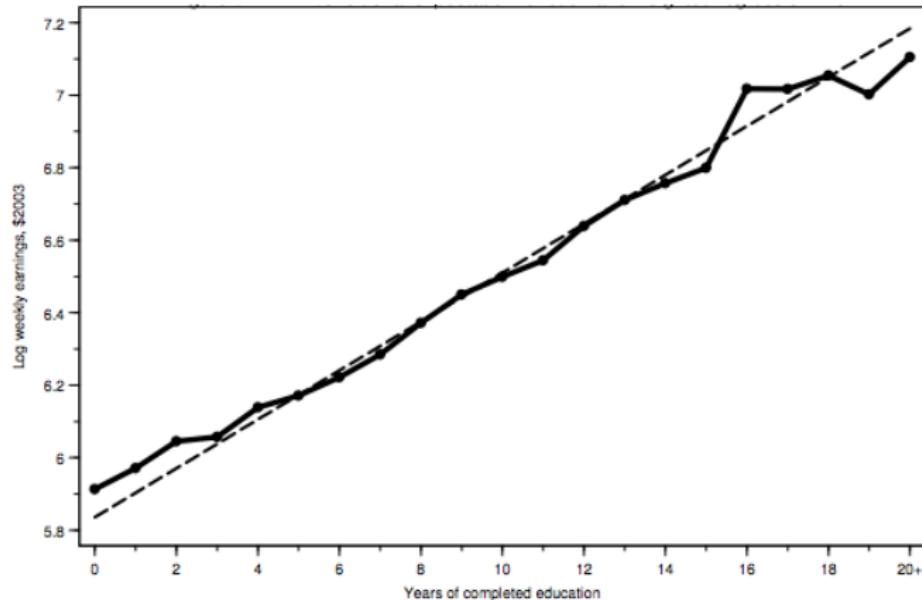
Porque usamos a Regressão Linear?

1. **Teorema da FEC Linear:** Se FEC linear, regressão = FEC
 - ▶ quando a FEC é linear? (i) $f_{x,y}$ normal, (ii) modelo saturado
2. **Teorema do Melhor Previsor Linear**
 - ▶ $X_i\beta$ é a melhor previsão linear de Y_i dado X_i
3. **Teorema da Regressão FEC**
 - ▶ $X_i\beta$ é a melhor aproximação linear (MMSE) de $E[Y_i|X_i]$

FEC é a melhor previsão irrestrita de Y_i

Regressão é a melhor aproximação linear de $E[Y_i|X_i]$

Porque usamos a Regressão Linear?



Regressão e Causalidade

Regressão e Causalidade

- ▶ quando podemos considerar que um coeficiente de uma regressão aproxima o efeito causal verdadeiro?
- ▶ relação escolaridade e salário causal se:
 - ▶ mudarmos escolaridade das pessoas de maneira perfeitamente controlada
 - ▶ escolaridade for distribuída aleatoriamente
- ▶ Hipótese da Independência Condicional (HIC)
 - ▶ controlando por observáveis, tratamento é aleatório
 - ▶ seleção em observáveis

Hipótese da Independência Condicional

- Resultado Potencial $Y_i = 0, 1$

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

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

- onde D_i é a decisão de fazer faculdade
- nunca observamos ambos $Y_{0,i}$ e $Y_{1,i}$

Hipótese da Independência Condisional

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{diferença observada em salário médio}} = \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{efeito tratamento médio nos tratados}} + \underbrace{E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]}_{\text{viés de seleção}}$$

- ▶ viés de seleção aqui deve ser positivo (superestima)
- ▶ HIC = { $Y_{0,i}, Y_{1,i}\}$ $\perp D_{0,i}|X_i$
- ▶ $E[Y_{0,i}|X_i, D_i = 1] = E[Y_{0,i}|X_i, D_i = 0]$

Hipótese da Independência Condicional

Vamos expandir para o caso de uma variável contínua:

$$Y_{si} = f_i(s)$$

$$Y_{0i} = f_i(12); Y_{1i} = f_i(16)$$

HIC = $Y_{si} \perp s_i | X_i$ para qualquer s

Hipótese da Independência Condisional

- ▶ efeito do ensino médio completo:

$$E[Y_i|X_i, s_i = 12] - E[Y_i|X_i, s_i = 11]$$

$$E[f_i(12)|X_i, s_i = 12] - E[f_i(11)|X_i, s_i = 11]$$

- ▶ isso nos dá um efeito causal para cada $X=x$
- ▶ Pela Lei das Expectativas Iteradas, o efeito incondicional:

$$E[E[Y_i|X_i, s_i = 12] - E[Y_i|X_i, s_i = 11]]$$

$$E[E[f_i(12) - E[f_i(11)]]|X_i]$$

$$E[f_i(12) - E[f_i(11)]]$$

- ▶ qual é a contraparte empírica? qual o estimador?

HIC na regressão

- ▶ Seleção nas observáveis

$$Y_i = \alpha + \beta X_i + \eta_i$$

$$\eta_i = A'_i \gamma + \xi_i$$

$$Y_i = \alpha + \beta X_i + A'_i \gamma + \xi_i$$

- ▶ mas e se não observarmos A_i ?

Viés de Variável Omitida

- ▶ regressão correta é $Y_i = \alpha + \beta X_i + A_i \gamma + \xi_i$
- ▶ como não observamos A_i , rodamos $Y_i = \alpha + \beta X_i + \xi_i$

$$\hat{\beta} = (X_i' X_i)^{-1} (X_i' Y_i)$$

$$\hat{\beta} = (X_i' X_i)^{-1} X_i' X_i \beta + (X_i' X_i)^{-1} X_i' A_i \gamma + (X_i' X_i)^{-1} X_i' \xi_i$$

$$E[\hat{\beta}] = \beta + \underbrace{E[(X_i' X_i)^{-1} X_i' A_i] \gamma}_{\text{viés}}$$

- ▶ $E[(X_i' X_i)^{-1} X_i' A_i] = Cov(X, A) / V(X) = \text{delta}_{AX}$
- ▶ quando o viés é nulo? 2 situações...

Viés de Variável Omitida

Table 3.2.1: Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupation dummies
	0.132 (0.007)	0.131 (0.007)	0.114 (0.007)	0.087 (0.009)	0.066 (0.010)

- ▶ variável omitida é observável sempre?
- ▶ qualquer controle é um bom controle?

Quando/Quais controles devem ser incluídos?

- ▶ todas que foram fixadas ao momento que o regressor foi determinado
- ▶ NÃO incluir se a variável for também dependente!!

- ▶ ex: ensino superior e salário (white/blue collar)
- ▶ devemos controlar por ocupação?
- ▶ e se ensino superior for aleatório para a amostra como um todo?

Proxies são controles bons?

- ▶ regressão de interesse é $Y_i = \alpha + \beta s_i + a_i \gamma + \epsilon_i$
 - ▶ onde a_i é habilidade (QI) medida antes da faculdade
- ▶ não temos dados sobre a_i , mas temos sobre uma medida de QI a_i^* de uma prova dada ao fim da faculdade
- ▶ em geral, teste de QI reflete inteligência e escolaridade:
 - ▶ $a_i^* = \pi_0 + \pi_1 s_i + \pi_2 a_i$
- ▶ tentando evitar o viés da variável omitida, incluímos a_i^* :
 - ▶ regressão correta é $Y_i = (\alpha - \gamma \frac{\pi_0}{\pi_2}) + (\beta - \gamma \frac{\pi_1}{\pi_2}) s_i + (\frac{\gamma}{\pi_2}) a_i^*$

Proxies são controles bons?

- ▶ proxy são bons controles somente se determinados antes do regressor de interesse
 - ▶ em outras palavras, se $\pi_1 = 0$
 - ▶ timing é muito importante aqui (nem sempre óbvio)
- ▶ caso π_1 positivo, sabemos que o coeficiente verdadeiro está entre a estimativa sem controles e a com a proxy mal especificada

Matching Estimators I

Raphael Corbi

Universidade de São Paulo

May 2021

Introduction to potential outcomes model

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 & \text{if hospitalized at time } t \\ 0 & \text{if not hospitalized at time } t \end{cases}$$

where i indexes an individual observation, such as a person

- Potential outcomes:

$$Y_{i,t}^j = \begin{cases} 1 & \text{health if hospitalized at time } t \\ 0 & \text{health if not hospitalized at time } t \end{cases}$$

where j indexes a counterfactual state of the world

Decomposition of difference in means

$$\underbrace{E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D = 1] - E[Y^0|D = 0]}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

where $E_N[Y|D = 1] \rightarrow E[Y^1|D = 1]$,
 $E_N[Y|D = 0] \rightarrow E[Y^0|D = 0]$ and $(1 - \pi)$ is the share of the population in the control group.

Random Assignment Solves the Selection Problem

$$\underbrace{E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D = 1] - E[Y^0|D = 0]}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

- If treatment is independent of potential outcomes, then swap out equations and **selection bias** zeroes out:

$$E[Y^0|D = 1] - E[Y^0|D = 0] = 0$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D=1] - E[Y^0|D=1]$$

$$\text{ATU} = E[Y^1|D=0] - E[Y^0|D=0]$$

- Rewrite the third row bias after $1 - \pi$:

$$\begin{aligned}\text{ATT} - \text{ATU} &= E[Y^1 | D=1] - E[Y^0 | D=1] \\ &\quad - E[Y^1 | D=0] + E[Y^0 | D=0] \\ &= 0\end{aligned}$$

- If treatment is independent of potential outcomes, then:

$$\begin{aligned}E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0] &= E[Y^1] - E[Y^0] \\ SDO &= ATE\end{aligned}$$

What if you can't conduct randomized experiment?

- Problems with the experimental design itself:
 - non-compliance by administrators
 - non-compliance by members of the treatment group
 - non-compliance by members of the control group
- Experiments may be impractical due to:
 - Too expensive
 - Unethical
 - Not feasible for some other reason

Figure 1
Lung Cancer at Autopsy: Combined Results from 18 Studies

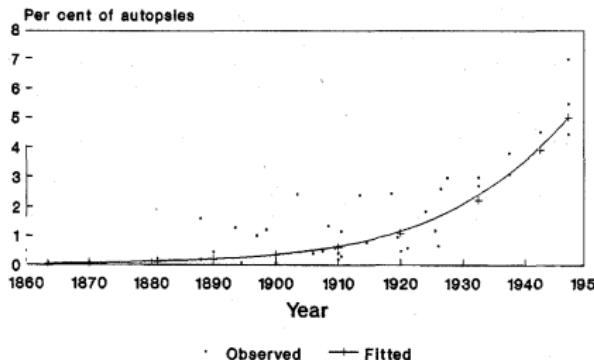


Figure 2(a)
Mortality from Cancer of the Lung in Males

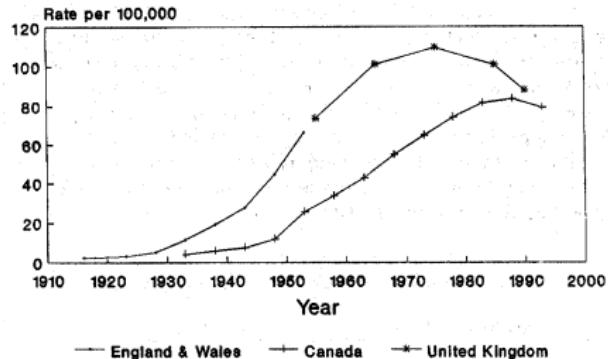


Figure 4
Smoking and Lung Cancer Case-control Studies

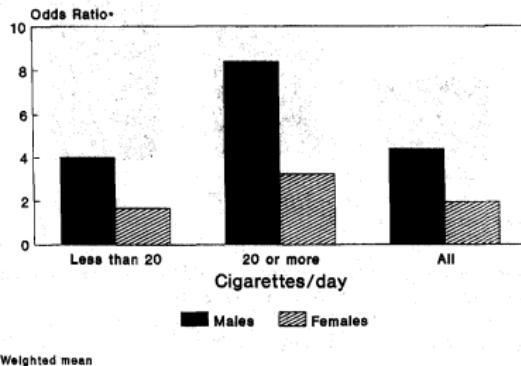
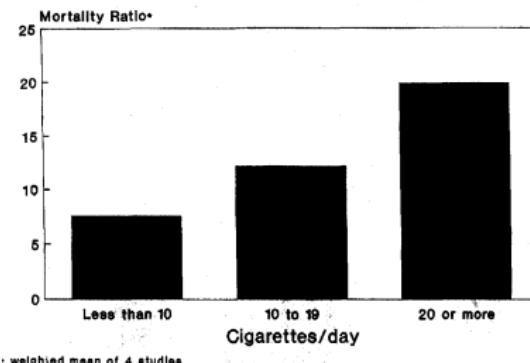
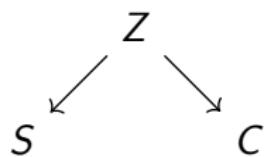


Figure 5
Smoking and Lung cancer Cohort Studies in Males



Does Smoking Cause Cancer?

Smoking, S , causes lung cancer, C ($S \rightarrow C$) versus spurious correlation due to backdoor path:



Nature of the criticism

Criticisms from Joseph Berkson, Jerzy Neyman and Ronald Fisher:
(Hill, Millar and Connelly 2003)

- ① Correlation b/w smoking and lung cancer is spurious due to biased selection of subjects (e.g., conditioning on collider problem)
- ② Functional form complaints about using “risk ratios” and “odds ratios”
- ③ Confounder, Z , creates backdoor path between smoking and cancer
- ④ Implausible magnitudes
- ⑤ No experimental evidence to incriminate smoking as a cause of lung cancer

Fisher's confounding theory

- Fisher, equally famous as a geneticist, argued from logic, statistics and genetic evidence for a hypothetical confounding genome, Z , and therefore smokers and non-smokers were not exchangeable (violation of independence assumption)
- Other studies showed that cigarette smokers and non-smokers were different on observables – more extraverted than non-smokers and pipe smokers, differed in age, differed in income, differed in education, etc.

Hindsight is 20/20

- Fisher was a chain smoking pipe smoker, he died of cancer, and he was a paid expert witness for the tobacco industry.
- But cynicism aside, it is easy to criticize Fisher because we look back with more information to when the smoking/lung cancer link was not universally accepted, and evidence for the *causal* link was shallow:

“the [the epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions.” Robert Hooke’s (1983)

Motivation: Smoking and Mortality

Table: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Are cigars dangerous?

Non-smokers and smokers differ in mortality and age

Table: Mean ages, years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

- Older people die at a higher rate, and for reasons other than just smoking cigars
- Maybe cigar smokers higher observed death rates is because they're older on average

Subclassification

- One way to think about the problem is that the covariates are *not balanced* – their mean values differ for treatment and control group. So let's try to balance them.
- Worth a pause - blocking on confounders vs controlling for covariates. The latter reduces residual variance, but shouldn't affect the bias of the estimator. *Ceteris paribus* vs blocking
- Subclassification (also called stratification): Compare mortality rates across the different smoking groups *within* age groups so as to neutralize covariate imbalances in the observed sample

Subclassification

- Divide the smoking group samples into age groups
- For each of the smoking group samples, calculate the mortality rates for the age group
- Construct probability weights for each age group as the proportion of the sample with a given age
- Compute the weighted averages of the age groups mortality rates for each smoking group using the probability weights

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What is the average death rate for pipe smokers?

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What is the average death rate for pipe smokers?

$$15 \cdot \left(\frac{11}{40}\right) + 35 \cdot \left(\frac{13}{40}\right) + 50 \cdot \left(\frac{16}{40}\right) = 35.5$$

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

$$15 \cdot \left(\frac{29}{40} \right) + 35 \cdot \left(\frac{9}{40} \right) + 50 \cdot \left(\frac{2}{40} \right) = 21.2$$

Table: Adjusted death rates using 3 age groups (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Covariates

Definition: Predetermined Covariates

Variable X is predetermined with respect to the treatment D (also called “pretreatment”) if for each individual i , $X_i^0 = X_i^1$, i.e., the value of X_i does not depend on the value of D_i . Such characteristics are called *covariates*.

Comment I: Does not imply X and D are independent

Comment II: Predetermined variables are often time invariant (e.g., sex, race), but time invariance is not a necessary condition

Comment III: Beware of colliders

Outcomes

Definition: Outcomes

Those variables, Y , that are (possibly) not predetermined are called *outcomes* (for some individual i , $Y_i^0 \neq Y_i^1$)

Adjustment for Observables

- Subclassification (Cochran 1968)
- Nearest Neighbor matching (Abadie and Imbens 2006, 2008)
- Propensity score (Rosenbaum and Rubin 1973)
- Multivariate regression

Identification under independence

- Recall that randomization implies

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

- and therefore:

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{by the switching equation}} \\ &= \underbrace{E[Y^1] - E[Y^0]}_{\text{by independence}} \\ &= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} \end{aligned}$$

- As well as that $ATT = ATE$:

$$E[Y^1 - Y^0] = E[Y^1 - Y^0|D=1]$$

Identification under conditional independence

Identification assumptions:

- ① $(Y^1, Y^0) \perp\!\!\!\perp D | X$ (conditional independence)
- ② $0 < Pr(D = 1 | X) < 1$ with probability one (common support)

Identification result:

- Given assumption 1:

$$\begin{aligned} E[Y^1 - Y^0 | X] &= E[Y^1 - Y^0 | X, D = 1] \\ &= E[Y | X, D = 1] - E[Y | X, D = 0] \end{aligned}$$

- Given assumption 2:

$$\begin{aligned} \delta_{ATE} &= E[Y^1 - Y^0] \\ &= \int E[Y^1 - Y^0 | X, D = 1] dPr(X) \\ &= \int (E[Y | X, D = 1] - E[Y | X, D = 0]) dPr(X) \end{aligned}$$

Identification under conditional independence

Identification assumptions:

- ① $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
- ② $0 < Pr(D = 1|X) < 1$ with probability one (common support)

Identification result:

- Similarly

$$\begin{aligned}\delta_{ATT} &= E[Y^1 - Y^0 | D = 1] \\ &= \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X|D = 1)\end{aligned}$$

- To identify δ_{ATT} the conditional independence and common support assumptions can be relaxed to:
 - ① $Y^0 \perp\!\!\!\perp D|X$
 - ② $Pr(D = 1|X) < 1$ (with $Pr(D = 1) > 0$)

Subclassification estimator

- The identification result is:

$$\delta_{ATE} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X)$$

$$\delta_{ATT} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X|D=1)$$

- Assume X takes on K different cells $\{X^1, \dots, X^k, \dots, X^K\}$.
Then the analogy principle suggests the following estimators:

$$\hat{\delta}_{ATE} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$$

$$\hat{\delta}_{ATT} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$$

where N^k is the number of obs. and N_T^k is the number of treatment observations in cell k ; $\bar{Y}^{1,k}$ is the mean outcome for the treated in cell k ; $\bar{Y}^{0,k}$ is the mean outcome for the control in cell k

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N}\right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N}\right)$?

$$4 \cdot \left(\frac{13}{30}\right) + 6 \cdot \left(\frac{17}{30}\right) = 5.13$$

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{7}{10} \right) = 5.4$$

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Not identified!

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 5 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{4}{10} \right) = 5.1$$

Curse of Dimensionality

- Subclassification may become less feasible in finite samples as the number of covariates grows (e.g., $K = 4$ was too many for this sample)
- Assume we have k covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low, medium, high, etc.)
- The number of sub classification cells (or “strata”) is 3^k . For $k = 10$, then it’s $3^{10} = 59,049$

Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both. If so, we cannot use sub classification.
- Subclassification is also a problem if the cells are “too coarse”. We can always use “finer” classifications, but finer cells worsens the dimensional problem, so we don’t gain much from that. ex: using 10 variables and 5 categories for each, we get $5^{10} = 9,765,625$.

Nearest Neighbor Matching

- See Abadie and Imbens (2006). “Large sample properties of matching estimators for average treatment effects”.
Econometrica
- We could also estimate δ_{ATT} by *imputing* the missing potential outcome of each treatment unit i using the observed outcome from that outcome’s “nearest” neighbor j in the control set

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where $Y_{j(i)}$ is the observed outcome of a control unit such that $X_{j(i)}$ is the **closest** value to X_i among all of the control observations (eg match on X)

Matching

- We could also use the average observed outcome over M closest matches:

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

- Works well when we can find good matches for each treatment group unit, so M is usually defined to be small (i.e., $M = 1$ or $M = 2$)

Alternative distance metric: Euclidean distance

When the vector of matching covariates, $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ has more than one dimension ($k > 1$) we will need a new definition of distance to measure “closeness”.

Definition: Euclidean distance

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

Comment: The Euclidean distance is not invariant to changes in the scale of the X 's. For this reason, alternative distance metrics that are invariant to changes in scale are used

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$$

where $\hat{\Sigma}_X$ is the sample variance-covariance matrix of X .

Avoiding dimensionality problems

- Curse of dimensionality makes matching on K covariates challenging
- Rubin (1977) and Rosenbaum and Rubin (1983) develop a method that can contain those K covariates used for adjusting
- Insofar as treatment is random conditional on K covariates, then one can use the propensity score to adjust for confounders

The Idea behind propensity scores

- Earlier we matched on X 's to compare units “near” one another based on some distance but matching discrepancies and sparseness created problems
- Propensity scores summarize covariate information about treatment selection into a single number bounded between 0 and 1 (i.e., a probability)
- Now we compare units with similar *estimated probabilities* of treatment
- And once we adjust using the propensity score, we no longer need to adjust for X

Identifying assumptions

- We need two assumptions for propensity scores to help us identify causal effects
 - ① Conditional independence, or unconfoundedness
 - ② Common support or overlap
- The first is based on state of the art and institutional details sufficient to warrant such a judgment call, making propensity scores arguably more, not less, advanced
- The latter is testable

Identifying assumption I: Conditional independence

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. There exists a set X of observable covariates such that after controlling for these covariates, treatment assignment is *independent of potential outcomes*.

- Conditional on X , treatment assignment is ‘as good as random’.
- ‘As good as random’ is English for “independent of potential outcomes” potential outcomes jargon
- Also sometimes called ‘ignorable treatment assignment’, ‘unconfoundedness’, ‘selection on observables’, ‘exogeneity’, ‘conditional zero mean’
- CIA is assumed, **not tested**, bc potential outcomes are *missing*. Consult your doctor

Identifying assumption II: Common support

For ranges of X , there is a positive probability of being both treated and untreated

- We'll talk about the propensity score in just a second; for now this assumption is only about X
- Assumption requires that there are units in both treatment and control for the range of propensity score
- Recall, RDD did not have common support so relied on extrapolation sensitive to functional form assumptions
- Common support ensures we can find similar enough donors in the control pool
- Unlike CIA, common support is **testable**

Formal Definition

Definition of Propensity score

A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables: $p(X) = \Pr(D = 1|X)$

Two Necessary Identification Assumptions:

- ① $(Y^0, Y^1) \perp\!\!\!\perp D|X$ (CIA)
- ② $0 < \Pr(D = 1|X) < 1$ (common support)

Steps

- ① Estimate the propensity score using logit/probit
- ② Estimate a particular ATE incorporating the propensity score using stratification, imputation, regression, or inverse probability weighting
- ③ Estimate standard errors

Estimating the propensity score

- Estimate the conditional probability of treatment using probit or logit model

$$Pr(D_i = 1|X_i) = F(\beta X_i)$$

- Use the estimated coefficients to calculate the propensity score for each unit i

$$\hat{\rho}_i = \hat{\beta} X_i$$

- Propensity score is the predicted conditional probability of treatment, or the fitted value for each unit – *same thing*

Propensity score theorem

If $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (CIA), then $(Y^1, Y^0) \perp\!\!\!\perp D|\rho(X)$ where $\rho(X) = Pr(D = 1|X)$, the propensity score

- Conditioning on the propensity score is enough to have independence between D and (Y^1, Y^0) (Rosenbaum and Rubin 1983)
- Valuable theorem because of dimension reduction and convergence rate issues which can introduce biases
- Big picture:** You can toss X out if you have $\hat{\rho}$ because all information from X have been absorbed into $\hat{\rho}$

Unbiased estimation of the ATE

Exact methods to do this to be discussed later, but until then, we can say this:

Corollary: Estimating the ATE

If $(Y^1, Y^0) \perp\!\!\!\perp D|X$, we can estimate average treatment effects:

$$E[Y^1 - Y^0 | \rho(X)] = E[Y|D=1, \rho(X)] - E[Y|D=0, \rho(X)]$$

Balancing property

- Because the propensity score is a function of X , we know:

$$\begin{aligned} \Pr(D = 1|X, \rho(X)) &= \Pr(D = 1|X) \\ &= \rho(X) \end{aligned}$$

- Conditional on $\rho(X)$, the probability that $D = 1$ does not depend on X .
- D and X are independent conditional on $\rho(X)$:

$$D \perp\!\!\!\perp X | \rho(X)$$

Balancing property

- So we obtain the **balancing property** of the propensity score:

$$Pr(X|D = 1, p(X)) = Pr(X|D = 0, p(X))$$

conditional on the property score, the distribution of the covariates is the same for treatment and control group units

- We can use this to check if our estimated propensity score actually produces balance:

$$Pr(X|D = 1, \hat{p}(X)) = Pr(X|D = 0, \hat{p}(X))$$

Propensity score theorem

- This theorem tells us the *only* covariate we need to adjust for is the conditional probability of treatment itself (i.e., the propensity score)
- It does not tell us which method we should use to do that adjustment, though, which is an estimation question
- There are options: inverse probability weighting, forms of imputation, stratification, and sometimes even regressions will incorporate the score as weights

Checking the common support assumption

- We can summarize the propensity scores in the treatment and control group and count how many units are off-support
- Crump, et al. (2009) offer a rule of thumb: keep scores on interval [0.1,0.9].
- Tossing out observations beyond those min and max scores
- A histogram of propensity scores by treatment and control group also highlights the overlap problem; software also can help such as teffects overlap

Inverse probability weighting

- I really like the simple method of inverse probability weighting aesthetically because there are no black boxes; it's all non-parametric averaging done through a particular kind of weights based on the propensity score
- IPW involves fewer implementation choices like number of neighbors, common support, etc.
- And because IPW is a smooth estimator, the bootstrap is valid for inference unlike covariate nearest neighbor matching which Abadie and Imbens (2008) show is not valid

Inverse probability weighting

- IPW is basically a reweighting of the outcomes by the propensity score developed in Robins and Rotnitzky (1995), Imbens (2000), Hirano and Imbens (2001)
- The weights can be expressed in two ways – without normalization (Horvitz and Thompson 1952) or normalized (Hajek 1971) – the difference being how well either approach can handle extreme values of the propensity score; the differences come out of the survey sampling literature
- Notation is far far scarier than in fact what we are doing, so I'll show you this in a Stata and R simulation to help pin down the intuition a little better
- We'll start with the basic idea using the Horvitz and Thompson (1952) expression of the weights as it's not as messy.

Inverse Probability Weighting

Proposition

If $Y^1, Y^0 \perp\!\!\!\perp D|X$, then

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - \rho(X)}{\rho(X) \cdot (1 - \rho(X))}\right] \\ \delta_{ATT} &= E[Y^1 - Y^0 | D = 1] \\ &= \frac{1}{Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \rho(X)}{1 - \rho(X)}\right]\end{aligned}$$

IPW Proof

Proof.

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

and the results follow from integrating over $P(X)$ and $P(X|D = 1)$.



Weighting on the propensity score

Previous formulas used population concepts. Switching to samples, we use a two-step estimator:

- ① Estimate the propensity score: $\hat{\rho}(X)$
- ② Use estimated score to produce analog estimators. Let $\hat{\delta}_{ATE}$ and $\hat{\delta}_{ATT}$ be an estimate of the ATE and ATT parameter:

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{\rho}(X_i)}{\hat{\rho}(X_i) \cdot (1 - \hat{\rho}(X_i))}$$

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{\rho}(X_i)}{1 - \hat{\rho}(X_i)}$$

Propensity score matching

- Matching, or what I like to call “imputation”, is another way that utilizes the \hat{p}
- They all use the same first stage, but differ on their second and third stages
- Part of the second stage may be imposing common support through “trimming”, but for different reasons because now this idea of distance is entering and maybe you think some units are “too far away” to be relevant counterfactuals

Standard matching strategy

- Pair each treatment unit i with one or more *comparable* control group unit j , where comparability is in terms of proximity to the estimated propensity score
- Impute the unit's missing counterfactual outcome $Y_{i(j)}$ based on the unit or units chosen in the previous step
- If more than one are “nearest neighbors”, then use the neighbors' weighted outcomes

$$Y_{i(j)} = \sum_{j \in C(i)} w_{ij} Y_j$$

where $C(i)$ is the set of neighbors with $W = 0$ of the treatment unit i and w_{ij} is the weight of control group units j with $\sum_{j \in C(i)} w_{ij} = 1$

Imputing the counterfactuals

A parameter of interest:

$$E[Y_i^1 | D_i = 1] - E[Y_i^0 | D_i = 1]$$

We estimate it as follows

$$\widehat{ATT} = \frac{1}{N_T} = \sum_{i: W_i=1} \left[Y_i - Y_{i(j)} \right]$$

where N_T is the number of matched treatment units in the sample.
Note the difference between *imputation* and weighting

Matching methods

- The probability of observing two units with *exactly* the same propensity score is in principle zero because $p(x)$ is continuous
- Several matching methods have been proposed in the literature, but the most widely used are:
 - Stratification matching
 - Nearest-neighbor matching (with or without caliper)
 - Radius matching
 - Kernel matching
- Typically, one treatment unit i is matched to several control units j , but sometimes one-to-one matching is used

Beating a dead horse

- The propensity score can make groups comparable **but** only on the variables used to estimate the propensity score in the first place. There is **NO** guarantee you are balancing on unobserved covariates.
- If you know that there are important unobservable variables, you may need another tool.
- Remember: randomization ensure that both observable and **unobservable** variables are balanced

Discrete Dependent Variables

Raphael Corbi

Universidade de São Paulo

May 2021

The Idea behind propensity scores

- Earlier we matched on X 's to compare units “near” one another based on some distance but matching discrepancies and sparseness created problems
- Propensity scores summarize covariate information about treatment selection into a single number bounded between 0 and 1 (i.e., a probability)
- Now we compare units with similar *estimated probabilities* of treatment
- And once we adjust using the propensity score, we no longer need to adjust for X

Identifying assumption I: Conditional independence

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. There exists a set X of observable covariates such that after controlling for these covariates, treatment assignment is *independent of potential outcomes*.

- Conditional on X , treatment assignment is ‘as good as random’.
- ‘As good as random’ is English for “independent of potential outcomes” potential outcomes jargon
- Also sometimes called ‘ignorable treatment assignment’, ‘unconfoundedness’, ‘selection on observables’, ‘exogeneity’, ‘conditional zero mean’
- CIA is assumed, **not tested**, bc potential outcomes are *missing*. Consult your doctor

Identifying assumption II: Common support

For ranges of X , there is a positive probability of being both treated and untreated

- We'll talk about the propensity score in just a second; for now this assumption is only about X
- Assumption requires that there are units in both treatment and control for the range of propensity score
- Recall, RDD did not have common support so relied on extrapolation sensitive to functional form assumptions
- Common support ensures we can find similar enough donors in the control pool
- Unlike CIA, common support is **testable**

Formal Definition

Definition of Propensity score

A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables: $p(X) = \Pr(D = 1|X)$

Two Necessary Identification Assumptions:

- ① $(Y^0, Y^1) \perp\!\!\!\perp D|X$ (CIA)
- ② $0 < \Pr(D = 1|X) < 1$ (common support)

Steps

- ① Estimate the propensity score using logit/probit
- ② Estimate a particular ATE incorporating the propensity score using stratification, imputation, regression, or inverse probability weighting
- ③ Estimate standard errors



Multiple Regression Analysis with Qualitative Information

Linear Probability Model (LPM)

- our dependent variable, y , takes on only two values: zero and one.
(e.g. whether an adult has a high school education ($y=1$) or not ($y=0$); whether a college student used illegal drugs during a given school year ($y=1$) or not ($y=0$) ; or whether a firm was taken over by another firm during a given year ($y=1$) or not ($y=0$)).
- What does it mean to write down a regression model such as...

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$$

when y is a binary variable? As y can take on only two values, β_j cannot be interpreted as the change in y given a one-unit increase in x_j , holding all other factors fixed: y either changes from zero to one or from one to zero (or does not change).

Multiple Regression Analysis with Qualitative Information

Linear regression when the dependent variable is binary

Nevertheless, β_j still has useful interpretations. If we assume that the zero conditional mean assumption MLR.4 holds, that is, $E(u|x_1, \dots, x_k)=0$, then we have, as always,

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$$

$$\Rightarrow E(y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

If the dependent variable only takes on the values 1 and 0, a simple property of condition expectation gives us...

$$E(y|x) = 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x)$$



Multiple Regression Analysis with Qualitative Information

Linear regression when the dependent variable is binary

The key point is that when y is a binary variable taking on the values zero and one, it is always true that $P(y = 1|x) = E(y|x)$.

The probability of “success” — that is, the probability that $y = 1$ — is the same as the expected value of y . Thus, we have the important equation

$$\Rightarrow E(y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

$$E(y|x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x)$$

Linear Probability Model (LPM)

$$\Rightarrow P(y = 1|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$



Multiple Regression Analysis with Qualitative Information

Linear regression when the dependent variable is binary

$$\Rightarrow P(y = 1|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

The probability of success, say, $p(x) = P(y = 1|x)$, is a linear function of the x_j , hence the “linear” in the name.

In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\Rightarrow \beta_j = \Delta P(y = 1|x) / \Delta x_j$$

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that $y=1$

Multiple Regression Analysis with Qualitative Information

- Example: Labor force participation of married women

=1 if in labor force, =0 otherwise

$$\widehat{inlf} = .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper} - .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt6} + .0130 \text{ kidsge6}, n = 753, R^2 = .264$$

(.154) (.0014) (.007) (.006)

(.00018) (.002) (.034)

(.0132)

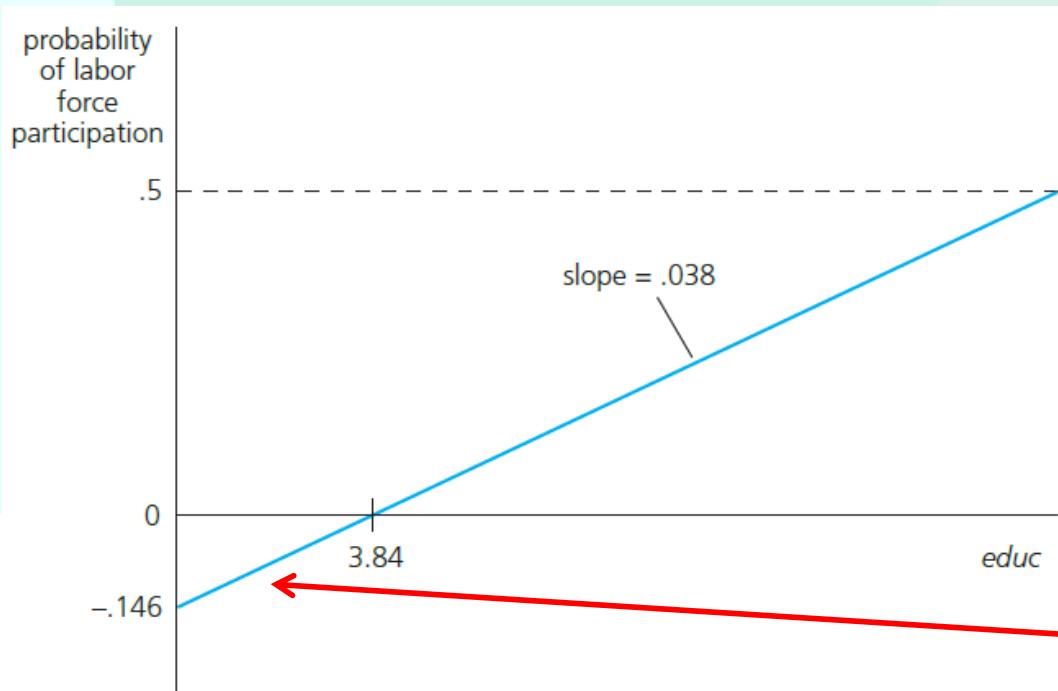
Non-wife income (in thousand dollars per year)

Diminishing returns for experience.
Levels off at exper=.039/.0012=32 year

If the number of kids under six years increases by one, the probability that the woman works falls by 26.2%

Multiple Regression Analysis with Qualitative Information

- Example: Female labor participation of married women (cont.)



Graph for nwifeinc=50, exper=5, age=30, kindslt6=1, and kidsge6=0

The maximum level of education in the sample is *educ*=17. For the given case, this leads to a predicted probability to be in the labor force of about 50%.

Negative predicted probability but no problem because no woman in the sample has *educ* < 5.



Multiple Regression Analysis with Qualitative Information

- **Advantages of the linear probability model**
 - Easy estimation and interpretation
 - Estimated effects and predictions are often reasonably good in practice
 - It usually works well for values of the independent variables that are near the averages in the sample.
 - In the labor force participation example, no women in the sample have four young children; in fact, only three women have three young children. Over 96% of the women have either no young children or one small child, and so we should probably restrict attention to this case when interpreting the estimated equation.



Multiple Regression Analysis with Qualitative Information

- **Disadvantages of the linear probability model**
 - Predicted probabilities may be larger than one or smaller than zero
 - what would it mean to predict that a woman is in the labor force with a probability of -.10? Marginal probability effects sometimes logically impossible
 - In fact, of the 753 women in the sample, 16 of the fitted values are less than zero, and 17 of the fitted values are greater than 1.
 - The linear probability model is necessarily heteroskedastic due to the binary nature of y .

$$Var(y|x) = P(y = 1|x) [1 - P(y = 1|x)] \quad \text{Variance of Bernoulli variable}$$

- Heteroskedasticity consistent standard errors need to be computed



Limited Dependent Variable Models and Sample Selection Corrections

- **So far...**

- we studied the LPM, which is simply an application of the multiple regression model to a **binary** dependent variable.
 - we have discussed its limitations
-
- Binary dep.var. is an example of a **limited dep.var variable** (LDV)
 - An LDV is broadly defined as a dependent variable whose range of values is substantively restricted
 - Nonnegative variables, e.g. wages, prices, interest rates
 - Nonnegative variables with excess zeros, e.g. labor supply
 - Count variables, e.g. the number of arrests in a year
 - Censored variables, e.g. charity donations



Limited Dependent Variable Models and Sample Selection Corrections

- **Logit and Probit models for binary response**
- **Disadvantages of the LPM for binary dependent variables**
 - 1) Predictions sometimes lie outside the unit interval
 - 2) Partial effects of explanatory variables are constant (*linear*)
 - 3) Heterokedasticity by construction
- **These limitations of the LPM can be overcome by using more sophisticated binary response models.**
 - Logit and Probit Models
 - Response probability is a nonlinear function of explanat. variables
 - Efficiency under distributional assumptions



Limited Dependent Variable Models and Sample Selection Corrections

- **Nonlinear models for binary response**
 - In these models, interest lies primarily in the response probability
 - In the LPM, we assume that the **response probability** [$P(y=1|x)$] is linear in a set of parameters β_j .
 - To avoid the LPM limitations (1 and 2), consider a class of binary response models where response probability is a **nonlinear** function of explanat. variables

$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(x\beta)$$

Probability of a "success" given explanatory variables

A cumulative distribution function $0 < G(z) < 1$. The response probability is thus a function of the explanatory variables x .

Shorthand vector notation: the vector of explanatory variables x also contains the constant of the model.



Limited Dependent Variable Models and Sample Selection Corrections

- Various nonlinear functions have been suggested for the function G to make sure that the probabilities are between zero and one. The two we will cover here are used in the vast majority of applications (along with the LPM).
- Both are increasing in z and $G(z) \rightarrow 0$ as $z \rightarrow -\infty$, $G(z) \rightarrow 1$ as $z \rightarrow \infty$
- **Choices for the link function as**

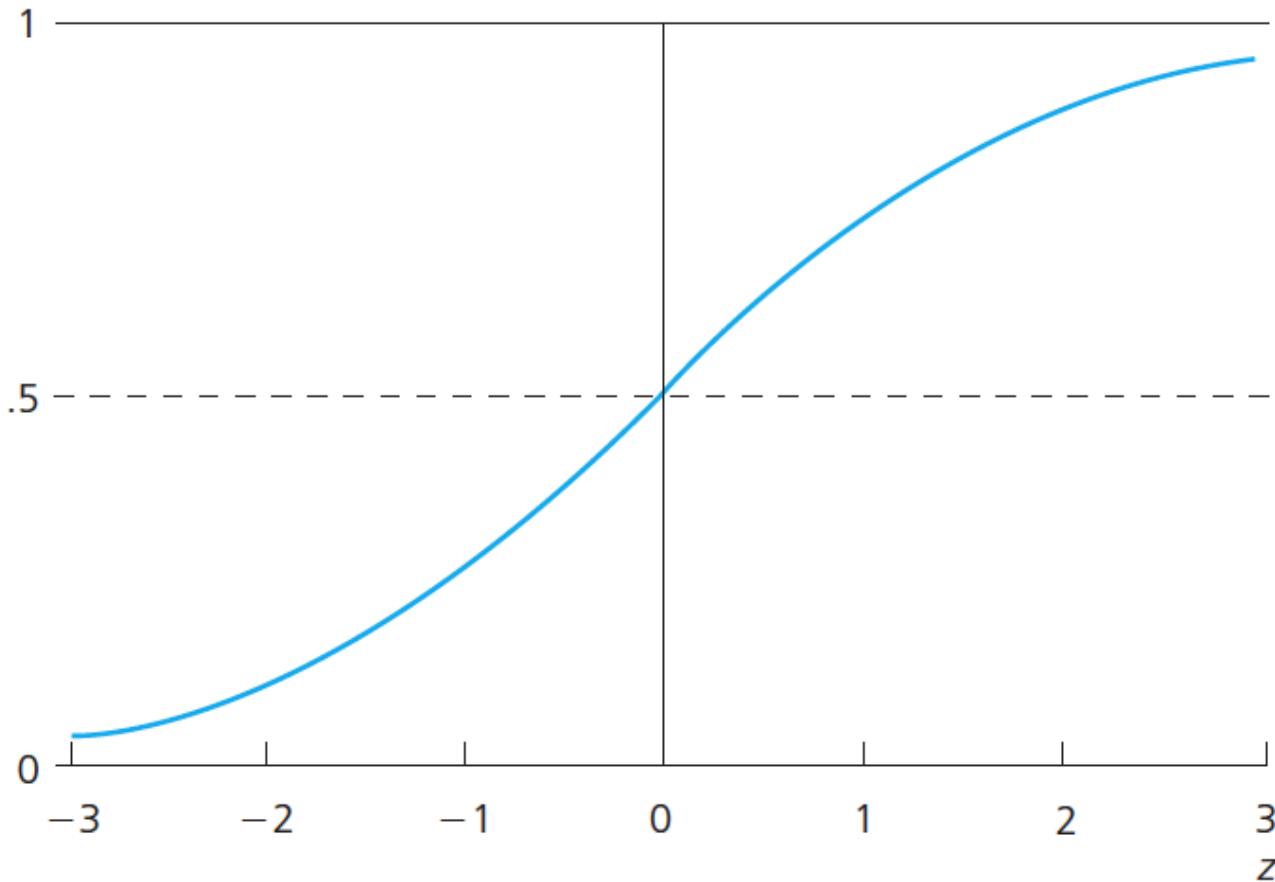
Probit: $G(z) = \Phi(z) \equiv \int_{-\infty}^z \phi(v)dv$ (normal distribution)

Logit: $G(z) = \Lambda(z) = \exp(z) / [1 + \exp(z)]$ (logistic function)



Limited Dependent Variable Models and Sample Selection Corrections

$$G(z) = \exp(z)/[1 + \exp(z)]$$





Limited Dependent Variable Models and Sample Selection Corrections

- **Latent variable formulation of the Logit and Probit models**
 - Logit and probit models can be derived from an underlying latent variable model.
 - Let y^* be an unobserved (latent) variable such that:

$$y^* = \mathbf{x}\beta + e \quad \text{and} \quad y = 1 [y^* > 0]$$

If the latent variable y^* is larger than zero, y takes on the value 1, if it is less or equal zero, y takes on 0 (y^* can thus be interpreted as the propensity to have $y = 1$)

- Assume e independent of \mathbf{x} and that e either has the logistic distribution or the standard normal distribution. In either case, e is symmetrically distributed about zero, which means $G(-z)=1-G(z)$
- We can derive the response probability for y :

$$P(y=1|x) = P(y^*>0|x) = P(e>-\mathbf{x}\beta) = 1 - G(-\mathbf{x}\beta) = G(\mathbf{x}\beta)$$



Limited Dependent Variable Models and Sample Selection Corrections

- **Interpretation of coefficients in Logit and Probit models**
 - As discussed before, Probit/Logit models overcome the problems of LPM: (1) linear effects and (2) probabilities in [0,1].
 - However they come at a cost. The interpretation of β coefficients is not the same as marginal effects as in standard OLS.
- **A)** As we will see, for logit and probit, the direction of the effect of x_j on $E(y^*|x) = \beta x$ and on $E(y|x) = P(y=1|x) = G(\beta x)$ is always the same.
- **B)** But the latent variable y^* rarely has a well-defined unit of measurement. (For example, y^* might be the difference in utility levels from two different actions.) Thus, the magnitudes of each β_j are not, by themselves, especially useful (in contrast to the linear probability model).

Limited Dependent Variable Models and Sample Selection Corrections

- **Interpretation of coefficients in Logit and Probit models**

Continuous explanatory variables:

$$\frac{\partial P(y = 1 | \mathbf{x})}{\partial x_j} = g(\mathbf{x}\boldsymbol{\beta})\beta_j \quad \text{where} \quad g(z) \equiv \partial G(z)/\partial z > 0$$

How does the probability for $y = 1$ change if explanatory variable x_j changes by one unit? **Note: same sign as β**

- **Partial effects are nonlinear and depend on the level of x .**

- however the relative effects of any two continuous explanatory variables do not depend on x : the ratio of the partial effects for x_j and x_h is $g(\mathbf{x}\boldsymbol{\beta})\beta_j / g(\mathbf{x}\boldsymbol{\beta})\beta_h = \beta_j / \beta_h$



Limited Dependent Variable Models and Sample Selection Corrections

- **Interpretation of coefficients in Logit and Probit models**

Discrete explanatory variables:

$$G[\beta_0 + \beta_1 x_1 + \dots + \beta_k(c_k + 1)] - G[\beta_0 + \beta_1 x_1 + \dots + \beta_k c_k]$$

For example, explanatory variable x_k increases by one unit.

- For example, if y is an employment indicator and x_k is a number of children, then eq. above is the change in probability of employment due to the job training program; this depends on other characteristics that affect employability, such as education and experience.
- Again, knowing the sign of β_k is sufficient for determining whether an extra child has a positive or negative effect. But to find the magnitude of the effect, we have to estimate the quantity above.



Limited Dependent Variable Models and Sample Selection Corrections

Maximum likelihood estimation of Logit and Probit models

- How should we estimate nonlinear binary response models?
- To estimate the LPM, we can use ordinary least squares, but because of the nonlinear nature of $E(y|x)$, OLS is not applicable.
- We can use the **Maximum Likelihood Estimation** (MLE). Remember that under the classical linear model assumptions, the OLS estimator is the maximum likelihood estimator. For estimating limited dependent variable models, maximum likelihood methods are indispensable.
- Because MLE is based on the distribution of y given x , the heteroskedasticity in $\text{Var}(y|x)$ is automatically accounted for.



Limited Dependent Variable Models and Sample Selection Corrections

Maximum likelihood estimation of Logit and Probit models

Assume that we have a random sample of size n. To obtain the maximum likelihood estimator, conditional on the explanatory variables, we need the density of y_i given x_i .

We can write this as:

$$f(y_i|x_i; \beta) = [G(x_i\beta)]^{y_i} [1 - G(x_i\beta)]^{1-y_i}$$

The probability that individual i's outcome is y_i given that his/her characteristics are x_i



Limited Dependent Variable Models and Sample Selection Corrections

Maximum likelihood estimation of Logit and Probit models

$$f(y_i|x_i; \beta) = [G(x_i\beta)]^{y_i} [1 - G(x_i\beta)]^{1-y_i}$$

- The log-likelihood function is a function of the parameters and the data (x_i, y_i) and is obtained by taking the log of $f(y_i|x_i;\beta)$.

$$\ell_i(\beta) = y_i \log[G(x_i\beta)] + (1 - y_i) \log[1 - G(x_i\beta)].$$

- Because $G(\cdot)$ is strictly between zero and one for logit and probit, $\ell_i(\beta)$ is well defined for all values of β .
- Finally, the **log-likelihood for a sample size of n** is obtained by summing $\ell_i(\beta)$ across all observations

$$\mathcal{L}(\beta) = \sum_{i=1}^n \ell_i(\beta)$$



Limited Dependent Variable Models and Sample Selection Corrections

Maximum likelihood estimation of Logit and Probit models

- In other words, log-likelihood function is a function of the parameters and the data (x_i, y_i) and is obtained by taking the log of $L(\beta)$.

$$\log L(\beta) = \log \left(\prod_{i=1}^n f(y_i | \mathbf{x}_i; \beta) \right) = \sum_{i=1}^n \log f(y_i | \mathbf{x}_i; \beta) \quad \text{Under random sampling}$$

The MLE of β , denoted $\hat{\beta}$, maximizes this log-likelihood $\log L(\beta)$. If $G(\cdot)$ is the standard logit (normal) cdf, then $\hat{\beta}$ is the logit (probit) estimator.

$$\max \log L(\beta) \rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k \quad \text{Maximum likelihood estimates}$$



Limited Dependent Variable Models and Sample Selection Corrections

Maximum likelihood estimation of Logit and Probit models

$$\max \log L(\beta) \rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k \leftarrow = \text{Maximum likelihood estimates}$$

Properties of maximum likelihood estimators

- As a nonlinear maximization problem, we cannot write formulas for the logit or probit maximum likelihood estimates. The formula for (asymptotic) standard error (chapter appendix).
- Once we have the standard errors, we can construct (asymptotic) t tests and confidence intervals, just as with OLS.
- ML estimators are consistent, asymptotically normal, and asymptotically efficient if the distributional assumptions hold



Limited Dependent Variable Models and Sample Selection Corrections

Hypothesis testing after maximum likelihood estimation

- The usual t-tests and confidence intervals can be used
- There are three alternatives to test multiple hypotheses.

We focus on the Likelihood-Ratio (LR) test.

It is based on the same concept as the F test in a linear model. The F test measures the increase in the sum of squared residuals when variables are dropped from the model. The LR test is based on the difference in the log-likelihood functions for the unrestricted and restricted models.



Limited Dependent Variable Models and Sample Selection Corrections

- **Hypothesis testing after maximum likelihood estimation**
 - The idea is this: Because the MLE maximizes the log-likelihood function, dropping variables generally leads to a smaller—or at least no larger—log-likelihood. (similar to the R-squared) The question is whether the fall in the log-likelihood is large enough to conclude that the dropped variables are important. We can make this decision once we have a test statistic and a set of critical values.

$$LR = 2(\log L_{ur} - \log L_r) \sim \chi_q^2$$

Chi-square distribution with q degrees of freedom

The null hypothesis that the q hypotheses hold is rejected if the growth in maximized likelihood is too large when going from the restricted to the unrestricted model



Limited Dependent Variable Models and Sample Selection Corrections

- **Goodness-of-fit measures for Logit and Probit models**

- Percent correctly predicted

$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\mathbf{x}_i \hat{\beta}) \geq .5 \\ 0 & \text{otherwise} \end{cases}$$

Individual i's outcome is predicted as one if the probability for this event is larger than .5, then percentage of correctly predicted $y = 1$ and $y = 0$ is counted

- Pseudo R-squared

$$\tilde{R}^2 = 1 - \log L_{ur} / \log L_0$$

Compare maximized log-likelihood of the model with that of a model that only contains a constant (and no explanatory variables)

- Correlation based measures

$$\text{Corr}(y_i, \tilde{y}_i), \text{Corr}(y_i, G(\mathbf{x}_i \hat{\beta}))$$

Look at correlation (or squared correlation) between predictions or predicted prob. and true values. Akin to standard R² in OLS.



Limited Dependent Variable Models and Sample Selection Corrections

- **Reporting partial effects of explanatory variables**
 - As discussed before, coefficient estimates are not the same as marginal effects of x on y (they are not constant but depend on x)
 - For a continuous x , then
$$\frac{\partial P(y = 1|x)}{\partial x_j} = g(x\beta)\beta_j$$
 - Average partial effects:
$$\widehat{APE}_j = n^{-1} \sum_{i=1}^n g(x_i\hat{\beta})\hat{\beta}_j$$

 - Analogous formulas hold for discrete explanatory variables

The partial effect of explanatory variable x_j is computed for each individual in the sample and then averaged across all sample members (makes more sense)

Regression Discontinuity

Raphael Corbi

Universidade de São Paulo

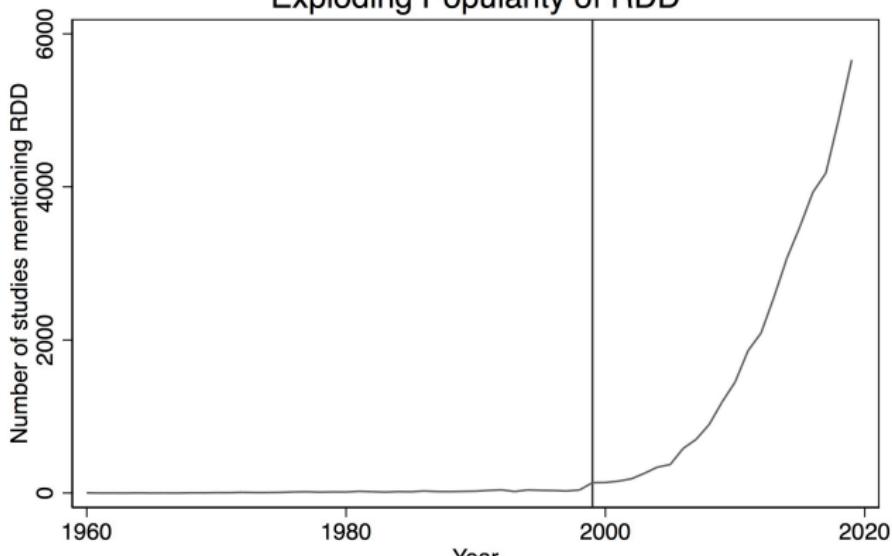
May 2021

What is regression discontinuity design?

Very popular particular type of research design known as *regression discontinuity design* (RDD). Cook (2008) has a fascinating history of thought on how and why.

- Donald Campbell, educational psychologist, invented regression discontinuity design (Thistlethwaite and Campbell, 1960), but then it went dormant for decades (Cook 2008).
- Angrist and Lavy (1999) and Black (1999) independently rediscover it. It's become incredibly popular in economics

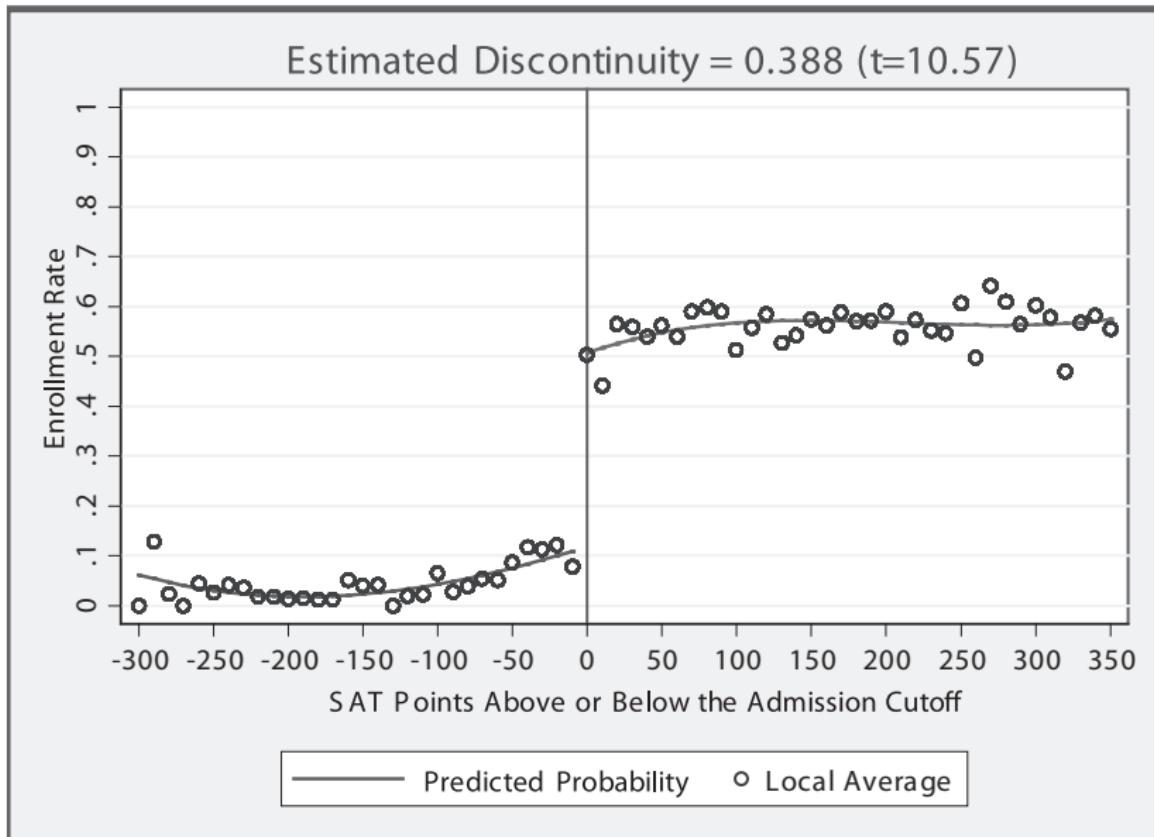
Exploding Popularity of RDD



Vertical bar is Angrist and Lavy (1999) and Black (1999)

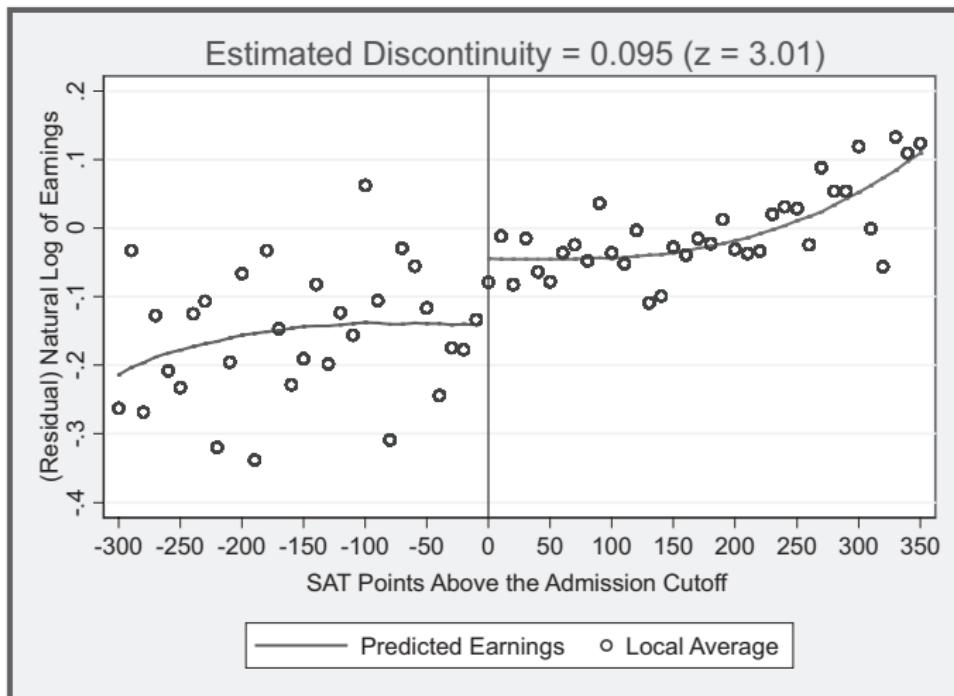
Tell me what you think is happening

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Tell me what you think is happening

FIGURE 2.—NATURAL LOG OF ANNUAL EARNINGS FOR WHITE MEN TEN TO FIFTEEN YEARS AFTER HIGH SCHOOL GRADUATION (FIT WITH A CUBIC POLYNOMIAL OF ADJUSTED SAT SCORE)



What is a regression discontinuity design?

- We want to estimate some causal effect of a treatment on some outcome, but we're worried about selection bias

$$E[Y^0|D = 1] \neq E[Y^0|D = 0]$$

due to self-selection into treatment

- RDD is based on a idea: if treatment assignment occurs abruptly when some underlying variable X called the “running variable” passes a cutoff c_0 , then we can use that to estimate the causal effect *even of a self-selected treatment*

Running and jumping

- Firms, schools and govt agencies have running variables that are used to assign treatments in their rules
- And consequently, probabilities of treatment will “jump” when that running variable exceeds a known threshold
- Most effective RDD studies involve programs where running variables assign treatments based on a “hair trigger”
- Good reasons; inexplicable reasons; arbitrary rules; a choice made by necessity and resource constraints; natural experiments

Selection examples and solutions from the literature

Think of these in light of a treatment where

$$E[Y^0|D = 1] \neq E[Y^0|D = 0]$$

- Yelp rounded a continuous score of ratings to generate stars which Anderson and Magruder 2011 used to study firm revenue
- US targeted air strikes in Vietnam using rounded risk scores which Dell and Querubin 2018 used to study the military and political activities of the communist state
- Card, Dobkin, and Maekas 2008 studied the effect of universal healthcare on mortality and healthcare usage exploiting jumps at age 65
- Almond, et al. 2010 studied the effect of intensive medical attention on health outcomes when a newborn's birthweight fell just below 1,500 grams

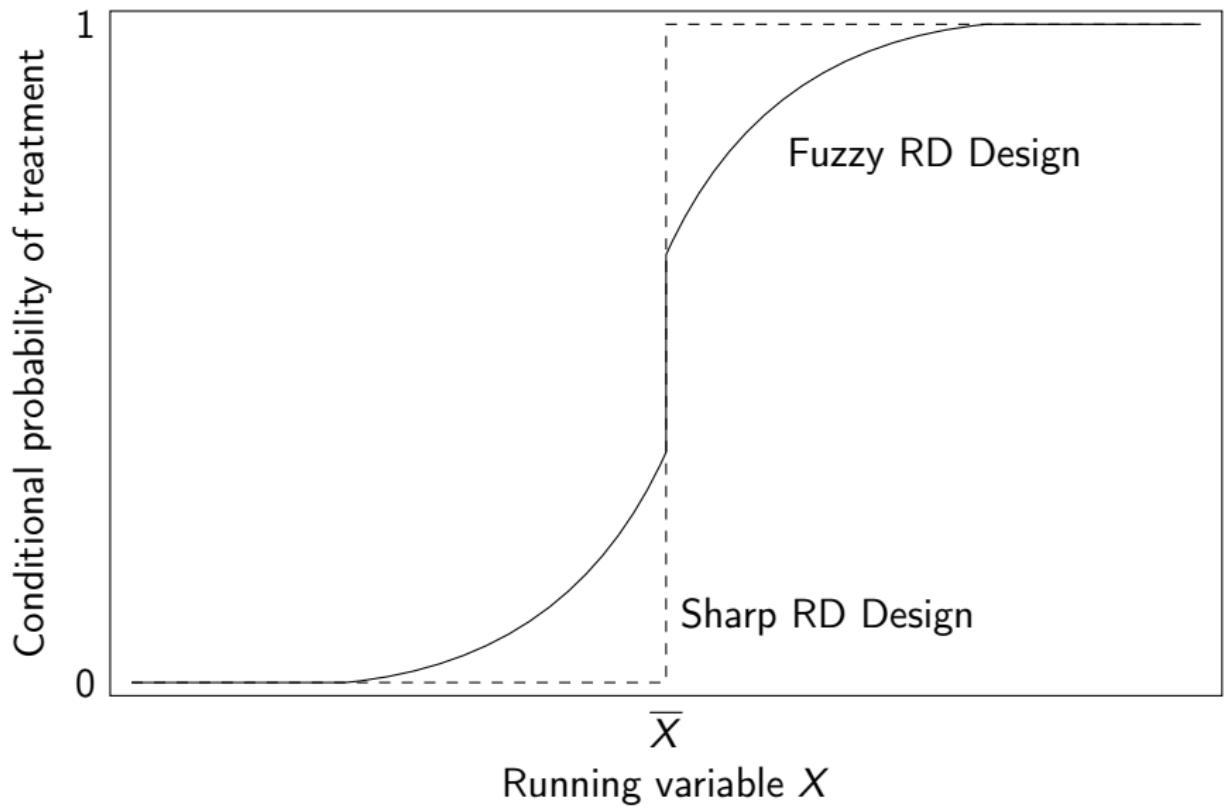


Figure: Sharp vs. Fuzzy RDD

Overlap

- Independence implies an equal distribution of characteristics across two groups guaranteeing overlap
 - In an RCT you can find 65 year olds treated and untreated
- But RDD doesn't have this feature bc you don't have groups with the same value of X in each group, so no overlap
 - 64 years olds are control, not treatment. 66 years olds are in treatment not control
- Some methods require overlap and therefore are off the table without it; but RDD has a workaround using extrapolation

Treatment assignment in the sharp RDD

Deterministic treatment assignment ("sharp RDD")

In Sharp RDD, treatment status is a deterministic and discontinuous function of a covariate, X_i :

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c_0 \\ 0 & \text{if } X_i < c_0 \end{cases}$$

where c_0 is a known threshold or cutoff. In other words, if you know the value of X_i for a unit i , you know treatment assignment for unit i with certainty.

Universal health insurance: Americans aged 64 are not eligible for Medicare, but Americans aged 65 ($X \geq c_{65}$) are eligible for Medicare (ignoring disability exemptions)

Treatment effect definition and estimation

Definition of treatment effect

The treatment effect parameter, δ , is the discontinuity in the conditional expectation function:

$$\begin{aligned}\delta &= \lim_{X_i \rightarrow c_0} E[Y_i^1 | X_i = c_0] - \lim_{c_0 \leftarrow X_i} E[Y_i^0 | X_i = c_0] \\ &= \lim_{X_i \rightarrow c_0} E[Y_i | X_i = c_0] - \lim_{c_0 \leftarrow X_i} E[Y_i | X_i = c_0]\end{aligned}$$

The sharp RDD estimation is interpreted as an average causal effect of the treatment at the discontinuity

$$\delta_{SRD} = E[Y_i^1 - Y_i^0 | X_i = c_0]$$

D is correlated with X and deterministic function of X ; overlap only occurs in the limit and thus the treatment effect is in the limit as X approaches c_0

Extrapolation

Estimation methods attempt to approximate the limiting parameter using units left and right of the cutoff

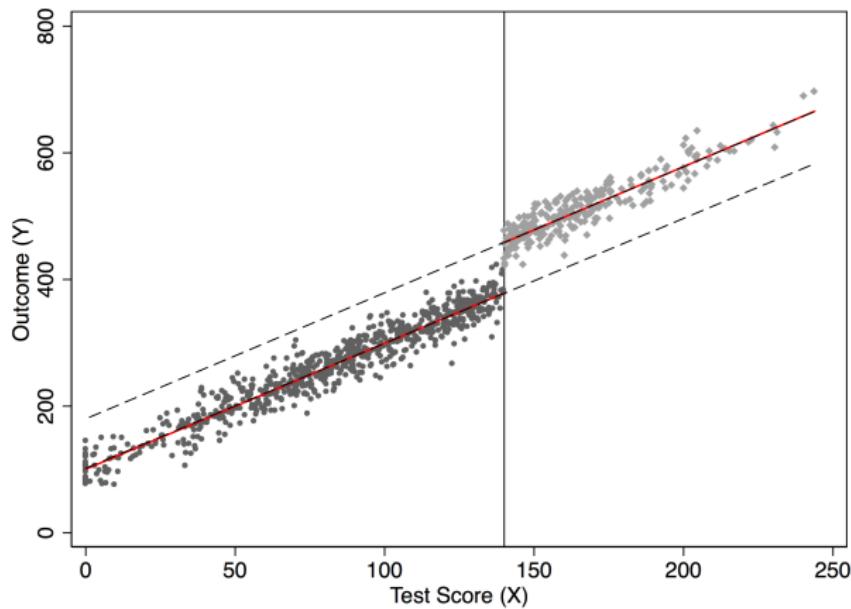


Figure: Dashed lines are extrapolations

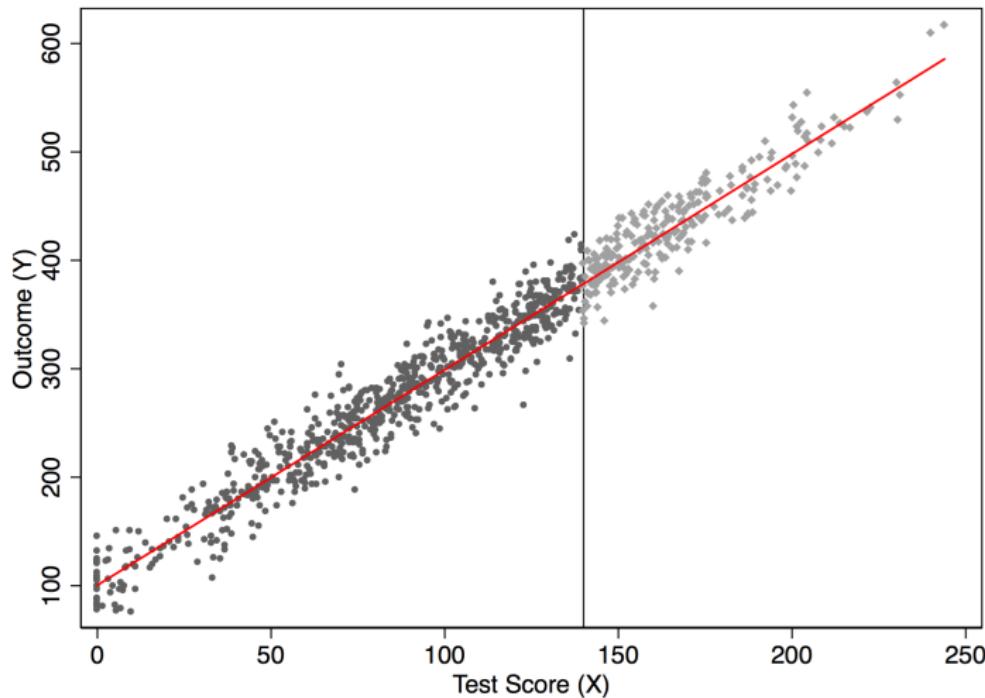
Key identifying assumption

Smoothness (or continuity) of conditional expectation functions
(Hahn, Todd and Van der Klaauw 2001)

$E[Y_i^0|X = c_0]$ and $E[Y_i^1|X = c_0]$ are continuous (smooth) in X at c_0 .

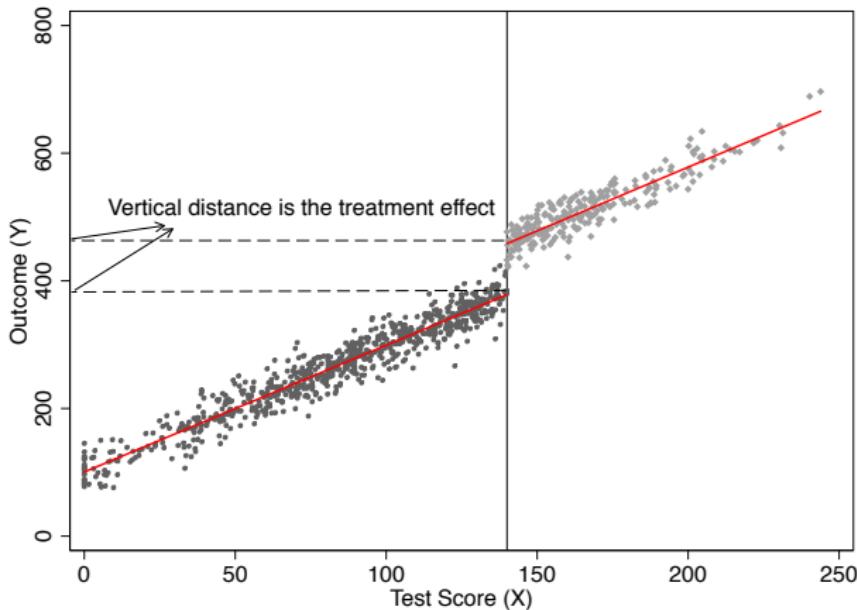
- Potential outcomes not actual outcomes
- If population average *potential outcomes*, Y^1 and Y^0 , are smooth functions of X through the cutoff, c_0 , then potential average outcomes *won't* jump at c_0 .
- Implies the cutoff is exogenous – i.e., nothing else changes related to potential outcomes at c_0
- Unobservables are evolving smoothly, too, through the cutoff

Graphical example of the smoothness assumption



Note these are *potential* not *actual* outcomes

Graphical example of the treatment effect, not the smoothness assumption

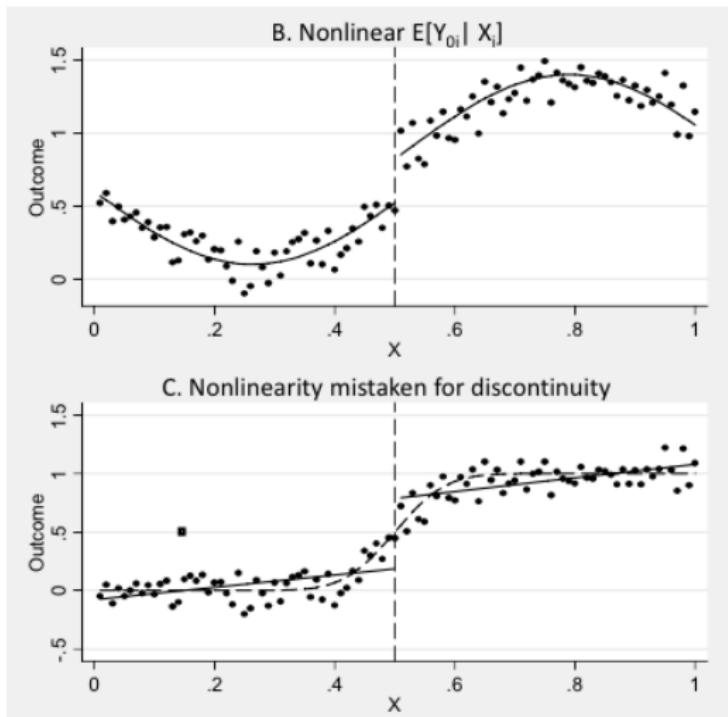


Note that these are *actual*, not *potential* outcomes

Nonlinearity bias

- Smoothness and *linearity* are different things.
- What if the trend relation $E[Y_i^0|X_i]$ does not jump at c_0 but rather is simply nonlinear?
- Then your linear model will identify a treatment effect when there isn't because the functional form had poor predictive properties beyond the cutoff
- ~~Let's look at a simulation~~

Importance of functional forms



Sharp RDD: Nonlinear Case

- Suppose the nonlinear relationship is $E[Y_i^0|X_i] = f(X_i)$ for some reasonably smooth function $f(X_i)$ (drumroll – like a cubic!)
- In that case we'd fit the regression model:

$$Y_i = f(X_i) + \delta D_i + \eta_i$$

- Since $f(X_i)$ is counterfactual for values of $X_i > c_0$, how will we model the nonlinearity?
- There are 2 common ways of approximating $f(X_i)$

Nonlinearities

People until Gelman and Imbens 2018 favored “higher order polynomials” but this is problematic due to overfitting. Gelman and Imbens 2018 recommend at best a quadratic

- ① Use global and local regressions with $f(X_i)$ equalling a p^{th} order polynomial

$$Y_i = \alpha + \delta D_i + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \eta_i$$

- ② Or use some nonparametric kernel method which I'll cover later

Different polynomials on the 2 sides of the discontinuity

- We can generalize the function, $f(x_i)$, by allowing it to differ on both sides of the cutoff by including them both individually and interacting them with D_i .
- In that case we have:

$$E[Y_i^0|X_i] = \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \cdots + \beta_{0p}\tilde{X}_i^p$$

$$E[Y_i^1|X_i] = \alpha + \delta + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \cdots + \beta_{1p}\tilde{X}_i^p$$

where \tilde{X}_i is the centered running variable (i.e., $X_i - c_0$).

Lines to the left, lines to the right of the cutoff

- Re-centering at c_0 ensures that the treatment effect at $X_i = c_0$ is the coefficient on D_i in a regression model with interaction terms
- As Lee and Lemieux (2010) note, allowing different functions on both sides of the discontinuity should be the main results in an RDD paper

Robustness against what?

- Are you done now that you have your main results? No
- Your main results are only causal insofar as smoothness is a credible belief, and since smoothness isn't guaranteed by "the science" like an RCT, you have to build your case
- You must now scrutinize alternative hypotheses that are consistent with your main results through sensitivity checks, placebos and alternative approaches

Main Challenges

Classify your concern regarding smoothness violations into two categories:

- Manipulation on the running variable
- Endogeneity of the cutoff

Most robustness is aimed at building credibility around these,

Manipulation of your running variable score

- Treatment is not as good as randomly assigned around the cutoff, c_0 , when agents are able to manipulate their running variable scores. This happens when:
 - ① the assignment rule is known in advance
 - ② agents are interested in adjusting
 - ③ agents have time to adjust
 - ④ administrative quirks like nonrandom heaping along the running variable

Examples include re-taking an exam, self-reported income, certain types of non-random rounding.

- Since necessarily treatment assignment is no longer independent of potential outcomes, it's likely this implies smoothness has been violated

Test 1: Manipulation of the running variable

Manipulation of the running variable

Assume a desirable treatment, D , and an assignment rule $X \geq c_0$. If individuals sort into D by choosing X such that $X \geq c_0$, then we say individuals are manipulating the running variable.

Also can be called “sorting on the running variable” – same thing

McCrary Density Test

~~We would expect waiting room A to become crowded. In the RDD context, sorting on the running variable implies heaping on the “good side” of c_0 .~~

- McCrary (2008) suggests a formal test: under the null the density should be continuous at the cutoff point.
- Under the alternative hypothesis, the density should increase at the kink (where D is viewed as good)
 - ① Partition the assignment variable into bins and calculate frequencies (i.e., number of observations) in each bin
 - ② Treat those frequency counts as dependent variable in a local linear regression
- This is oftentimes visualized with confidence intervals illustrating the effect of the discontinuity on density - you need no jump to pass this test

McCrary density test

- The McCrary Density Test has become **mandatory** for every analysis using RDD.
 - If you can estimate the conditional expectations, you evidently have data on the running variable. So in principle you can always do a density test
 - You can download the (no longer supported) Stata ado package, DCdensity, to implement McCrary's density test (<http://eml.berkeley.edu/~jmccrary/DCdensity/>)
 - You can install `rdrobust` for Stata and R too, and it will implement the test

Caveats about McCrary Density Test

- For RDD to be useful, you already need to know something about the mechanism generating the assignment variable and how susceptible it could be to manipulation. Note the rationality of economic actors that this test is built on.
- A discontinuity in the density is “suspicious” – it *suggests* manipulation of X around the cutoff is probably going on. In principle one doesn’t need continuity.
- This is a high-powered test. You need a lot of observations at c_0 to distinguish a discontinuity in the density from noise.

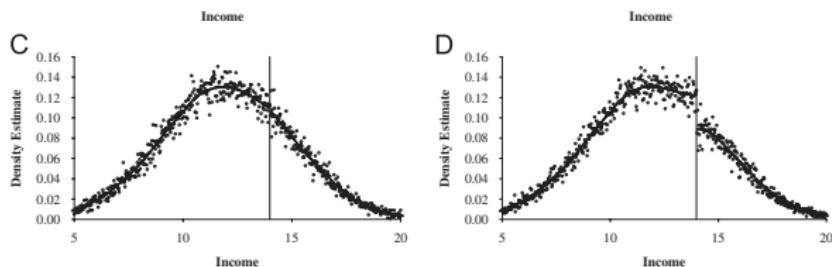


Figure: Panel C is density of income when there is no pre-announcement and no manipulation. Panel D is the density of income when there is pre-announcement and manipulation. From McCrary (2008).

Evaluating smoothness through balance

- Balance tests and placebo tests are related but distinct
- We can't directly test smoothness bc we are missing counterfactuals
- Ask yourself: why should average values of exogenous covariates jump if potential outcomes are smooth through the cutoff?
- If there are exogenous (non collider) covariates strongly associated with potential outcomes but exogenous to them, then they should be the same on either side of the cutoff if smoothness holds
- In this sense, balance tests are indirect searching for evidence supporting smoothness

Balance implementation

Don't make it hard – do what you did to Y , only to Z

- Choose other noncolliders associated with potential outcomes, Z
- Create similar graphical plots as you did for Y
- Could also conduct the parametric and nonparametric estimation on Z
- You do **not** want to see a jump around the cutoff, c_0

Placebos at non-discontinuous points

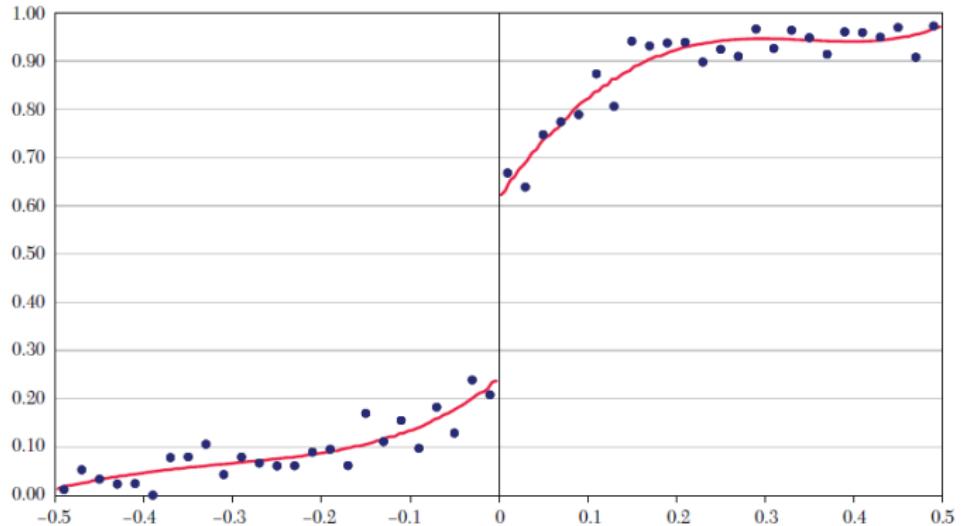
- Placebos in time are common with panels; placebo in running variables are their equivalent in RDD
- Imbens and Lemieux (2010) suggest we look at one side of the discontinuity (e.g., $X < c_0$), take the median value of the running variable in that section, and pretend it was a discontinuity, c'_0
- Then test whether in reality there is a discontinuity at c'_0 . You do **not** want to find anything.
- Remember though: smoothness at placebo points is neither necessary nor sufficient for smoothness in the potential outcomes at the cutoff
- So there are Type I and Type II risks of error with this

Outcomes

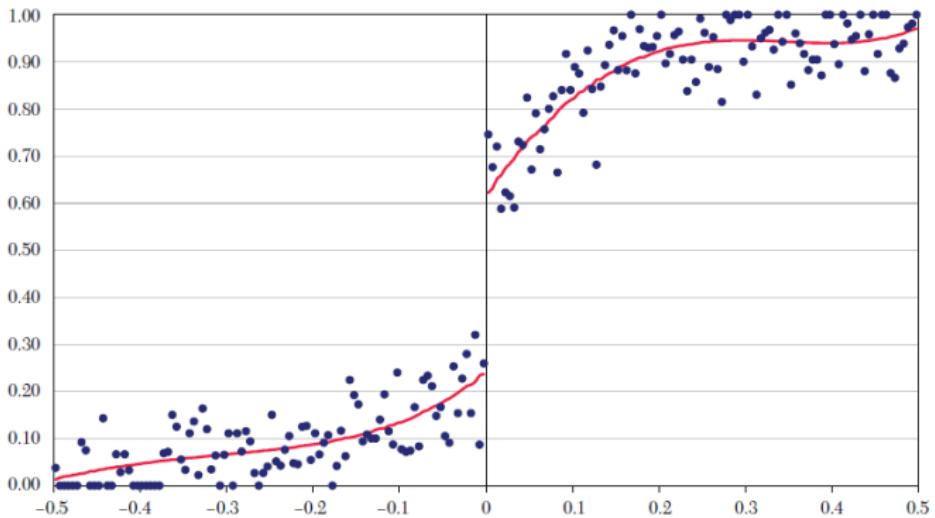
① Outcome by running variable, (X_i):

- Construct bins and average the outcome within bins on both sides of the cutoff
- Look at different bin sizes when constructing these graphs
- Plot the running variables, X_i , on the horizontal axis and the average of Y_i for each bin on the vertical axis
- Consider plotting a relatively flexible regression line on top of the bin means, but some readers prefer an eyeball test without the regression line to avoid “priming”

Example: Outcomes by Running Variables



Example: Outcomes by Running Variables with smaller bins



McCrary Density

③ Density of the running variable

- One should plot the number of observations in each bin.
- This plot allows to investigate whether there is a discontinuity or heaping in the distribution of the running variable at the threshold
- Heaping or discontinuities in the density suggest that people can manipulate their running variable score
- This is an indirect test of the identifying assumption that each individual has imprecise control over the assignment variable, which may violate smoothness

Density of the running variable

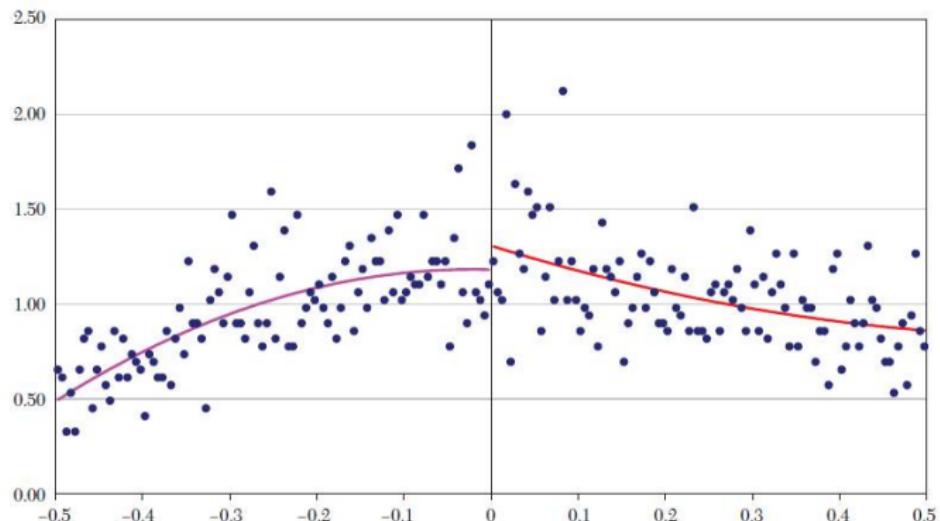


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Balance pictures

④ Covariates by a running variable

- Construct a similar graph to the outcomes graph but use a noncollider covariate as the “outcome”
- Balance implies smoothness through the cutoff, c_0 .
- If noncollider covariates jump at the cutoff, one is probably justified to reject that potential outcomes aren’t also probably jumping there

Example: Covariates by Running Variable

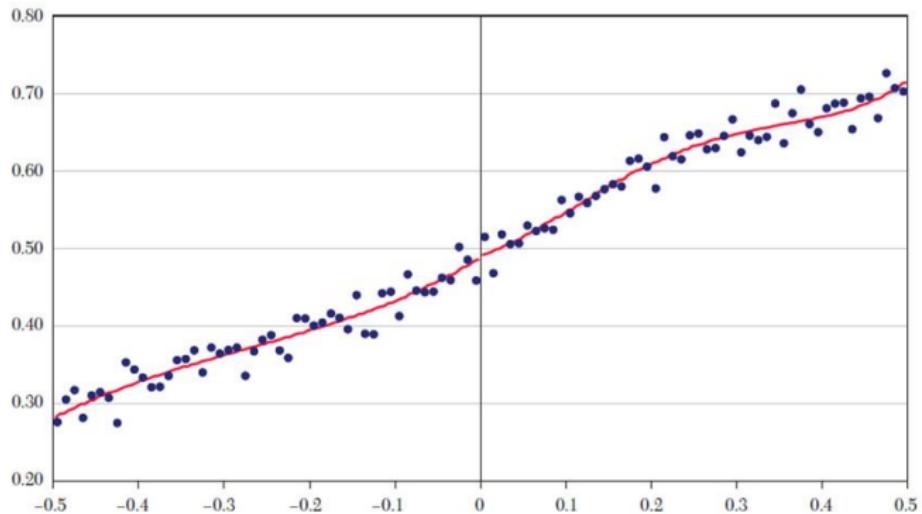
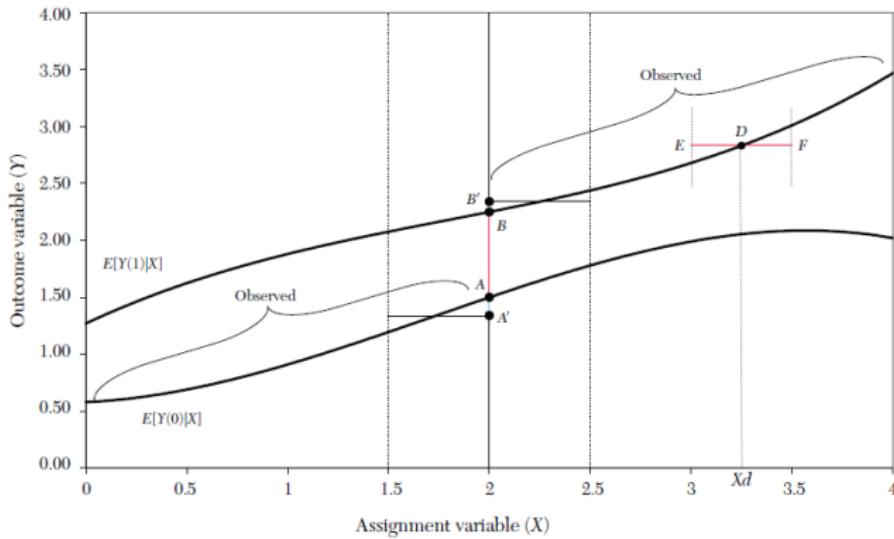


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

Parametric vs. nonparametric approaches

- Least squares approaches, because it models the counterfactual using functional forms, is parametric
- As a result, it can have poor predictive properties on counterfactuals above/below the cutoff
- Another way of approximating $f(X_i)$ is to use a nonparametric kernel which has its own problems; just not that one.

Kernel regression



- While the “true” effect is AB , with a certain bandwidth a rectangular kernel would estimate the effect as $A'B'$
- There is therefore systematic bias with the kernel method if the $f(X)$ is upwards or downwards sloping

Bandwidths

- Several methods for choosing the optimal bandwidth (window), but it's always a trade off between bias and variance
- In practical applications, you want to check for balance around that window
- Standard error of the treatment effects can be bootstrapped but there are also other alternatives
- You could add other variables to nonparametric methods.

DO VOTERS AFFECT OR ELECT POLICIES? EVIDENCE FROM THE U. S. HOUSE*

DAVID S. LEE
ENRICO MORETTI
MATTHEW J. BUTLER

There are two fundamentally different views of the role of elections in policy formation. In one view, voters can *affect* candidates' policy choices: competition for votes induces politicians to move toward the center. In this view, elections have the effect of bringing about some degree of policy compromise. In the alternative view, voters merely *elect* policies: politicians cannot make credible promises to moderate their policies, and elections are merely a means to decide which one of two opposing policy views will be implemented. We assess which of these contrasting perspectives is more empirically relevant for the U. S. House. Focusing on elections decided by a narrow margin allows us to generate quasi-experimental estimates of the impact of a "randomized" change in electoral strength on subsequent representatives' roll-call voting records. We find that voters merely *elect* policies: the degree of electoral strength has no effect on a legislator's voting behavior. For example, a large *exogenous* increase in electoral strength for the Democratic party in a district does not result in shifting both parties' nominees to the left. Politicians' inability to credibly commit to a compromise appears to dominate any competition-induced convergence in policy.

Public choice

There are two fundamentally different views of the role of voters in a representative democracy.

- ① **Convergence:** Voters force candidates to become relatively moderate depending on their size in the distribution (Downs 1957).

“Competition for votes can force even the most partisan Republicans and Democrats to moderate their policy choices. In the extreme case, competition may be so strong that it leads to ‘full policy convergence’: opposing parties are forced to adopt identical policies” – Lee, Moretti, and Butler 2004.

- ② **Divergence:** Voters pick the official and after taking office, she pursues her most-preferred policy.

Falsification of either hypothesis had been hard

- Very difficult to test either one of these since you don't observe the counterfactual votes of the loser for the same district/time
- Winners in a district are selected based on their policy's conforming to unobserved voter preferences, too
- Lee, Moretti and Butler (2004) develop the "close election RDD" which has the aim of determining whether convergence, while theoretically appealing, has any explanatory power in Congress
- The metaphor of the RCT is useful here: maybe close elections are being determined by coin flips (e.g., a few votes here, a few votes there)

Outcome is Congress person's liberal voting score

- **Liberal voting score** is a report card from the Americans for Democratic Action (ADA) for the House election results 1946-1995
 - Authors use the ADA score for all US House Representatives from 1946 to 1995 as their voting record index
 - For each Congress, ADA chooses about twenty high-profile roll-call votes and creates an index varying 0 and 100 for each Representative of the House measuring liberal voting record

Democratic “voteshare” is the running variable

- **Voteshare** from the same races
 - The running variable is voteshare which is the share of all votes that went to a Democrat.
 - They use a close Democratic victory to check whether convergence or divergence is correct (what's smoothness here?)
 - Discontinuity in the running variable occurs at $\text{voteshare} = 0.5$. When $\text{voteshare} > 0.5$, the Democratic candidate wins.
- I'll show `lmb1.do` to `lmb10.do` (and R) at times just so we can all see the simple estimation methods ourselves.

Nonparametric estimation

- Hahn, Todd and Van der Klaauw (2001) emphasized using local polynomial regressions
- Estimate $E[Y|X]$ in such a way that doesn't require committing to a functional form
- That model would be something general like

$$Y = f(X) + \varepsilon$$

Contemporaneous liberal voting score

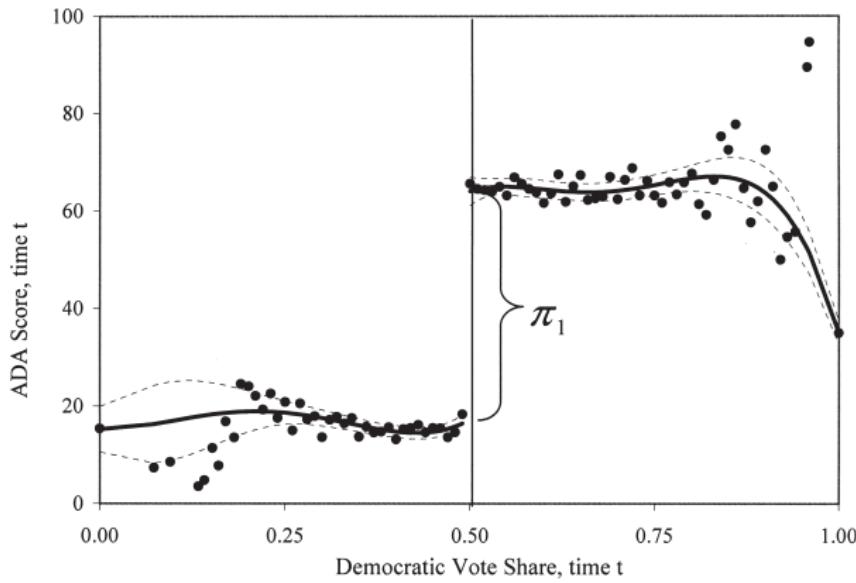


FIGURE IIa
Effect of Party Affiliation: π_1

Figure: Lee, Moretti, and Butler 2004, Figure IIa. $\pi_1 \approx 45$

Future liberal voting score

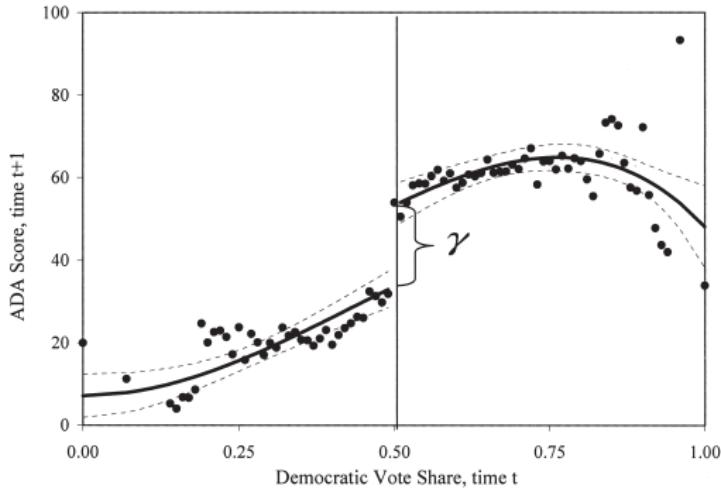


FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t . Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Affect"}} + \underbrace{\pi_1(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Elect"}}$$

Figure: Lee, Moretti, and Butler 2004, Figure I. $\gamma \approx 20$

Inc incumbency advantage

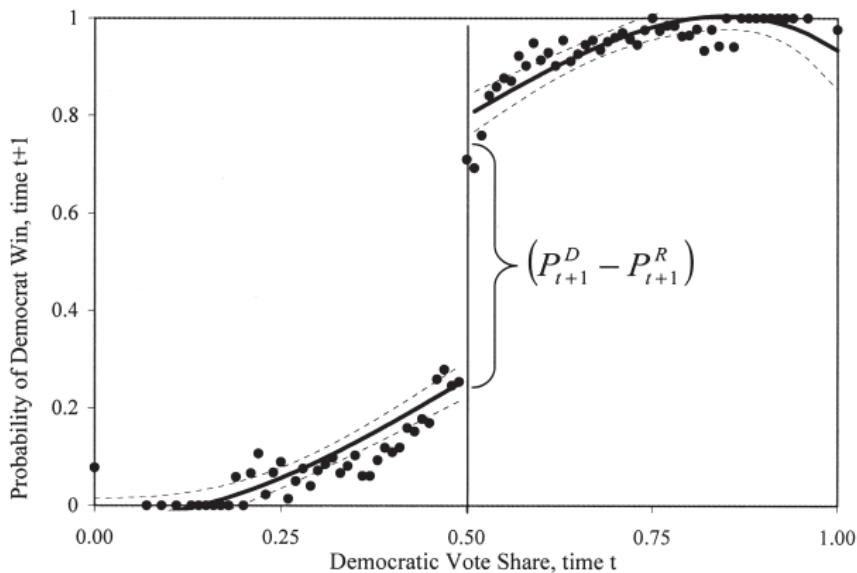


FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P_{t+1}^D - P_{t+1}^R)$

Figure: Lee, Moretti, and Butler 2004, Figure IIb. $(P_{t+1}^D - P_{t+1}^R) \approx 0.50$

Concluding remarks

- Caughey and Sekhon (2011) questioned the finding (not the design per se) saying that bare winners and bare losers in the US House elections differed considerably on pretreatment covariates (imbalance), which got worse in the closest elections
- Eggers, et al. (2014) evaluated 40,000 close elections including the House in other time periods, mayor races, and other types of US races including nine other countries
- They couldn't find another instance where Caughey and Sekhon's critique applied
- Assumptions behind close election design therefore probably holds and is one of the best RD designs we have

Instrumental Variables I

Raphael Corbi

Universidade de São Paulo

June 2021

Instrumental variables

- If treatment is tied to an unobservable, then conditioning strategies, even RDD, are invalid
- Instrumental variables offers some hope at recovering the causal effect of D on Y
- The best instruments come from deep knowledge of institutional details (Angrist and Krueger 1991)
- Certain types of natural experiments can be the source of such opportunities and may be useful

When is IV used?

Instrumental variables methods are typically used to address the following kinds of problems encountered in naive regressions

- ① Omitted variable bias
- ② Measurement error
- ③ Simultaneity bias
- ④ Reverse causality
- ⑤ Randomized control trials with noncompliance

Elasticity of demand is unidentified

- James Stock notes that his publications had a theme regarding identification
- He knew, for instance, that he couldn't simply look at correlations between price and quantity if he wanted the elasticity of demand due to simultaneous shifts in supply and demand
- The pairs of quantity and price weren't demand, or supply – they were demand and supply equilibrium values and therefore didn't reflect the demand or the supply curve, both of which are counterfactuals
- Those points are nothing more than a bunch of numbers – no more, no less – that have no practical use, scientific or otherwise

Exhibit 1

The Graphical Demonstration of the Identification Problem in Appendix B (p. 296)

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.

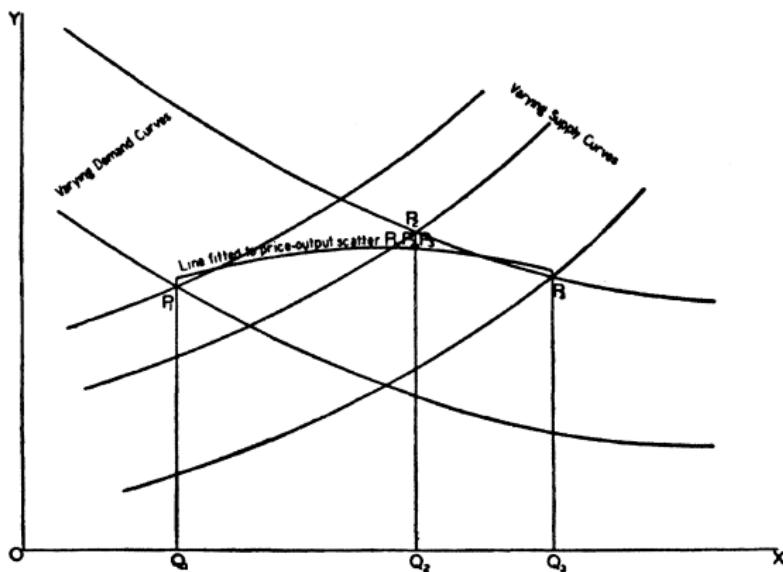


Figure: Wright's graphical demonstration of the identification problem

Constant treatment effects

- Constant treatment effects (i.e., β is constant across all individual units)
 - Constant treatment effects is the traditional econometric pedagogy when first learning instrumental variables, and doesn't need the potential outcomes model or notation to get the point across
 - Constant treatment effects is identical to assuming that $ATE=ATT=ATU$ because constant treatment effects assumes $\beta_i = \beta_{-i} = \beta$ for all units

Heterogenous treatment effects

- Heterogeneous treatment effects (i.e., β_i varies across individual units)
 - Heterogeneous treatment effects means that the $ATE \neq ATT \neq ATU$ because β_i differs across the population
 - This is equivalent to assuming the coefficient, β_i , is a random variable that varies across the population
 - Heterogenous treatment effects is based on work by Angrist, Imbens and Rubin (1996) and Imbens and Angrist (1994) which introduced the “local average treatment effect” (LATE) concept

Data requirements

- Your data isn't going to come with a codebook saying "instrumental variable". So how do you find it?
- Well, sometimes the researcher just *knows*.
- That is, the researcher knows of a variable (Z) that actually *is* randomly assigned and that affects the endogenous variable but not the outcome (except via the endogenous variable)
- Such a variable is called an "instrument".

Picking a good instrument

- The best instruments you think of first, then you seek the data second (but often students go in the reverse order which is basically guaranteed to be a crappy instrument)
- If you want to use IV, then ask:

What moves around the covariate of interest that might be plausibly random?

- Is there any element in the treatment that could be construed as random?
- If you were to find that random piece, then you have found an instrument
- Once you have identified such a variable, begin to think about what data sets might have information on an outcome of interest, the treatment, and the instrument you have put your finger on.

Does family size reduce labor supply or is it selection?

Angrist and Evans (1998), "Children and their parents' labor supply" *American Economic Review*,

- They want to know the effect of family size on labor supply, but need exogenous changes in family size
- So what if I told you if the first two children born were of the same gender, then you're less likely to work. What?!

Angrist and Evans cont.

- Many parents have a preference for having at least one child of each gender
 - Consider a couple whose first two kids were both boys; they will often have a third, hoping to have a girl
 - Consider a couple whose first two kids were girls; they will often have a third, hoping for a boy
 - Consider a couple with one boy and one girl; they will often not have a third kid
- The gender of your kids is arguably randomly assigned (maybe not exactly, but close enough)

Our causal model: Returns to schooling again

$$Y = \alpha + \delta S + \gamma A + \nu$$

where Y is log earnings, S is years of schooling, A is unobserved ability, and ν is the error term

- Suppose there exists a variable, Z_i , that is correlated with S_i .
- We can estimate δ with this variable, Z :

Before we use Z...

Since ability A_i is unobserved, we could run the following shorter equation instead...

$$Y_i = \alpha + \delta S + \eta_i$$

where η_i is a composite error term equalling $\gamma A_i + \epsilon_i$.

We assume that schooling is correlated with ability, so therefore it is correlated with η_i , making it endogenous in the second, shorter regression. Only ϵ_i is uncorrelated with the regressors, and that is by definition.

Omitted variable bias

We know from the derivation of the least squares operator that the estimated value of $\hat{\delta}$ is:

$$\hat{\delta} = \frac{C(Y, S)}{V(S)} = \frac{E[YS] - E[Y]E[S]}{V(S)}$$

Plugging in the true value of Y (from the longer model), we get the following:

$$\begin{aligned}\hat{\delta} &= \frac{E[\alpha S + S^2\delta + \gamma SA + \varepsilon S] - E(S)E[\alpha + \delta S + \gamma A + \varepsilon]}{V(S)} \\ &= \frac{\delta E(S^2) - \delta E(S)^2 + \gamma E(AS) - \gamma E(S)E(A) + E(\varepsilon S) - E(S)E(\varepsilon)}{V(S)} \\ &= \delta + \gamma \frac{C(AS)}{V(S)}\end{aligned}$$

If $\gamma > 0$ and $C(A, S) > 0$, then $\hat{\delta}$, the coefficient on schooling, is upward biased. And that is probably the case given that it's likely that ability and schooling are positively correlated.

How can IV be used to obtain consistent estimates?

$$\begin{aligned}\text{Cov}(Y, Z) &= \text{Cov}(\alpha + \delta S + \gamma A + \nu, Z) \\&= E[(\alpha + \delta S + \gamma A + \nu)Z] - E[\alpha + \delta S + \gamma A + \nu]E[Z] \\&= \{\alpha E(Z) - \alpha E(Z)\} + \delta\{E(SZ) - E(S)E(Z)\} \\&\quad + \gamma\{E(AZ) - E(A)E(Z)\} + E(\nu Z) - E(\nu)E(Z) \\ \text{Cov}(Y, Z) &= \delta \text{Cov}(S, Z) + \gamma \text{Cov}(A, Z) + \text{Cov}(\nu, Z)\end{aligned}$$

Divide both sides by $\text{Cov}(S, Z)$ and the first term becomes δ , the LHS becomes the ratio of the reduced form to the first stage, plus two other scaled terms.

Consistency

- What conditions must hold for a valid IV design?
 - $\text{Cov}(S, Z) \neq 0$ – “first stage” exists. S and Z are correlated
 - $\text{Cov}(A, Z) = \text{Cov}(\nu, Z) = 0$ – “exclusion restriction”. This means Z that orthogonal to the factors in ν , such as unobserved ability, A , as well as the structural disturbance term, ν
- Assuming the first stage exists and that the exclusion restriction holds, then we can estimate δ with δ_{IV} :

$$\begin{aligned}\delta_{IV} &= \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)} \\ &= \delta\end{aligned}$$

IV is Consistent if IV Assumptions are Satisfied

- The IV estimator is consistent if the IV assumptions are satisfied. Substitute true model for Y :

$$\begin{aligned}\delta_{IV} &= \frac{\text{Cov}([\alpha + \rho S + \gamma A + \nu], Z)}{\text{Cov}(S, Z)} \\ &= \delta \frac{\text{Cov}([S], Z)}{\text{Cov}(S, Z)} + \gamma \frac{\text{Cov}([A], Z)}{\text{Cov}(S, Z)} + \frac{\text{Cov}([\nu], Z)}{\text{Cov}(S, Z)} \\ &= \delta + \gamma \frac{\text{Cov}(\eta, Z)}{\text{Cov}(S, Z)}\end{aligned}$$

Identifying assumptions and consistency

- Taking the probability limit which is an asymptotic operation to show consistency:

$$\begin{aligned}\text{plim } \widehat{\delta}_{IV} &= \text{plim } \delta + \gamma \frac{\text{Cov}(\eta, Z)}{\text{Cov}(S, Z)} \\ &= \delta\end{aligned}$$

because $\text{Cov}([A], Z) = 0$ and $\text{Cov}([\nu], Z) = 0$ due to the exclusion restriction, and $\text{Cov}(S, Z) \neq 0$ (due to the first stage)

IV Assumptions

- But, if Z is *not* independent of η (either correlated with A or ν), *and* if the correlation between S and Z is “weak”, then the second term blows up.
- We will explore the problems created by weak instruments in just a moment.
- ~~First, let's look at a DAG summarizing all this information~~

Two-stage least squares

- The two-stage least squares estimator was developed by Theil (1953) and Basman (1957) independently
- Note, while IV is a research design, 2SLS is a specific estimator.
- Others include LIML, the Wald estimator, jacknife IV, two sample IV, and more

Two-stage least squares concepts

- Causal model. Sometimes called the structural model:

$$Y_i = \alpha + \delta S_i + \eta_i$$

- First-stage regression. Gets the name because of two-stage least squares:

$$S_i = \gamma + \rho Z_i + \zeta_i$$

- Second-stage regression. Notice the fitted values, \hat{S} :

$$Y_i = \beta + \delta \hat{S}_i + \nu_i$$

Reduced form

- Some people like a simpler approach because they don't want to defend IV's assumptions
- Reduced form a regression of Y onto the instrument:

$$Y_i = \psi + \pi Z_i + \varepsilon_i$$

- ~~This would be like regressing hell onto Sunday Candy, as opposed to regressing hell onto church with Sunday Candy instrumenting for church~~

Two-stage least squares

Suppose you have a sample of data on Y , X , and Z . For each observation i we assume the data are generated according to

$$Y_i = \alpha + \delta S_i + \eta_i$$

$$S_i = \gamma + \rho Z_i + \zeta_i$$

where $\text{Cov}(Z, \eta_i) = 0$ and $\rho \neq 0$.

Two-stage least squares

Plug in covariance and write out the following:

$$\begin{aligned}\widehat{\delta_{2sls}} &= \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, S)} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(S_i - \bar{S})} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})S_i}\end{aligned}$$

Two-stage least squares

Substitute the causal model definition of Y to get:

$$\begin{aligned}\widehat{\delta_{2sls}} &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}) \{\alpha + \delta S_i + \eta_i\}}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}) S_i} \\ &= \delta + \frac{\frac{1}{n} (Z_i - \bar{Z}) \eta_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}) S_i} \\ &= \delta + \text{"small if } n \text{ is large"}$$

Where did the first term go? Why did the second term become δ ?

Two-stage least squares

- Calculate the ratio of “reduced form” (π) to “first stage” coefficient (ρ):

$$\hat{\delta}_{2sls} = \frac{Cov(Z, Y)}{Cov(Z, S)} = \frac{\frac{Cov(Z, Y)}{Var(Z)}}{\frac{Cov(Z, S)}{Var(Z)}} = \frac{\hat{\pi}}{\hat{\rho}}$$

- Rewrite $\hat{\rho}$ as

$$\begin{aligned}\hat{\rho} &= \frac{Cov(Z, S)}{Var(Z)} \\ \hat{\rho} Var(Z) &= Cov(Z, S)\end{aligned}$$

Intuition of 2SLS

- Two stage least squares is nice because in addition to being an estimator, there's also great intuition contained in it which you can use as a device for thinking about IV more generally.
- The intuition is that 2SLS estimator replaces S with the fitted values of S (i.e., \widehat{S}) from the first stage regression of S onto Z and all other covariates.
- By using the fitted values of the endogenous regressor from the first stage regression, our regression now uses *only* the exogenous variation in the regressor due to the instrumental variable itself

Intuition of IV in 2SLS

- ... but think about it – that variation was there before, but was just a subset of all the variation in the regressor
- Go back to what we said in the beginning - we need the endogenous variable to have pieces that are random, and IV finds them.
- Instrumental variables therefore reduces the variation in the data, but that variation which is left is *exogenous*
- ~~"With a long enough [instrument], you can [estimate any causal effect]" – Scott Cunningham paraphrasing Archimedes~~

Estimation with software

- One manual way is just to estimate the reduced form and first stage coefficients and take the ratio of the respective coefficients on Z
- But while it is always a good idea to run these two regressions, don't compute your IV estimate this way

Estimation with software

- It is often the case that a pattern of missing data will differ between Y and S
- In such a case, the usual procedure of “casewise deletion” is to keep the subsample with non-missing data on Y , S , and Z .
- But the reduced form and first stage regressions would be estimated off of different sub-samples if you used the two step method before
- The standard errors from the second stage regression are also wrong

Estimation with software

- Estimate this in Stata using -ivregress 2sls-.
- Estimate this in R -ivreg()- which is in the AER package

Summing up...

Let our true model of returns to schooling be

$$Y_i = \alpha + \delta S + \gamma A_i + \epsilon_i$$

where Y is log earnings, S is years of schooling, A is unobserved ability, and ϵ is the error term. Since ability A_i is unobserved, we could run the following shorter equation instead...

$$Y_i = \alpha + \delta S + \eta_i$$

where η_i is a composite error term equalling $\gamma A_i + \epsilon_i$.

Schooling is correlated with ability, so therefore it is correlated with η_i , making it endogenous in the second, shorter regression. Only ϵ_i is uncorrelated with the regressors, and that is by definition.

Summing up...

Estimating δ by simple OLS will lead to wrong estimates due to the omitted variable bias.

$$\delta^{OLS} = \delta + \gamma \frac{cov(A, S)}{var(S)}$$

We can use a variable Z as instrument in order to consistently estimate δ .
Two conditions must hold for a valid IV design:

1. $cov(S, Z) \neq 0$ “first-stage”
2. $cov(A, Z) = cov(\epsilon, Z) = 0$ “exclusion restriction”

Illustrative example

BY running a simple simulation, we can illustrate the two conditions needed for IV consistency and alternative ways to calculate δ^{IV} .

1. $\delta^{IV} = \text{cov}(Y, Z)/\text{cov}(S, Z) \neq \delta^{OLS} = \text{cov}(Y, S)/\text{cov}(S, S)$
2. estimate first stage, predict \hat{Y} and run second stage
3. ratio of RF by FS coefficient, or $\delta^{IV} = \frac{\frac{\text{cov}(Y, Z)}{\text{var}(Z)}}{\frac{\text{cov}(S, Z)}{\text{var}(Z)}} = \frac{\hat{\pi}^{RF}}{\hat{\rho}^{FS}}$
4. using command **ivreg** (which also gives us the correct standard deviation)

Suppose there is an scholarship program that incentivizes students to continue school. Eligibility is defined by a lottery.

Can Z be used as an instrument?

Weak instruments

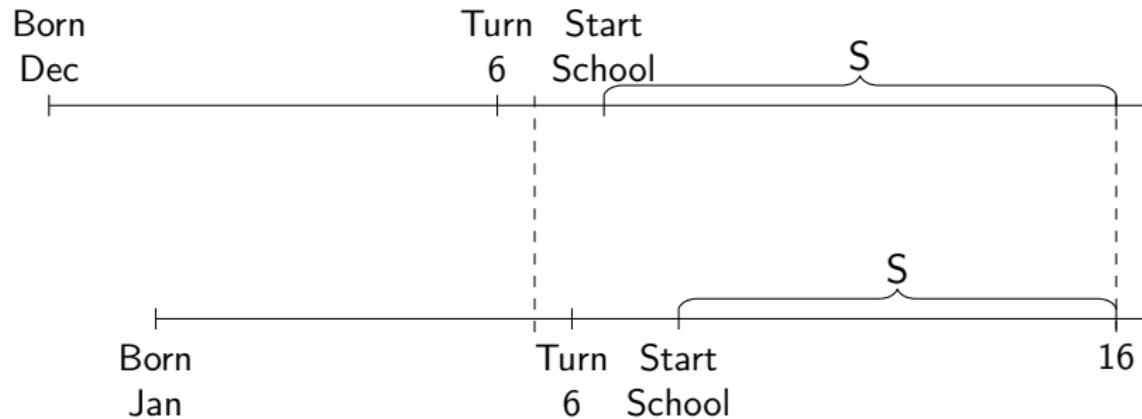
- A weak instrument is one that is not strongly correlated with the endogenous variable in the first stage
- This can happen if the two variables are independent or the sample is small
- If you have a weak instrument, then the bias of 2SLS is centered on the bias of OLS and the cure ends up being worse than the disease
- We knew this was a problem, but it was brought into sharp focus with Angrist and Krueger (1991) and some papers that followed

Angrist and Krueger (1991)

- In practice, it is often difficult to find convincing instruments – usually because potential instruments don't satisfy the exclusion restriction
- But in an early paper in the causal inference movement, Angrist and Krueger (1991) wrote a very interesting and influential study instrumental variable
- They were interested in schooling's effect on earnings and instrumented for it with *which quarter of the year you were born*
- Remember Chance quote - what the heck would birth quarter have to do with earnings such that it was an excludable instrument?

Compulsory schooling

- In the US, you could drop out of school once you turned 16
- "School districts typically require a student to have turned age six by January 1 of the year in which he or she enters school"
(Angrist and Krueger 1991, p. 980)
- Children have different ages when they start school, though, and this creates different lengths of schooling at the time they turn 16 (potential drop out age):



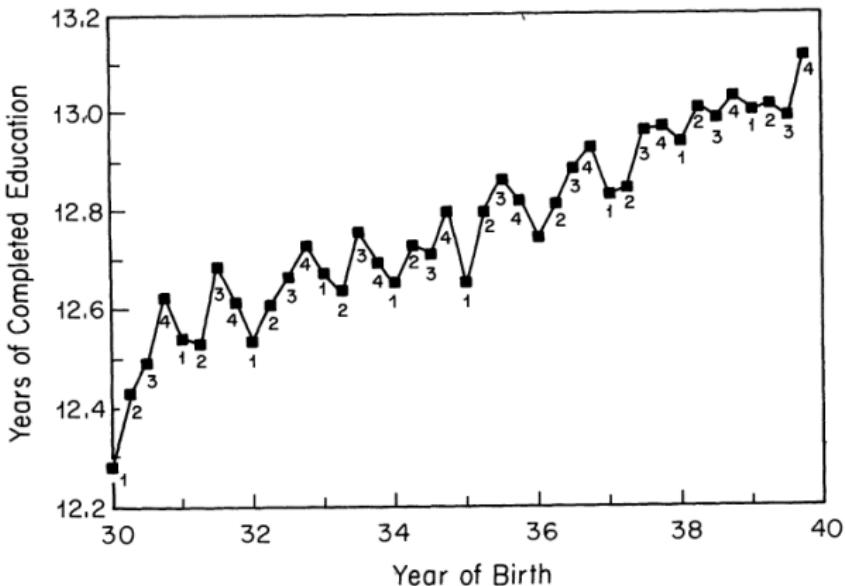
If you're born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

Visuals

- You need good data visualization for IV partly because of the scrutiny around the design
- The two pieces you should be ready to build pictures for are the first stage and the reduced form
- Angrist and Krueger (1991) provide simple, classic and compelling pictures of both

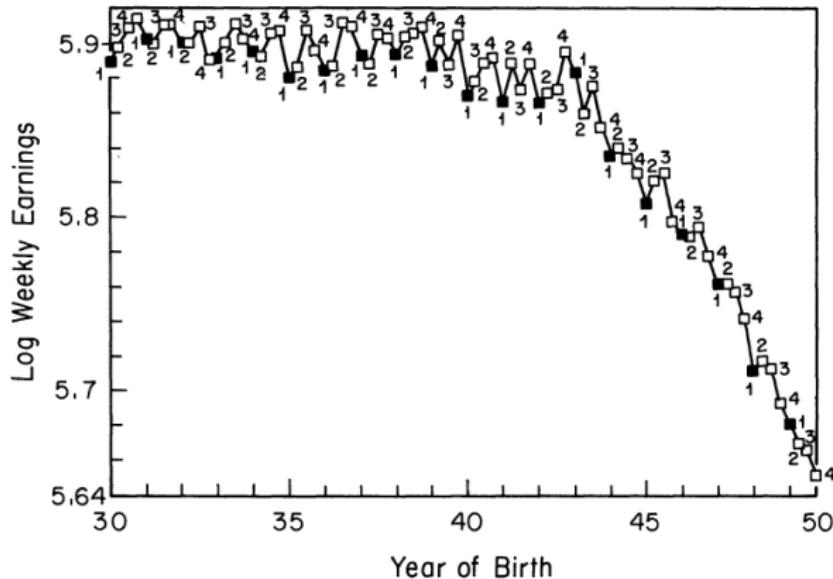
First Stage

Men born earlier in the year have lower schooling. This indicates that there is a first stage. Notice all the 3s and 4s at the top. But then notice how it attenuates over time . . .



Reduced Form

Do differences in schooling due to different quarter of birth translate into different earnings?



Two Stage Least Squares model

- The causal model is

$$Y_i = \delta S_i + \varepsilon$$

- The first stage regression is:

$$S_i = X\pi_{10} + \pi_{11}Z_i + \eta_{1i}$$

- The reduced form regression is:

$$Y_i = X\pi_{20} + \pi_{21}Z_i + \eta_{2i}$$

- The covariate adjusted IV estimator is the sample analog of the ratio, $\frac{\pi_{21}}{\pi_{11}}$

Two Stage Least Squares

- Angrist and Krueger instrument for schooling using three quarter of birth dummies: a dummies for 2nd, 3rd and 4th qob
- Their estimated first-stage regression is:

$$S_i = X\pi_{10} + Z_{1i}\pi_{11} + Z_{2i}\pi_{12} + Z_{3i}\pi_{13} + \eta_1$$

- The second stage is the same as before, but the fitted values are from the new first stage

First stage regression results

Quarter of birth is a strong predictor of total years of education

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect ^a			F-test ^b [P-value]
			I	II	III	
Total years of education	1930–1939	12.79	-0.124 (0.017)	-0.086 (0.017)	-0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	-0.085 (0.012)	-0.035 (0.012)	-0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	-0.019 (0.002)	-0.020 (0.002)	-0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	-0.015 (0.001)	-0.012 (0.001)	-0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	-0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	-0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	-0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	-0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

First stage regression results: Placebos

Completed master's degree	1930–1939	0.09	−0.001	0.002	−0.001	1.7	
			(0.001)	(0.001)	(0.001)	[0.1599]	
Completed doctoral degree	1940–1949	0.11	0.000	0.004	0.001	3.9	
			(0.001)	(0.001)	(0.001)	[0.0091]	
	1930–1939	0.03	0.002	0.003	0.000	2.9	
			(0.001)	(0.001)	(0.001)	[0.0332]	
	1940–1949	0.04	−0.002	0.001	−0.001	4.3	
			(0.001)	(0.001)	(0.001)	[0.0050]	

a. Standard errors are in parentheses. An $MA(+2, -2)$ trend term was subtracted from each dependent variable. The data set contains men from the 1980 Census, 5 percent Public Use Sample. Sample size is 312,718 for 1930–1939 cohort and is 457,181 for 1940–1949 cohort.

b. F-statistic is for a test of the hypothesis that the quarter-of-birth dummies jointly have no effect.

IV Estimates Birth Cohorts 20-29, 1980 Census

Independent variable	(1) OLS	(2) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)
Race (1 = black)	—	—
SMSA (1 = center city)	—	—
Married (1 = married)	—	—
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No
Age	—	—
Age-squared	—	—
χ^2 [dof]	—	25.4 [29]

Mechanism

- In addition to log weekly wage, they examined the impact of compulsory schooling on log annual salary and weeks worked
- The main impact of compulsory schooling is on the log weekly wage – not on weeks worked

More instruments

To incorporate the cross-state seasonal variation in education, we computed TSLS estimates that use as instruments for education a set of three quarter-of-birth dummies interacted with fifty state-of-birth dummies, in addition to three quarter-of-birth dummies interacted with nine year-of-birth dummies.¹⁸ The estimates also include fifty state-of-birth dummies in the wage equation, so the variability in education used to identify the return to education in the TSLS estimates is solely due to differences by season of birth. Unlike the previous TSLS estimates, the seasonal differences are now allowed to vary by state as well as by birth year.

Problem enters with many quarter of birth interactions

- They want to increase the precision of their 2SLS estimates, so they load up their first stage with more instruments
- Specifications with 30 (quarter of birth \times year) dummy variables and 150 (quarter of birth \times state) instruments
 - What's the intuition here? The effect of quarter of birth may vary by birth year or by state
- It reduced the standard errors, but that comes at a cost of potentially having a weak instruments problem

Weak Instruments

- For a long time, applied empiricists were not attentive to the small sample bias of IV
- But in the early 1990s, a number of papers highlighted that IV can be *severely* biased – in particular, when instruments have only a weak correlation with the endogenous variable of interest and when many instruments are used to instrument for one endogenous variable (i.e., there are many overidentifying restrictions).
- In the worst case, if the instruments are so weak that there is no first stage, then the 2SLS sampling distribution is centered on the probability limit of OLS

Weak Instruments

In particular, Bound, Jaeger, and Baker (1995) show that the bias of 2SLS centers on the previously defined OLS bias as the weakness of the instrument grows.

That bias can be written as a function of the first-stage F-statistic:

$$E[\hat{\delta}^{OLS} - \delta] = \frac{\text{cov}(S, \eta)}{\text{var}(S)} \frac{1}{F + 1}$$

where F is the population analogy of the F-statistic for the joint significance of the instruments in the first-stage regression. If the first stage is weak, and $F \rightarrow 0$, then the bias of 2SLS approaches $\frac{\text{cov}(S, \eta)}{\text{var}(S)}$. But if the first stage is very strong, $F \rightarrow \infty$, then the 2SLS bias goes to 0.

Popular IV Designs

Raphael Corbi

Universidade de São Paulo

June 2021

Popular IV Designs

IV is a general strategy to adopt when you have a good instrument.

Over the years, certain types of IV strategies have been used so many times that they constitute their own designs.

1. lottery design (randomized control)
2. judge fixed effect design

IV in Randomized Trials

- In many randomized trials, participation is nonetheless voluntary among those randomly assigned to treatment
- Consequently, noncompliance is not uncommon and without correcting for it, creates selection biases
- IV designs may even be helpful when evaluating a randomized trial, even though treatment was randomly assigned
- The solution is to instrument for treatment with whether you “won the lottery” and estimate LATE

Lottery designs

- The instrument is your randomized lottery
- Examples might be randomized lottery for attending charter schools to study effect of charter schools on educational outcomes, or a randomized voucher to encourage the collection of health information
- Recall Thornton (2008) instrumented for getting HIV results to estimate causal effect of learning one was HIV+ on condom purchases
- We'll discuss two papers from 2012 and 2014 evaluating a lottery-based expansion of Medicaid health insurance on Oregon on numerous health and financial outcomes

Overarching question

- What are the effects of expanding access to public health insurance for low income adults?
 - Magnitudes, and even the signs, associated with that question were uncertain
- Limited existing evidence
 - Institute of Medicine review of evidence was suggestive, but a lot of uncertainty
 - Observational studies are confounded by selection into health insurance
 - Quasi-experimental work often focuses on elderly and children
 - Only one randomized experiment in a developed country: the RAND health insurance experiment
 - 1970s experiment on a general population
 - Randomized cost-sharing, not coverage itself

The Oregon Health Insurance Experiment

Setting: Oregon Health Plan Standard

- Oregon's Medicaid expansion program for poor adults
- Eligibility
 - Poor (<100% federal poverty line) adults 19-64
 - Not eligible for other programs
 - Uninsured > 6 months
 - Legal residents
- Comprehensive coverage (no dental or vision)
- Minimum cost-sharing
- Similar to other states in payments, management
- Closed to new enrollment in 2004

The Oregon Medicaid Experiment

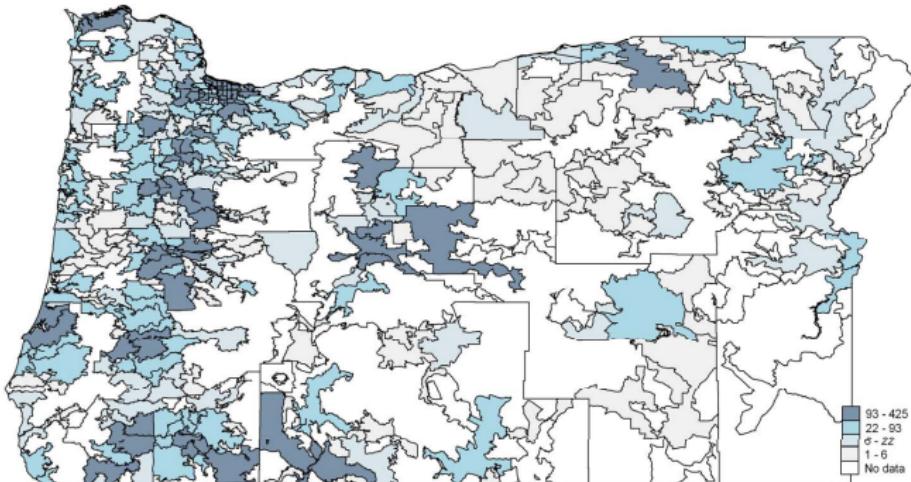
Oregon held a lottery

- Waiver to operate lottery
- 5-week sign-up period, heavy advertising (January to February 2008)
- Low barriers to sign up, no eligibility pre-screening
- Limited information on list
- Randomly drew 30,000 out of 85,000 on list (March–October 2008)
- Those selected given chance to apply
 - Treatment at household level
 - Had to return application within 45 days
 - 60% applied; 50% of those deemed eligible → 10,000 enrollees

Data

- Pre-randomization demographic information
 - From lottery sign-up
- State administrative records on Medicaid enrollment
 - Primary measure of first stage (i.e., insurance coverage)
- Outcomes
 - Administrative data (~16 months post-notification): Hospital discharge data, mortality, credit reports
 - Mail surveys (~15 months): some questions ask 6-month look-back; some ask current
 - In-person survey and measurements (~25 months): Detailed questionnaires, blood samples, blood pressure, body mass index

Lottery List Distribution Across Zip Codes



Empirical Framework

- They present reduced form estimates of the causal effect of lottery selection

$$Y_{ihj} = \beta_0 + \beta_1 LOTTERY_h + X_{ih}\beta_2 + V_{ih}\beta_3 + \varepsilon_{ihj}$$

- Validity of experimental design: randomization; balance on treatment and control. This is what readers expect

Empirical framework

- They also present IV results because they want to isolate the causal effect of insurance coverage

$$\begin{aligned} INSURANCE_{ihj} &= \delta_0 + \delta_1 LOTTERY_{ih} + X_{ih}\delta_2 + V_{ih}\delta_3 + \mu_{ihj} \\ y_{ihj} &= \pi_0 + \pi_1 \widehat{INSURANCE}_{ih} + X_{ih}\pi_2 + V_{ih}\pi_3 + v_{ihj} \end{aligned}$$

- Effect of lottery on coverage: about 25 percentage points
- We have independence guaranteed; now we need exclusion: the primary pathway of the lottery must be via being on Medicaid
 - Could affect participation in other programs, but actually small
 - "Warm glow" of winning – especially early
- Analysis plan, multiple inference adjustment

Effect of lottery on coverage (first stage)

	Full sample		Credit subsample		Survey respondents	
	Control mean	Estimated FS	Control mean	Estimated FS	Control mean	Estimated FS
Ever on Medicaid	0.141 (0.004)	0.256 (0.004)	0.135 (0.004)	0.255 (0.004)	0.135 (0.007)	0.290 (0.007)
Ever on OHP Standard	0.027 (0.003)	0.264 (0.003)	0.028 (0.004)	0.264 (0.004)	0.026 (0.005)	0.302 (0.005)
# of Months on Medicaid	1.408 (0.045)	3.355 (0.045)	1.352 (0.055)	3.366 (0.055)	1.509 (0.055)	3.943 -0.09
On Medicaid, end of study period	0.106 (0.003)	0.148 (0.003)	0.101 (0.004)	0.151 (0.004)	0.105 (0.006)	0.189 (0.006)
Currently have any insurance (self report)					0.325 (0.008)	0.179 (0.008)
Currently have private ins. (self report)					0.128 (0.005)	-0.008 (0.005)
Currently on Medicaid (self report)					0.117 (0.006)	0.197 (0.006)
Currently on Medicaid					0.093 (0.006)	0.177 (0.006)

Amy Finkelstein, et al. (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year", *Quarterly Journal of Economics*, vol. 127, issue 3, August.

Effects of Medicaid

Use primary and secondary data to gauge 1-year effects

- Mail surveys: 70,000 surveys at baseline, 12 months
- Administrative data
 - Medicaid enrollment records
 - Statewide Hospital discharge data, 2007-2010
 - Credit report data, 2007-2010
 - Mortality data, 2007-2010

Mail survey data

- **Fielding protocol**
 - ~70,000 people, surveyed at baseline and 12 months later
 - Basic protocol: three-stage male survey protocol, English/Spanish
 - Intensive protocol on a 30% subsample included additional tracking, mailings, phone attempts (done to adjust for non-response bias)
- **Response rate**
 - Effective response rate = 50%
 - Non-response bias always possible, but response rate and pre-randomization measures in administrative data were balanced between treatment and control

Administrative data

- Medicaid records
 - Pre-randomization demographics from list
 - Enrollment records to assess “first stage” (how many of the selected got insurance coverage)
- Hospital discharge data
 - Probabilistically matched to list, de-identified at Oregon Health Plan
 - Includes dates and source of admissions, diagnoses, procedures, length of stay, hospital identifier
 - Includes years before and after randomization
- Other data
 - Mortality data from Oregon death records
 - Credit report data, probabilistically matched, de-identified

Outcomes

- **Access and use of care**
 - Is access to care improved? Do the insured use more care? Is there a shift in the types of care being used?
 - Mail surveys and hospital discharge data
- **Financial strain**
 - How much does insurance protect against financial strain?
 - What are the out-of-pocket implications?
 - Mail surveys and credit reports
- **Health**
 - What are the short-term impacts on self-reported physical and mental health?
 - Mail surveys and vital statistics (mortality)

Effect of lottery on coverage

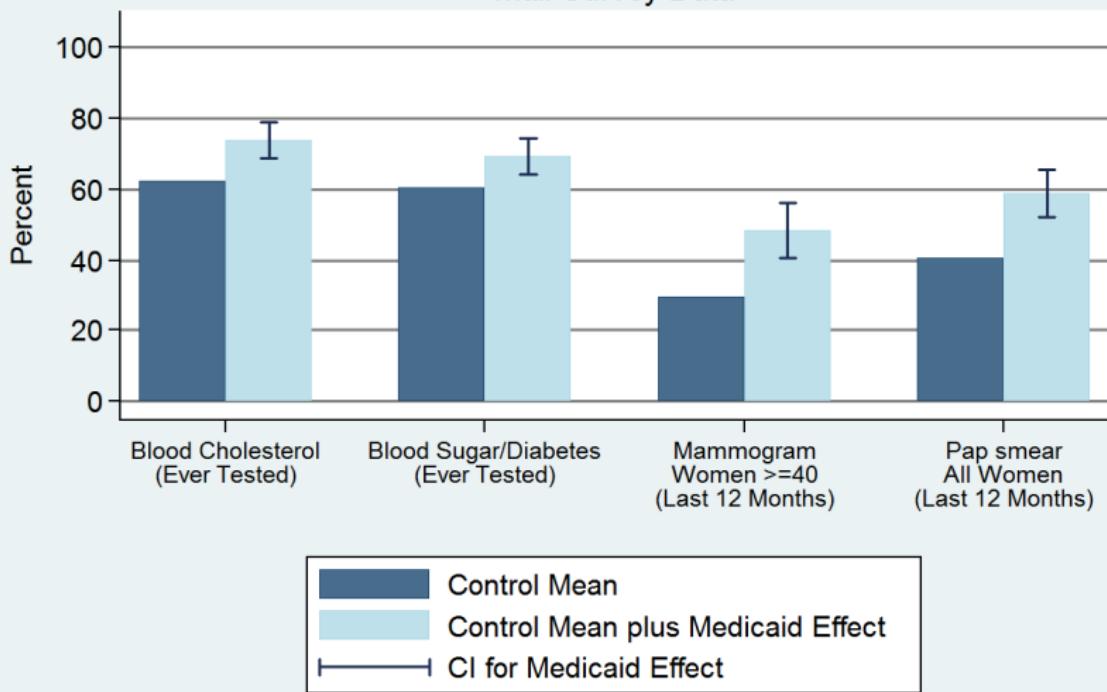
Gaining insurance resulted in better access to care and higher satisfaction with care (conditional on actually getting care)

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Have a usual place of care	49.9%	+9.9%	+33.9%	.0001
Have a personal doctor	49.0%	+8.1%	+28.0%	.0001
Got all needed health care	68.4%	+6.9%	+23.9%	.0001
Got all needed prescriptions	76.5%	+5.6%	+19.5%	.0001
Satisfied with quality of care	70.8%	+4.3%	+14.2%	.001

SOURCE: Survey data

Preventive Care

Mail Survey Data



Effect of lottery on coverage

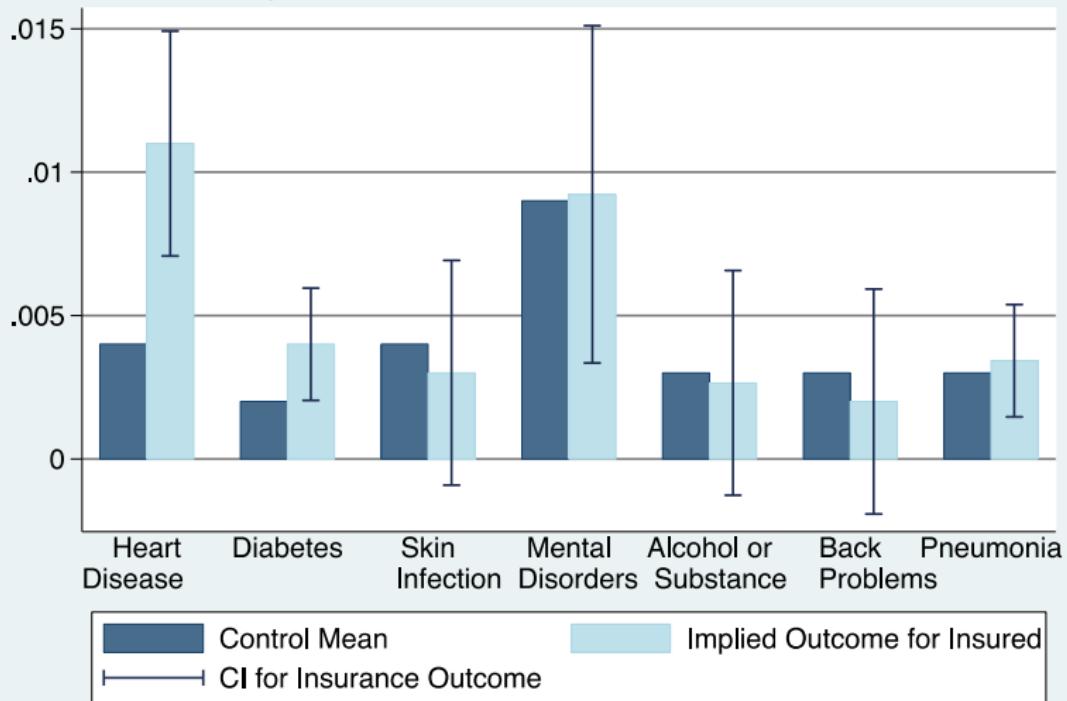
Gaining insurance resulted in increased probability of hospital admissions, primarily driven by non-emergency department admissions

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Any hospital admission	6.7%	+.50%	+2.1%	.004
--Admits through ED	4.8%	+.2%	+.7%	.265
--Admits NOT through ED	2.9%	+.4%	+1.6%	.002

SOURCE: Hospital Discharge Data

Overall, this represents a 30% higher probability of admission, although admissions are still rare events

Hospital Utilization for Selected Conditions



Summary: Access and use of care

- Overall, utilization and costs went up relative to controls
 - 30% increase in probability of an inpatient admission
 - 35% increase in probability of an outpatient visit
 - 15% increase in probability of taking prescription medications
 - Total \$777 increase in average spending (a 25% increase)
- With this increased spending, those who gained insurance were
 - 35% more likely to get all needed care
 - 25% more likely to get all needed medications
 - Far more likely to follow preventive care guidelines, such as mammograms (60%) and PAP tests (45%)

Results: Financial Strain

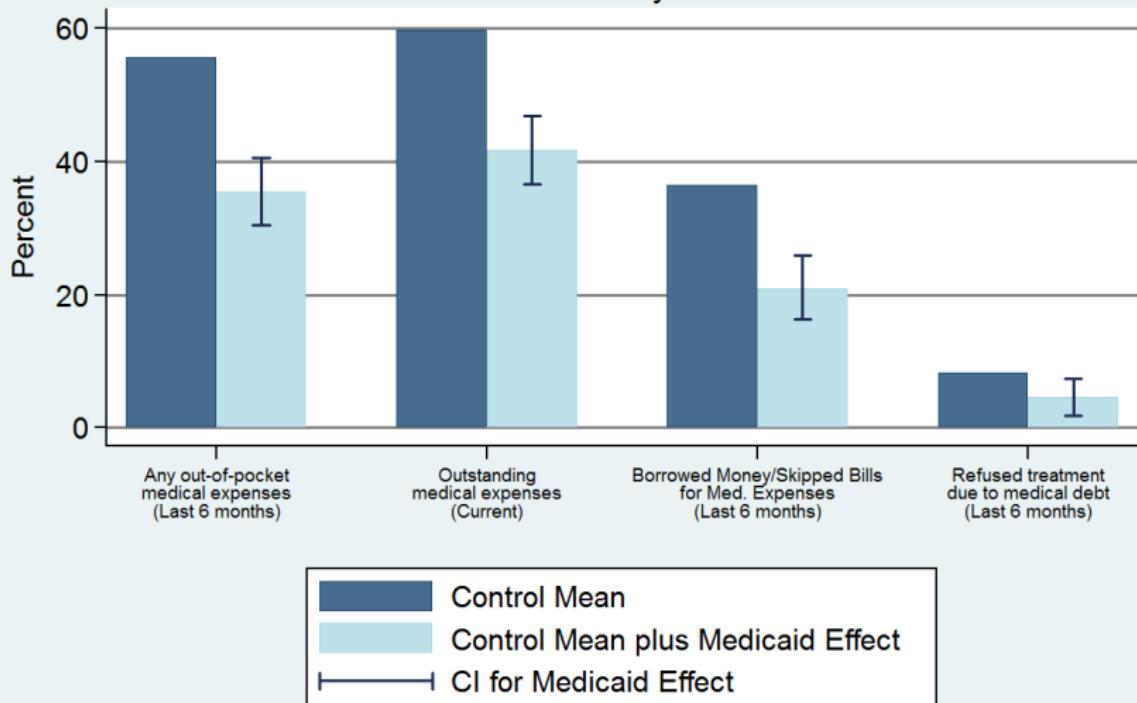
Gaining insurance resulted in a reduced probability of having medical collections in credit reports, and in lower amounts owed

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Had a bankruptcy	1.4%	+0.2%	+0.9%	.358
Had a collection	50.0%	-1.2%	-4.8%	.013
--Medical collections	28.1%	-1.6%	-6.4%	.0001
--Non-medical collections	39.2%	-0.5	-1.8%	.455
\$ owed medical collections	\$1,999	-\$99	-\$390	.025

Source: Credit report data

Self-reported Financial Strain

Mail Survey Data



Summary: Financial Strain

- Overall, reductions in collections on credit reports were evident
 - 25% decrease in probability of a medical collection
 - Those with a collection owed significantly less
- Household financial strain related to medical costs was mitigated
 - Substantial reduction across all financial strain measures
 - Captures “informal channels” people use to make it work
- Implications for both patients and providers
 - Only 2% of bills sent to collections are ever paid

Results: Self-reported health

Self-reported measures showed significant improvements one year after randomization

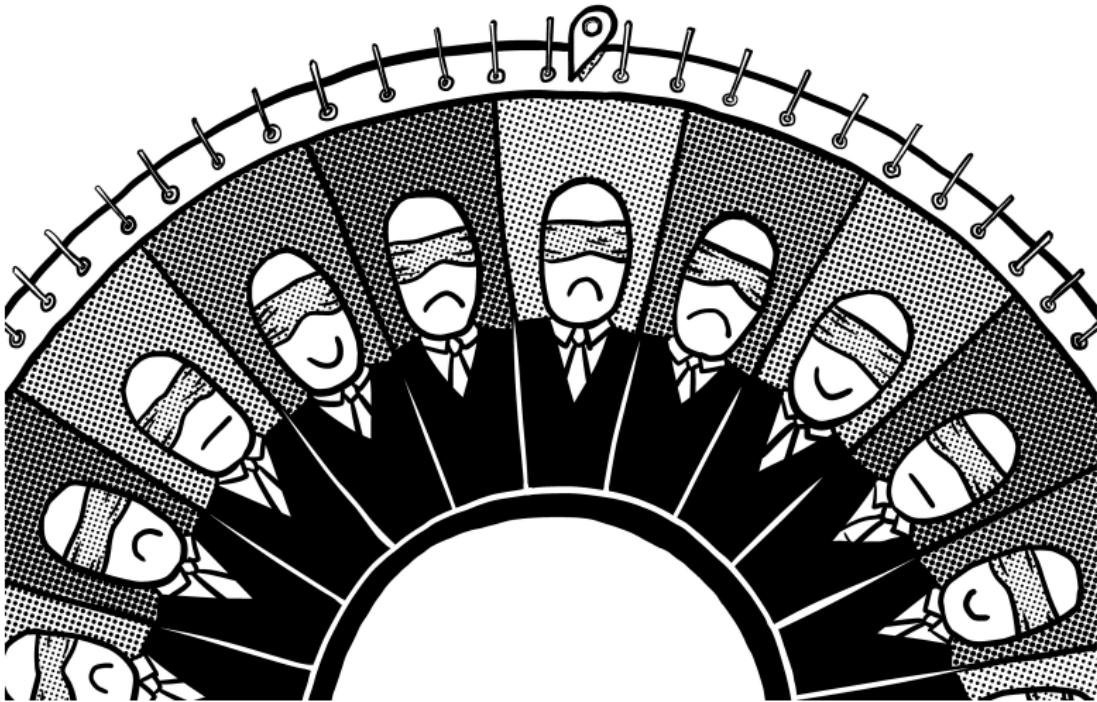
	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Health good, v good, excellent	54.8%	+3.9%	+13.3%	.0001
Health stable or improving	71.4%	+3.3%	+11.3%	.0001
Depression screen NEGATIVE	67.1%	+2.3%	+7.8%	.003
CDC Healthy Days (physical)	21.86	+.381	+1.31	.018
CDC Healthy Days (mental)	18.73	+.603	+2.08	.003

Source: Survey data

Summary: Self-reported health

- Overall, big improvements in self-reported physical and mental health
 - 25% increase in probability of good, very good or excellent health
 - 10% decrease in probability of screening for depression
- Physical health measures open to several interpretations
 - Improvements consistent with findings of increased utilization, better access, and improved quality
 - BUT in their baseline surveys, results appeared shortly after coverage ($\sim 2/3$ rds magnitude of full result)
 - May suggest increase in *perception* of well-being rather than physical health
- Biomarker data can shed light on this issue

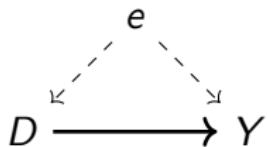
Judge Fixed Effect Design



Juvenile incarceration

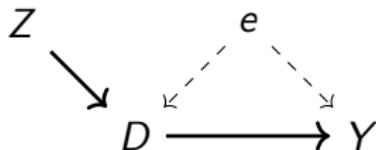
- Aizer and Doyle (2015) were interested in the causal effect of juvenile imprisonment on future crime and human capital accumulation
- Extremely important policy question given the US has the world's highest incarceration rate and prison population of any country in the world by a significant margin (500 prisoners per 100,000, over 2 million adults imprisoned, 4.8 million under supervision)
- High rates of incarceration extend to juveniles: in 2010, the stock of juvenile detainees stood at 70,792, a rate of 2.3 per 1,000 aged 10-19.
- Including supervision, US has a juvenile corrections rate 5x higher than the next highest country, South Africa

Confounding



- We are interested in the causal effect of juvenile incarceration (D) on life outcomes, like adult crime and high school completion
- But youth *choose* to commit crimes, and that choice may be due to unobserved criminogenic factors like poverty or underlying criminal propensities which are themselves causing those future outcomes

Leniency as an instrument



- Aizer and Doyle (2015) propose an instrument - the propensity to convict by the judge the youth is randomly assigned
- If judge assignment is random, and the various assumptions hold, then the IV strategy identifies the local average treatment effect of juvenile incarceration on life outcomes

The Main Idea

- “Plausibly exogenous” variation in juvenile detention stemming from the random assignment of cases to judges who vary in their sentencing
- Consider two juveniles randomly assigned to two different judges with different incarceration tendencies (Scott and Bob)
- Random assignment ensures that differences in incarceration between Scott and Bob are due to the judge, not themselves, because remember, they’re identical

Data

- 35,000 juveniles administrative records over 10 years who came before a juvenile court in Chicago (Juvenile Court of Cook County Delinquency Database)
- Data were linked to public school data for Chicago (Chicago Public Schools) and adult incarceration data for Illinois (Illinois Dept. of Corrections Adult Admissions and Exits)
- They wanted to know the effect of juvenile incarceration on high school completion (2nd data needed) and adult crime (3rd data needed) using randomized judge assignment (1st data needed)
- They need personal identifying information in each data set to make this link (i.e., name, DOB, address)

Preview of findings

- Juvenile incarceration decreased high school graduation by 13 percentage points (vs. 39pp in OLS)
- Increased adult incarceration by 23 percentage points (vs. 41pp in OLS)
- Marginal cases are high risk of adult incarceration and low risk of high school completion as a result of juvenile custody
- Unlikely to ever return to school after incarcerated, but when they do return, they are more likely to be classified as special ed students, and more likely to be classified for special ed services due to behavioral/emotional disorders (as opposed to cognitive disability)

“Plausibly” exogenous

- Very common in these studies for the assignment to some decision-maker to be *arbitrary* but not clearly random (i.e., not random no. generator)
- In this case, juveniles charged with a crime are assigned to a calendar corresponding to their neighborhood and calendars have 1-2 judges who preside over them
- 1/5 of hearings are presided over by judges who cover the calendar when the main judge can't, known as swing judges
- Judge assignment is a function of the sequence with which cases happen to enter into the system and judge availability that is set in advance
- No scope for which judge you see first; conversations with court administrators confirm its random

Structural equation

$$Y_i = \beta_0 + \beta_1 JI_i + \beta_2 X_i + \varepsilon_i$$

where X_i is controls and ε_i is an error term. In this, juvenile incarceration is likely correlated with the error term.

This is the “long” causal model. But note, from the prior DAG, we cannot control for e because it is unobserved. But it is confounding the estimation of juvenile incarceration’s effect on outcomes.

Incarceration Propensity as an Instrument

- The instrument is based on the randomized judge equalling the propensity to incarcerate from the randomly assigned judge
- “Leave-one-out mean”

$$Z_{j(i)} = \left(\frac{1}{n_{j(i)} - 1} \right) \left(\sum_{k \neq i}^{n_{j(i)} - 1} \tilde{J}_{l_k} \right)$$

- The $n_{j(i)}$ terms is the total number of cases seen by judge k , and \tilde{J}_{l_k} is equal to 1 if the juvenile was incarcerated during their first case
- Thus the instrument is the judge's incarceration among first cases based on all their other cases
- It's basically a judge fixed effect given the likelihood two judges have precisely the same propensity is small

Information about the instrument

- There are 62 judges in the data, and the average number of initial cases per judge is 607
- Substantial variation in the data - raw measure ranges from 4% to 21%
- Residualized measure based on controls still has substantial variation from 6% to 18%
- Variation comes from two sources: variation among the regular (nonswing) judges (80% of cases) and variation from the swing judges (20% of cases)

Distribution of IV

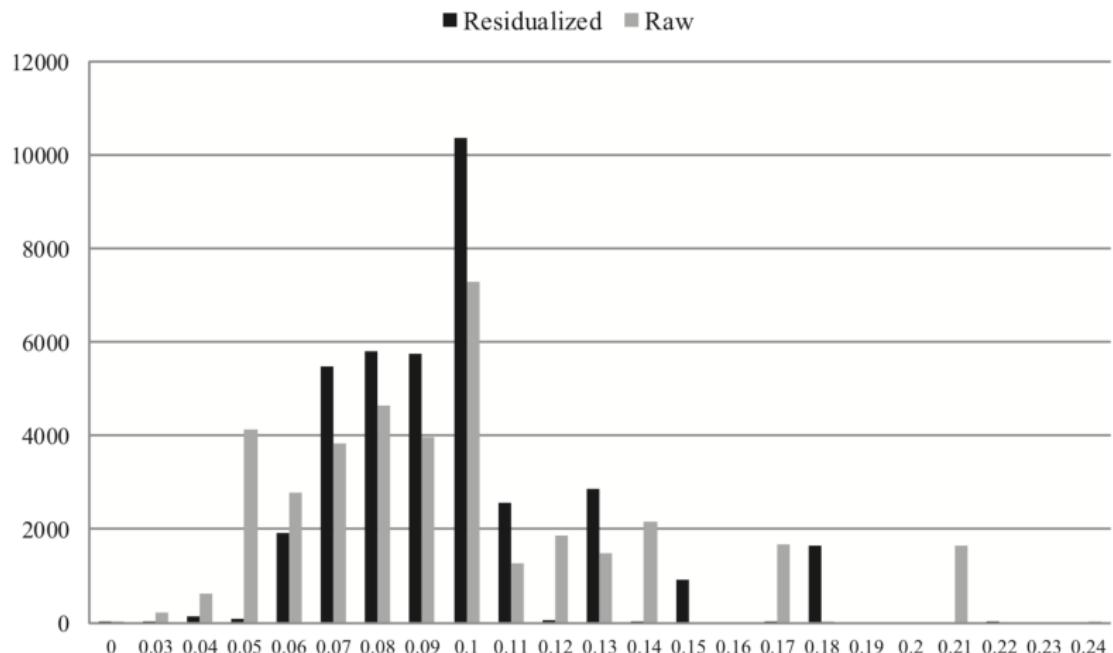


FIGURE I
Distribution of Z: Judge Incarceration Rate

Balance test

TABLE II
INSTRUMENT VERSUS JUVENILE CHARACTERISTICS

	Z distribution			Middle vs.	Top vs.
	Bottom tercile	Middle tercile	Top tercile	bottom <i>p</i> -value	bottom <i>p</i> -value
Z: first judge's leave-out mean incarceration rate in first cases	0.062	0.094	0.147	(.000)	(.000)
Juvenile characteristics					
Male	0.827	0.830	0.833	(.561)	(.311)
African American	0.724	0.737	0.742	(.096)	(.249)
Hispanic	0.189	0.176	0.172	(.061)	(.272)
White	0.078	0.079	0.078	(.833)	(.957)
Other race/ethnicity	0.009	0.008	0.007	(.352)	(.345)
Special education	0.241	0.237	0.252	(.549)	(.130)
U.S. census tract poverty rate	0.264	0.265	0.265	(.572)	(.696)
Age at offense	14.8	14.8	14.8	(.437)	(.434)
<i>P</i> (Juvenile incarceration <i>X</i>)	0.219	0.221	0.220	(.251)	(.516)
Observations	37,692				

First stage

TABLE III
FIRST STAGE

	(1)	(2)	(3)
Dependent variable: juvenile incarcerations		OLS	
First judge's leave-out mean incarceration rate among first cases	1.103 (0.102)	1.082 (0.095)	1.060 (0.097)
Demographic controls	No	Yes	Yes
Court controls	No	No	Yes
Observations	37,692		
Mean of dependent variable	0.227		

High school completion

TABLE IV
JUVENILE INCARCERATION AND HIGH SCHOOL GRADUATION

	Dependent variable: graduated high school						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample				Juvenile court sample		
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	-0.389 (0.0066)	-0.292 (0.0065)	-0.391 (0.0055)	-0.088 (0.0043)	-0.073 (0.0041)	-0.108 (0.044)	-0.125 (0.043)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440,797	440,797	420,033	37,692			
Mean of dependent variable	0.428	0.428	0.433	0.099			

Adult crime

TABLE V
JUVENILE INCARCERATION AND ADULT CRIME

	Dependent variable: entered adult prison by age 25						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample				Juvenile court sample		
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	0.407 (0.0082)	0.350 (0.0064)	0.219 (0.013)	0.200 (0.0072)	0.155 (0.0073)	0.260 (0.073)	0.234 (0.076)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440797	440797	420033	37692			
Mean of dependent variable	0.067	0.067	0.057	0.327			

Crime type

TABLE VI
JUVENILE INCARCERATION AND ADULT CRIME TYPE

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: entered adult prison by age 25 for crime type					
	Homicide			Violent		
	OLS	OLS	2SLS	OLS	OLS	2SLS
Juvenile incarceration	0.051 (0.0031)	0.021 (0.0030)	0.035 (0.030)	0.138 (0.0046)	0.061 (0.0050)	0.149 (0.041)
Sample	Full CPS	Juvenile court	Juvenile court	Full CPS	Juvenile court	Juvenile court
Mean of dep. var.: JI = 0	0.008	0.043	0.043	0.024	0.121	0.121
Observations	440,797	37,692	37,692	440,797	37,692	37,692
	Property			Drug		
Juvenile incarceration	0.079 (0.0040)	0.047 (0.0038)	0.142 (0.044)	0.183 (0.011)	0.078 (0.0068)	0.097 (0.052)
Sample	Full CPS	Juvenile Court	Juvenile Court	Full CPS	Juvenile Court	Juvenile Court
Mean of dep. var.	0.013	0.060	0.060	0.034	0.176	0.176
Observations	440,797	37,692	37,692	440,797	37,692	37,692

High school transfers

TABLE VIII
INTERMEDIATE SCHOOLING OUTCOMES: HIGH SCHOOL TRANSFERS

Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)
	Ever present in CPS school at least 1 year after Initial hearing		Transferred to another CPS high school in years after hearing		Ultimate transfer: adult correctional facility	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
Juvenile incarceration	-0.025 (0.0063)	-0.215 (0.069)	0.055 (0.010)	-0.115 (0.243)	0.127 (0.006)	0.243 (0.060)
Mean of dependent variable	0.666		0.242		0.175	
Observations	37,692		18,195		37,692	

Developing emotional problems

TABLE IX
INTERMEDIATE SCHOOLING OUTCOMES: SPECIAL EDUCATION STATUS

Dependent variable:	Special education type observed in years after initial hearing					
	Any Special Education		Emotional/ behavioral disorder		Learning disability	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
Juvenile incarceration	-0.024 (0.004)	-0.003 (0.037)	0.027 (0.003)	0.133 (0.043)	-0.040 (0.004)	-0.097 (0.039)
Mean of dependent variable	0.193		0.082		0.085	
Observations	29,794					

Concluding remarks

- Sad, but important, paper - the marginal kid shouldn't have been incarcerated
- More generally, leniency designs are very powerful and very common if you know how to look for them
- Bottleneck, influential decision-makers, discretion - these are the three elements of the design

John Snow

- John Snow was a practicing anesthesiologist in the mid 19th century London
- He was then famous for inventing a machine that would carefully deliver chloroform to patients in homogenous dosage which reduced mortality from anesthesia
- But he is now famous for providing convincing evidence that cholera was a waterborne disease during the 1854 outbreak
- Published two works on cholera – an essay in 1849, and a book in 1855
- Died of a stroke in 1858

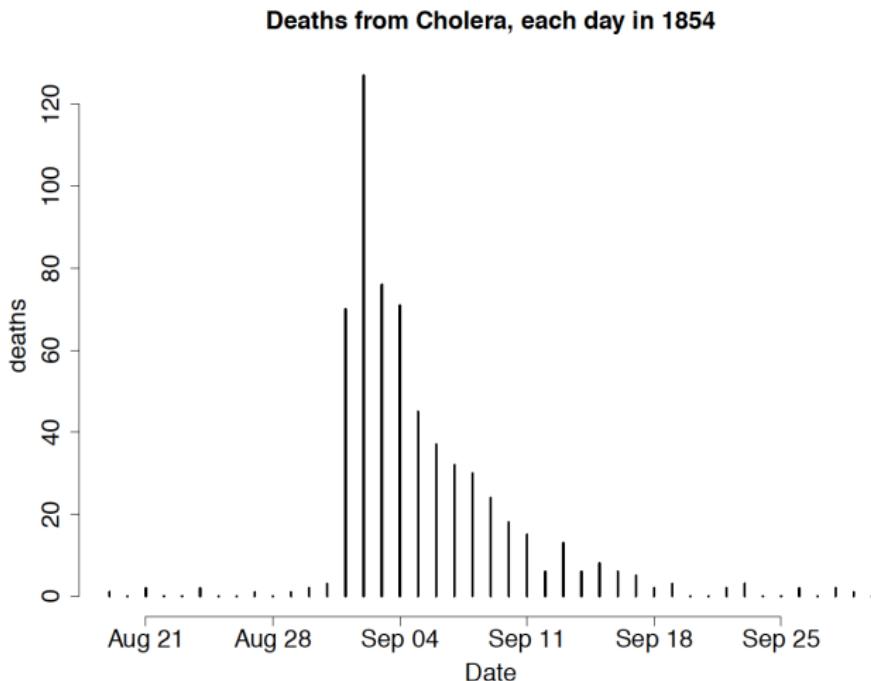


Figure: Daily cholera deaths, London (Coleman 2019)

Cholera background

- Cholera hits London three times in the early to mid 1800s causing large waves of tens of thousands of deaths
- Three London epidemics – 1831-1832, 1848-1849, 1853-1854
- Cholera attacked victims suddenly, with a 50% survival rate, and very painful symptoms included vomiting and acute diarrhea

Miasmis

- 19th century London was a filthy place with waste collecting in cesspools under houses or emptied into open ditches and sewers
- Majority opinion about disease was *miasmis*
- Miasmis hypothesized that disease transmission was caused by vapors and smells; unclear its relevance for person-to-person

Never before seen microorganism

- Microscopes were around but had horrible resolution
- Most human pathogens couldn't be seen
- Johnson (2007) reports Snow did track down a microscope but could only see blurry things moving around
- Isolating these microorganisms wouldn't occur for half a century

Thought Experiment

- How will he convince anyone that cholera is waterborne and not due to “bad air”?
- Consider the ideal experiment: randomize households by coin flip to receive water from runoff (control) vs. water without runoff (treatment)
- Unethical, impractical and unrealistic
- Even if the randomized experiment is not possible, the thought experiment suggests the observational equivalent

Multiple sources of evidence, not just one

Snow makes his argument with many pieces of evidence that when taken together are very compelling that water, not air, is the cause of the cholera epidemics. These can be categorized as:

- ① Observation
- ② Broad Street Pump
- ③ Grand Experiment

How he argues for the Broad street pump

- Famous map showing unusual mass of cholera deaths near the public Broad street pump
- He was looking for the source, but he was not inductively forming his theory with this map because he already knew the mechanism
- He was assembling evidence that would further refute the explanations of those who advocated an alternative explanation of the outbreak

Two companies fight for customers

- Southwark and Vauxhall Waterworks Company and the Lambeth Water Company competed over some of the regions south of the Thames
- In 16 sub-districts, with a population of 300,000, they competed directly, even supplying customers side-by-side

"In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies." Snow (1855) p 75

Lambeth moves its pipe

- During the 1849 epidemic, both companies drew water from Thames which was polluted with sewage and cholera
- London passes legislation requiring utility companies to move their pipes above the city
- In 1852, the Lambeth Company, a water utility company, changed supply from Hungerford Bridge
- It moved its intake pipe upstream to cleaner water and in response to legislation (SV delayed)
- This created a natural experiment because Southwark and Vauxhall left its intake pipe in place

Meticulous Data Collection

- Two types of data: DD uses aggregate deaths bc of mixing of customers whereas his Broad Street evidence focused on individuals
- Collected detailed information from households with cholera deaths on utility subscription (Lambeth or SV)
- Many residents didn't know their water company – distant landlords paid for it
- He knew Lambeth water was four times saltier, so he'd take a sample and test it using a saline test back at his office

Shoelather and knowledge of institutional details

- Careful balance checks – “the pipes of each Company go down all the streets into nearly all the courts and alleys”
- Concern for sample selection bias –“No fewer than 3000 people of both sexes [of all types affected]”
- Treatment assignment was arbitrary – “a few houses supplied by one Company and a few by the other”

Table XII

Modified Table XII (Snow 1854)

Company name	1849	1854
Southwark and Vauxhall	135	147
Lambeth	85	19

Estimated ATT using DD is 78 fewer deaths per 10,000

Failure to convince

"In spite of what has since been recognized as a classic exercise in data, analysis, and argument, Snow failed to convince the medical profession, the policy-making establishment, or the public." (Coleman 2019)

Final victory

- Another cholera outbreak in 1866, east of London, is when Snow's ideas were gradually and reluctantly accepted by public officials and the scientific community
- 1866 outbreak was confined only to the east of London, which was the last area not yet connected to the newly constructed sewage system which discharged sewage below the Thames
- The rest of London didn't have an outbreak
- This was the final piece of evidence that swayed skeptics and led to a more reasoned assessment of Snow's data and analysis

Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
Southwark and Vauxhall	Before	$Y = SV$		D
	After	$Y = SV + T$	T	

Sample averages

$$\widehat{\delta}_{kU}^{2 \times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Population expectations

$$\widehat{\delta}_{kU}^{2\times 2} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

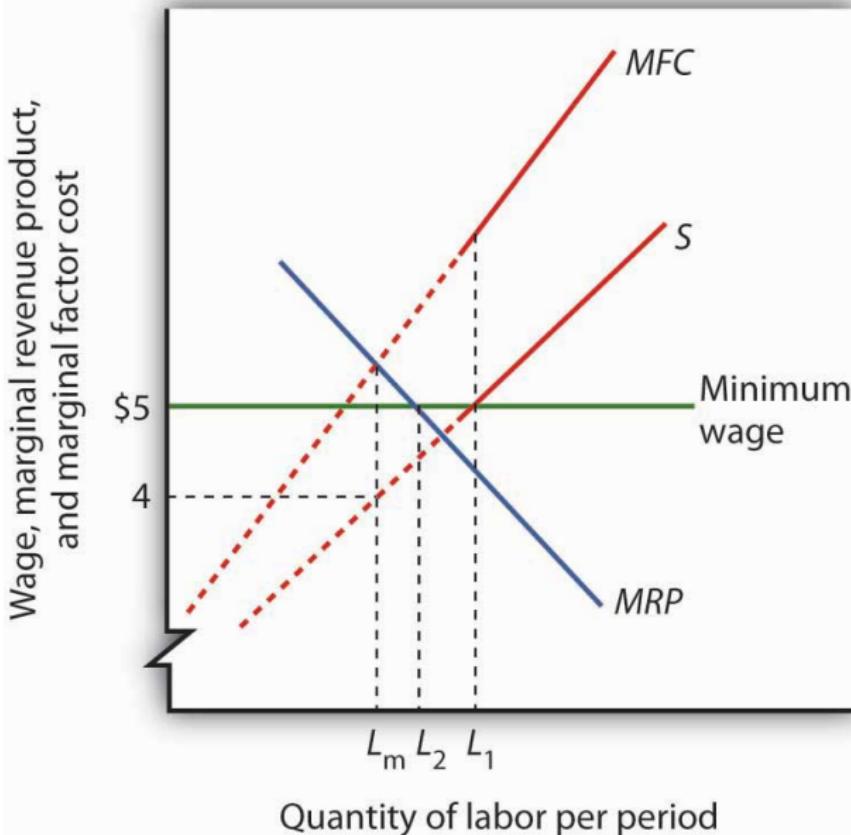
Another famous DD study

- Card and Krueger (1994) was a seminal study on the minimum wage both for the result and for the design
- Not the first time we saw DD in the modern period - there's Ashenfelter (1978) and Card (1991) - but got a lot of attention

Competitive vs noncompetitive markets

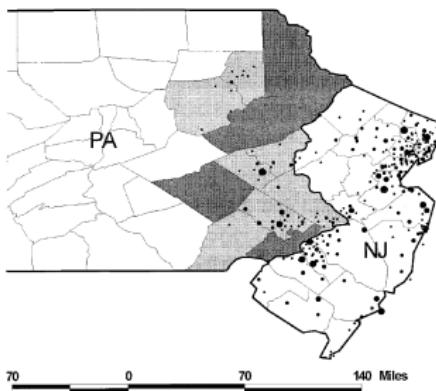
- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In a competitive input market, increases in the minimum wage would move us up a downward sloping labor demand curve → employment would fall
- Monopsony (imperfect labor markets) suggest the opposite effect whereby raising the minimum wage increases employment

Monopsony's minimum wage predictions



Card and Krueger (1994)

- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey - shoeleather!

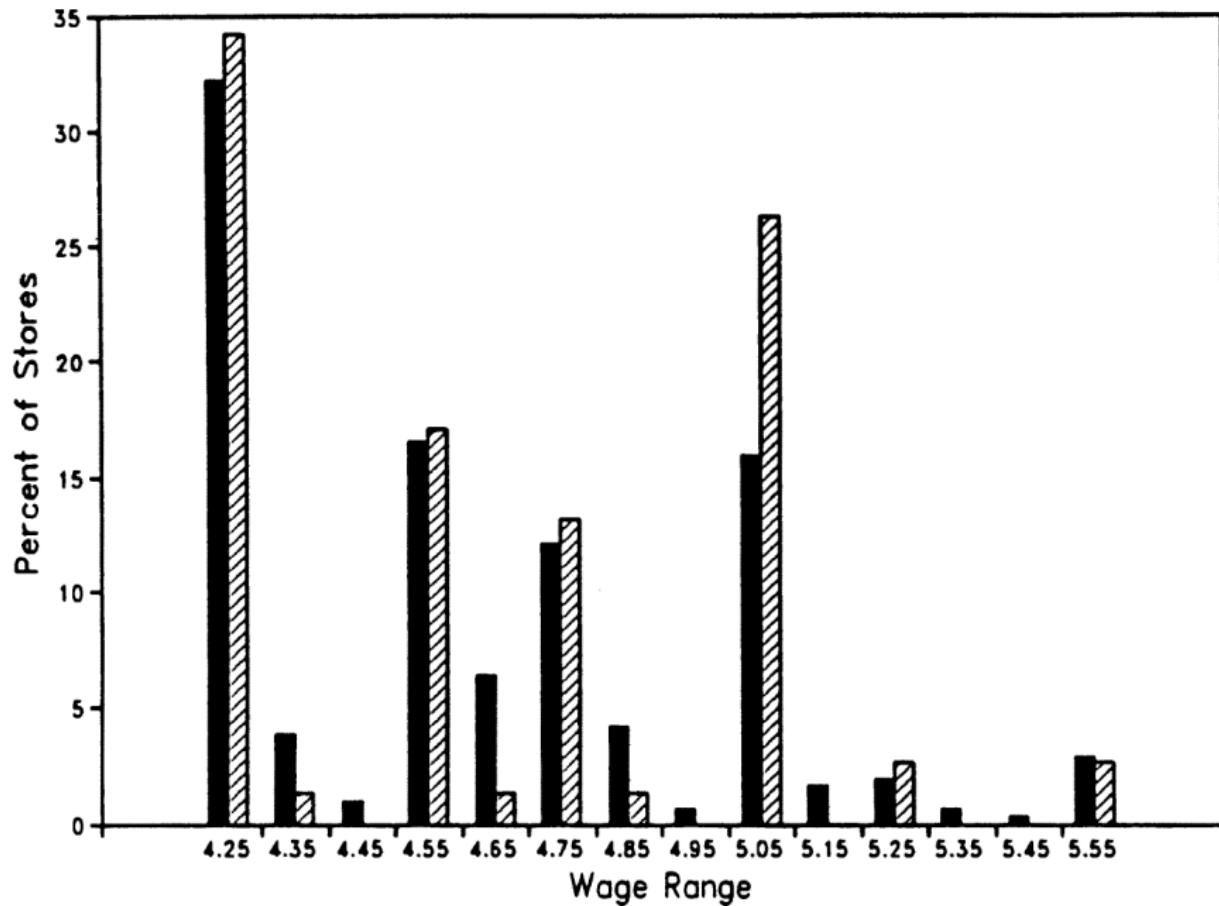
Parallel trends assumption

- Key identifying assumption is the “parallel trends” assumption

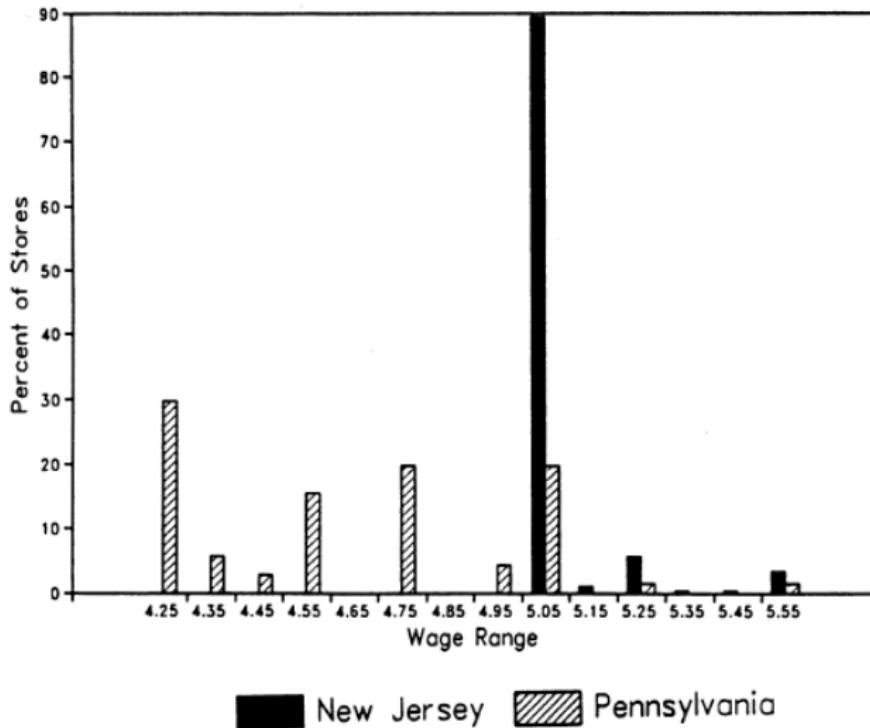
$$\underbrace{[E[Y_{NJ}^0 | Post] - E[Y_{NJ}^0 | Pre]] - [E[Y_{PA}^0 | Post] - E[Y_{PA}^0 | Pre]]}_{\text{Non-parallel trends bias}}$$

- Note the counterfactual - it is *not testable* no matter what someone tells you, bc New Jersey's post period potential employment in a world with a lower minimum wage is unobserved
- Let's look at this a couple of different ways, including a graphic showing the binding minimum wage

February 1992



November 1992



Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment *rose* in NJ relative to PA after the minimum wage change - consistent with monopsony theory

Regression DD

The typical regression model we estimate is

$$Y_{it} = \beta_1 + \beta_2 \text{Treat}_i + \beta_3 \text{Post}_t + \beta_4 (\text{Treat} \times \text{Post})_{it} + \varepsilon_{it}$$

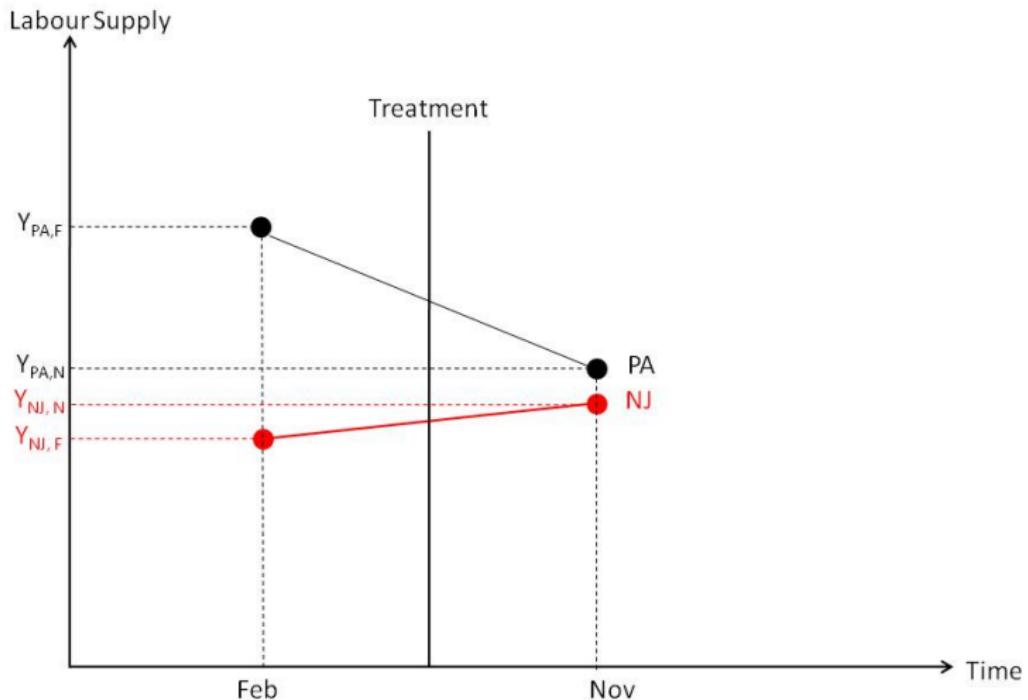
where Treat is a dummy if the observation is in the treatment group and Post is a post treatment dummy

Regression DD - Card and Krueger

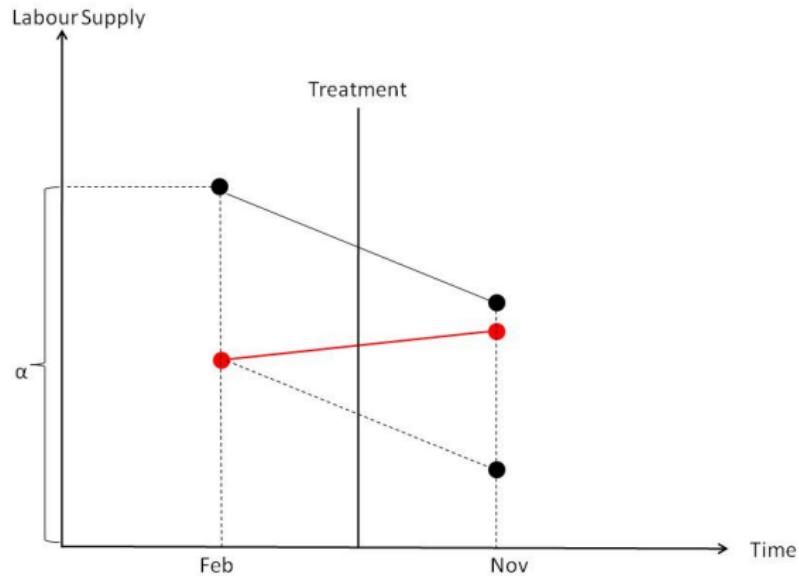
- In the Card and Krueger case, the equivalent regression would be:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{its}$$

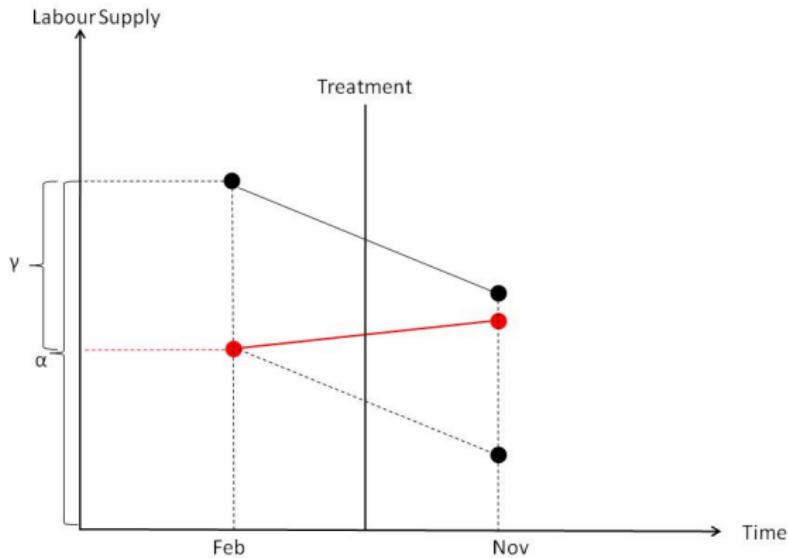
- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$



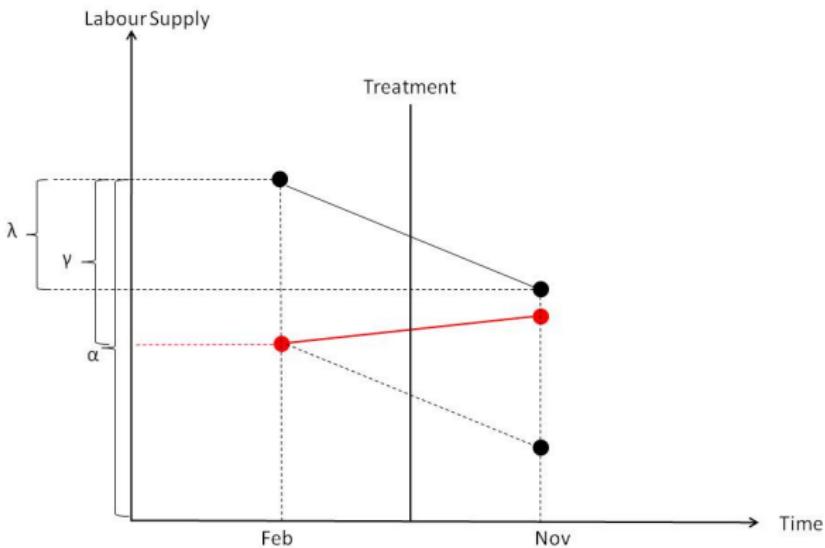
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



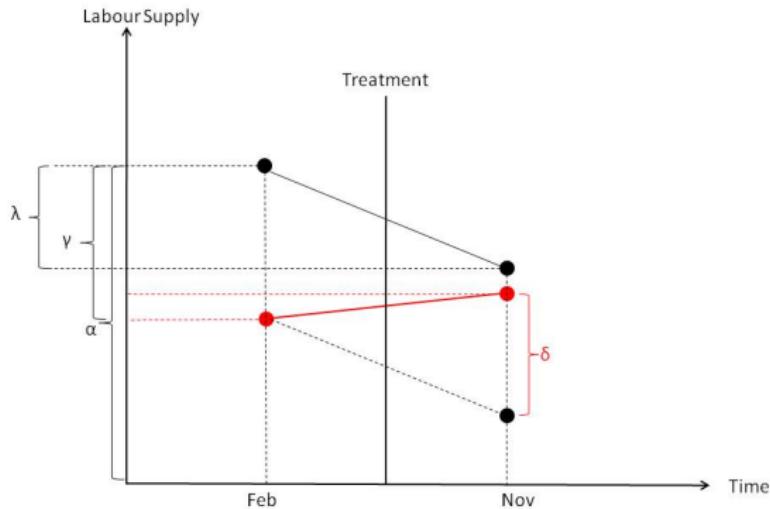
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Key assumption of any DD strategy: Parallel trends

- The key assumption for any DD strategy is that the outcome in treatment and control group would follow the same time trend in the absence of the treatment
- This doesn't mean that they have to have the same mean of the outcome
- But regardless of parallel trends, OLS always estimates the vertical bar on next slide

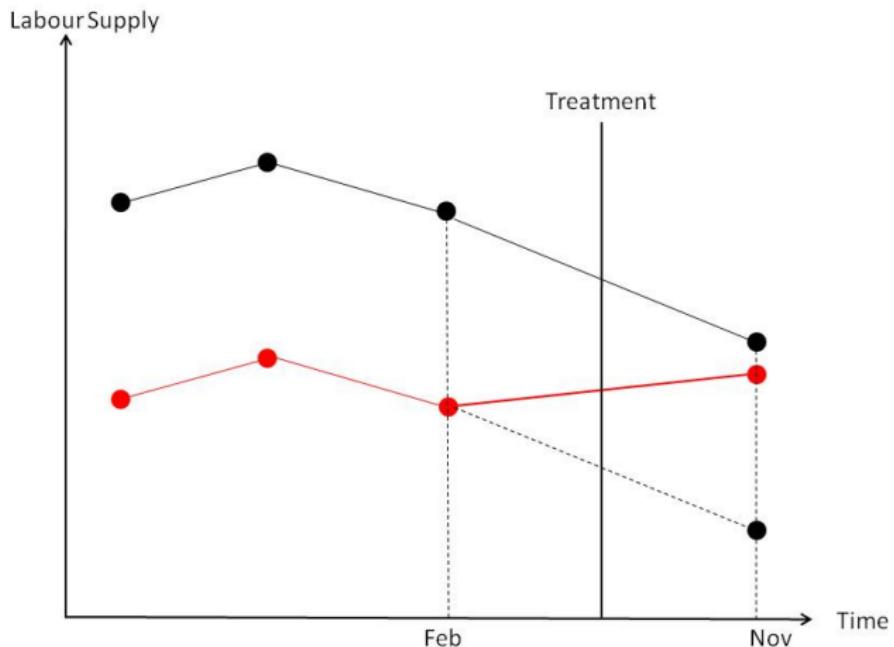
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Parallel leads, not trends

- The identifying assumption for all DD designs is some representation of a counterfactual parallel trend
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check using pre-treatment data to show that the trends had been the same prior to treatment
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

Plot the raw data when there's only two groups



Losing parallel trends

- If parallel trends doesn't hold, then ATT is not identified
- But, regardless of whether ATT is identified, OLS always estimates the same thing
- That's because OLS uses the slope of the control group to estimate the DD parameter, which is only unbiased if that slope is the correct counterfactual for the treatment group

Labor Supply

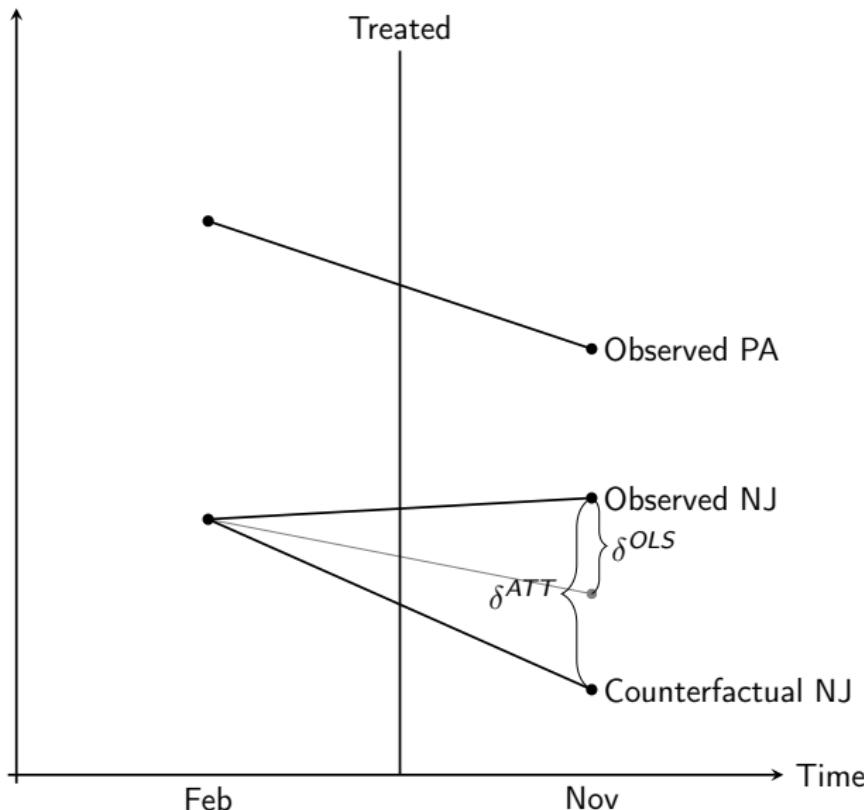


Figure: DD regression diagram without parallel trends

Differential timing makes pre-treatment undefined for untreated groups

- New Jersey treated in late 1992, New York in late 1993, Pennsylvania never treated
- Pre-treatment:
 - New Jersey: <1992
 - New York: <1993
 - Pennsylvania: undefined
- So how do we check parallel leads?

Event study regression

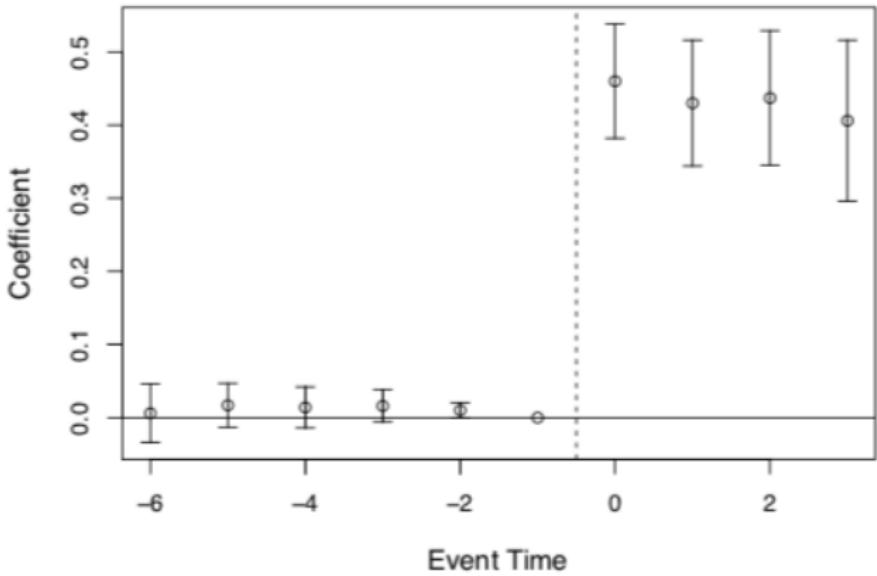
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-1}^{-q} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

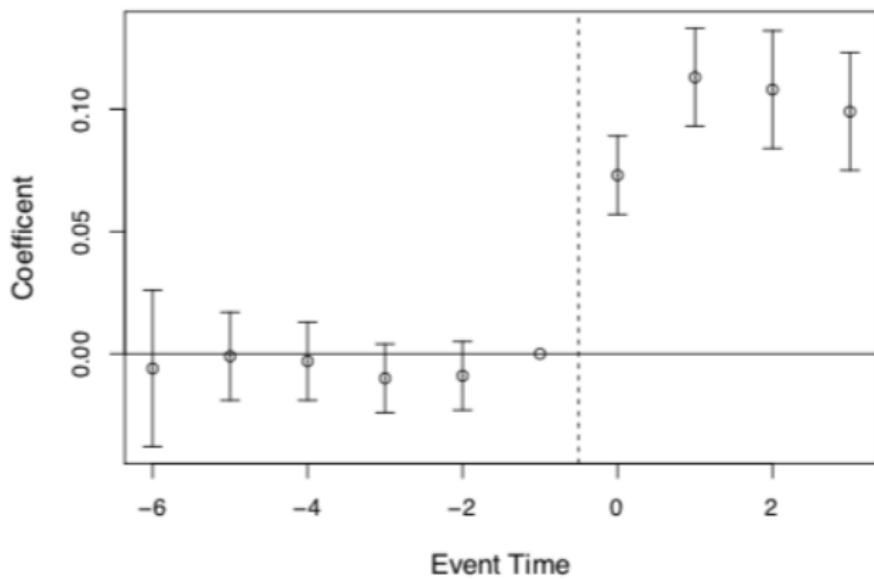
- Treatment occurs in year 0
- Includes q leads or anticipatory effects
- Includes m leads or post treatment effects

Medicaid and Affordable Care Act example

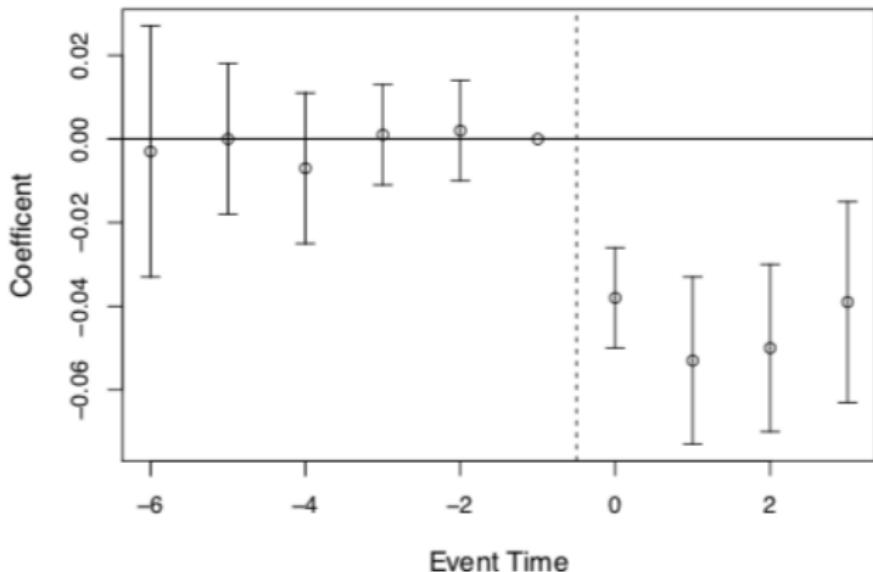
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

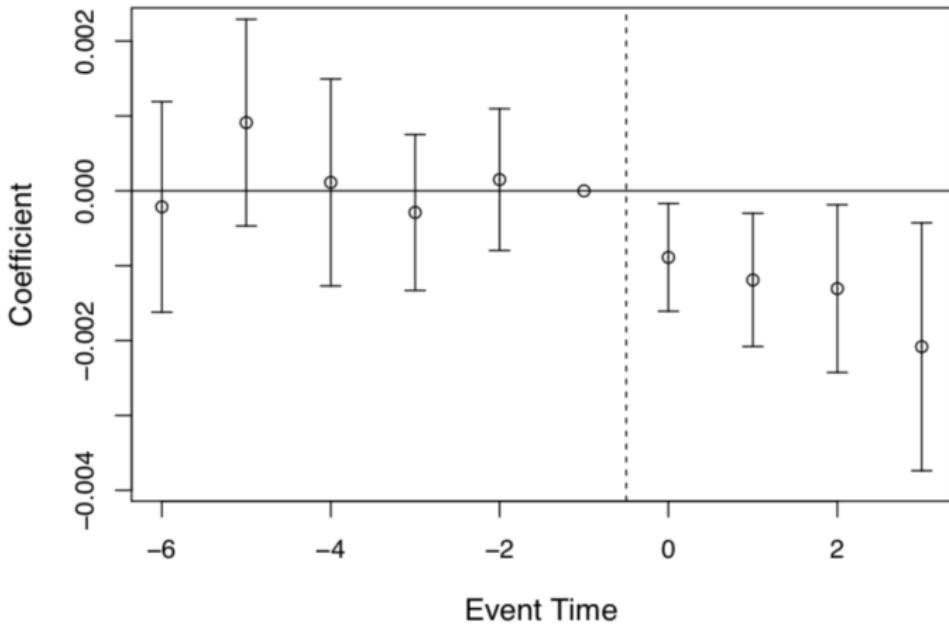


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on annual mortality

Juvenile Curfew Laws

- ▶ youth curfews: popular tool for combating juvenile delinquency
- ▶ 80% of 347 largest US cities have curfew laws in 1997
- ▶ little evidence that such policies are effective at reducing crime
- ▶ *The Impact of Juvenile Curfew Laws on Arrests of Youth and Adults (Patrick Kline, ALER; 2011)*
 - ▶ event study research design (different treatment dates)
 - ▶ arrest behavior of various age groups within a city before/after law
 - ▶ key contribution: separates 2 kinds of treatments

2 separate treatments

- ▶ 2 treatments, applying to different age groups
- ▶ both interesting to economists!
- ▶ "Statutory" Treatment
 - ▶ punishment for curfew violations by minors
 - ▶ affects only youth under curfew age
 - ▶ informs on substitutability of crime across time
- ▶ "Statistical Discrimination" Treatment
 - ▶ lower standards of probable cause due to *perceived youth*
 - ▶ *police unable to distinguish ex-ante people below/above curfew age*
 - ▶ *probability of being stopped is higher for both adjacent groups*
 - ▶ *American Civil Liberties Union = violation of civil liberties*
 - ▶ *informs elasticity of crime wrt prob(detection)*

How can we separate the two treatments?

- ▶ probability being stopped depends little on actual age
- ▶ "Statutory" Treatment
 - ▶ compare group just above/below curfew age
- ▶ "Statistical Discrimination"
 - ▶ compare group just above and older adults

Juvenile Curfew Laws

- ▶ Juvenile curfews = local ordinances proscribing minors from occupying public areas and streets during particular times
- ▶ 1880: First case in Omaha, NE
- ▶ 1990s: violent crime/victimization of minors rise
- ▶ cities learn how to craft the law in an amenable manner to courts
- ▶ Dallas, TX 1991 LAW as a model to other cities
- ▶ focused on specified age group/times, and with exceptions
- ▶ designed to deal with specific needs of the city, based on data
- ▶ US Court of Appeals and Supreme Court upheld the law

The Dallas Model

- ▶ all youth under 17 at 11pm(12pm)-6am weekdays(ends).
- ▶ extensive campaign to inform the public
- ▶ police used verbal warning, take home, fines up to \$500 or custody
- ▶ businesses could also be fined for having minors in its premises

- ▶ Three months: no arrests, but hundreds warnings and citations
- ▶ juvenile victimization fell by 17% from 1,950 in previous year

Event Study Methodology

- ▶ recent studies rely on variation in the date of adoption of city curfew laws to identify treatment effects on criminal behavior (Males and Macallair, 1999; McDowall, Loftin, and Wiersema, 2000).
- ▶ problem: not causal if curfew laws are enacted in response to city-specific trends in arrests.
- ▶ aim: research design capable of testing for such trends and recovering any dynamics of the impact of curfew enactment

Event Study Methodology

$$R_{cy} = \sum_t \beta_t D_{cy}^t + \psi_y + \theta_c + \epsilon_{cy}$$

where R_{cy} is the log(arrests in a age group) in city c and year y , ψ_y is a year effect, θ_c is a city effect, β_t coefficients represent the time path of arrest relative to the date of curfew enactments

- ▶ D_t = dummies for when curfew enactment is t period away in a city

$$D_{cy}^t = I[y - e_c = t]$$

Event Study Methodology

- ▶ if curfews are randomly assigned, $\beta_t = 0$ for any $t < 0$ should hold
- ▶ in other words, treatment is not preceded by trends in city-specific arrests
- ▶ restrictions on estimation:
 - ▶ (a) the D_{cy}^t set is collinear with θ_c , so normalize $\beta_{-1} = 0$
 - ▶ (b) restrict $t \in [-6, 6]$ so dynamics wear off after 6 years

Data

- ▶ 54 cities with curfew enactment in 1980-2014 period, all regions and sizes ($\geq 180,000$)
- ▶ sources: newspaper stories, Ruefle and Reynolds (1996), phone calls, municipal code
- ▶ FBI (type I offences) arrest data by gender, age, city and type of offence
- ▶ focus on serious felonies (unlikely reclassified as curfew violations)

Data

Table 1. Curfew Data by City

City	State	Population in 1990	Year Curfew Enacted	Statutory Curfew Age
Akron	OH	223,019	1990	17
Albuquerque	NM	384,736	1994	16
Anaheim	CA	266,406	1990	16
Anchorage	AK	226,338	1989	15
Atlanta	GA	394,017	1991	16
Austin	TX	465,577	1992	16
Baltimore	MD	736,014	1995	16
Baton Rouge	LA	219,531	1995	16
Birmingham	AL	265,852	1996	16
Buffalo	NY	328,123	1994	16
Charlotte	NC	396,003	1985	15
Cincinnati	OH	364,040	1994	17
Cleveland	OH	505,616	1993	17
Colorado Springs	CO	281,140	1992	17
Corpus Christi	TX	257,453	1991	16
Dallas	TX	1,006,831	1994	16

Results: curfew enactments and curfew violation arrest

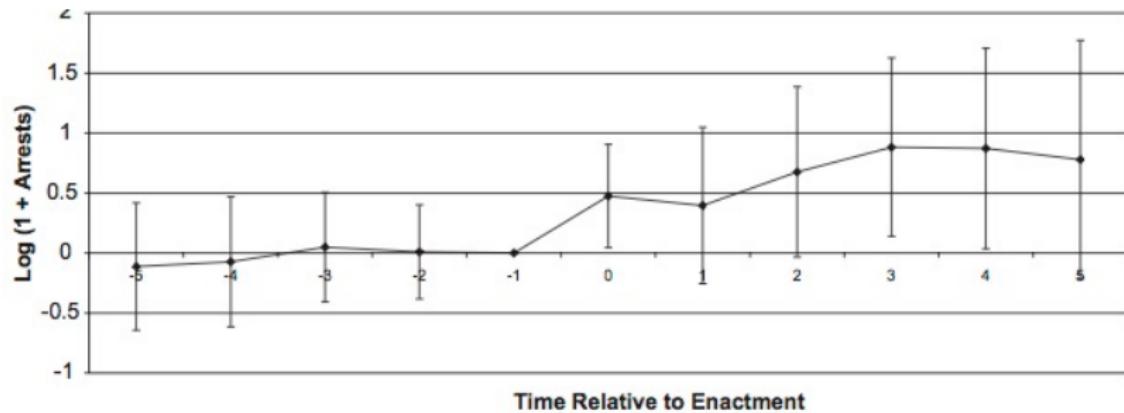


Figure 1. Arrests of Youth below Curfew Age for Curfew and Loitering Violations.

Curfew enactments and type I offence arrests

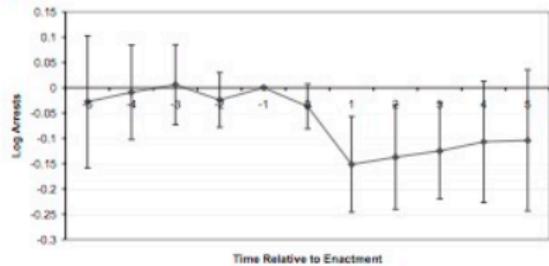
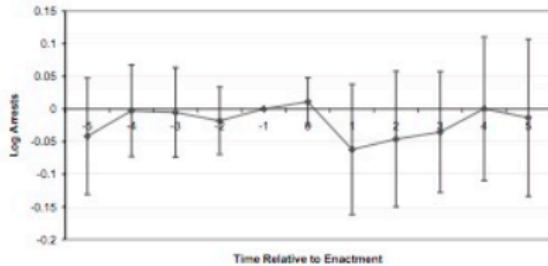
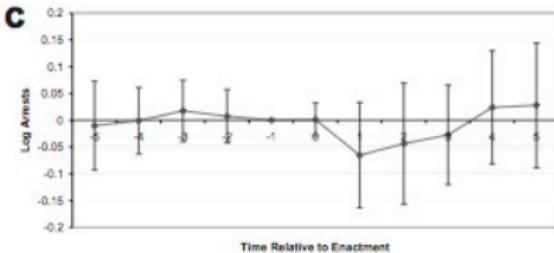
a**b****c**

Figure 2. (a) Youth below Curfew Age. (b) Young Adults above Curfew Age.
(c) Adults Age 25+.

Results

- ▶ sharp timing of results = data classification is accurate
- ▶ no pre-existing trend in youth or adults above age 25
- ▶ imprecise estimates
- ▶ evidence that arrest of underage youth drop by 15% and settles at 10% (135 per year)
- ▶ ~~effect on adults = potential influence of social programming efforts like midnight basketball~~
- ▶ ~~focus only on youth 1 year above/below curfew age~~

Further Results: officers per capita

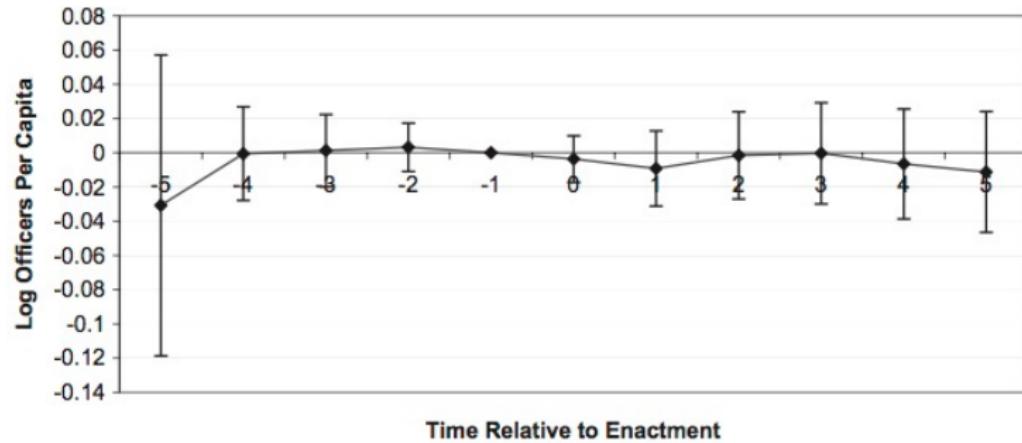
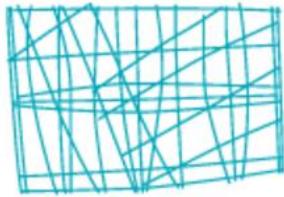


Figure 4. Log Officers per Capita.

Concluding Remarks

- ▶ curfew laws appear to have important effects on youth crime
- ▶ reduce arrests in 10% after 5 years for underage
- ▶ small reduction in adult arrests, similar across groups, suggesting no statistical discrimination effect, however precision for older age groups is poor
- ▶ youth crime is not perfectly substitutable across time

- ▶ what is the mechanism of this effect? police enforcement? parents?
- ▶ cost-benefit analysis? need cost information.



NEREUS

The University of São Paulo
Regional and Urban Economics Lab



FEAUSP

Rainforest Conservation policy assessment: The case of the Atlantic Forest in Brazil

Keyi Ando Ussami¹ & Ariaster Baumgratz Chimeli¹

¹ University of São Paulo, Department of Economics

- PART I: The Atlantic Forest Law: is it effective?
- PART II: Exploring the mechanisms behind the Atlantic Forest conservation

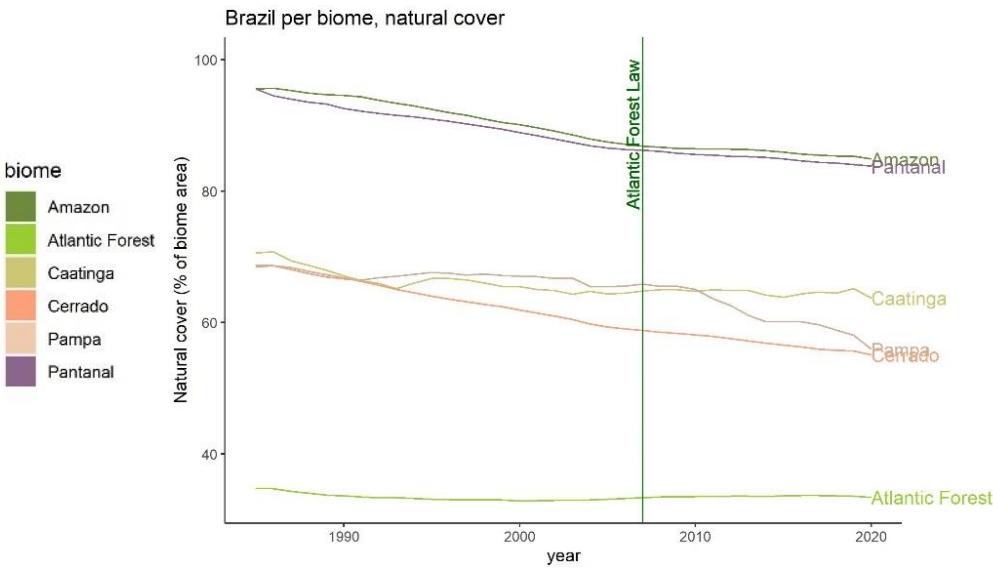
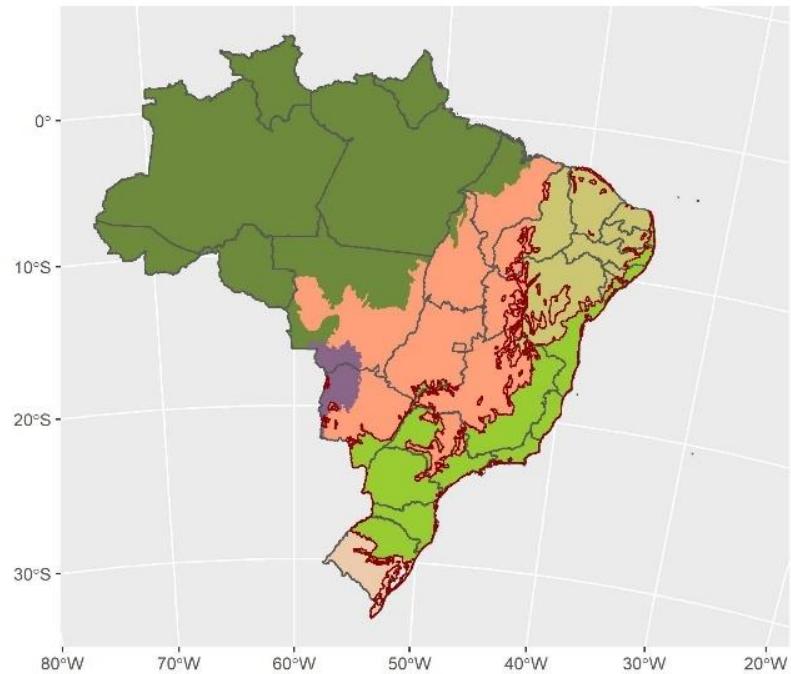
Atlantic Rainforest

- 112 Mha of highly heterogeneous regions, high biodiversity and endemism → world biodiversity hotspot
- Its natural forests is now reduced to 28% of its original area
- 70% of Brazilian population
- 80% of Brazilian GDP
- Concerns about the AF started in the 70's



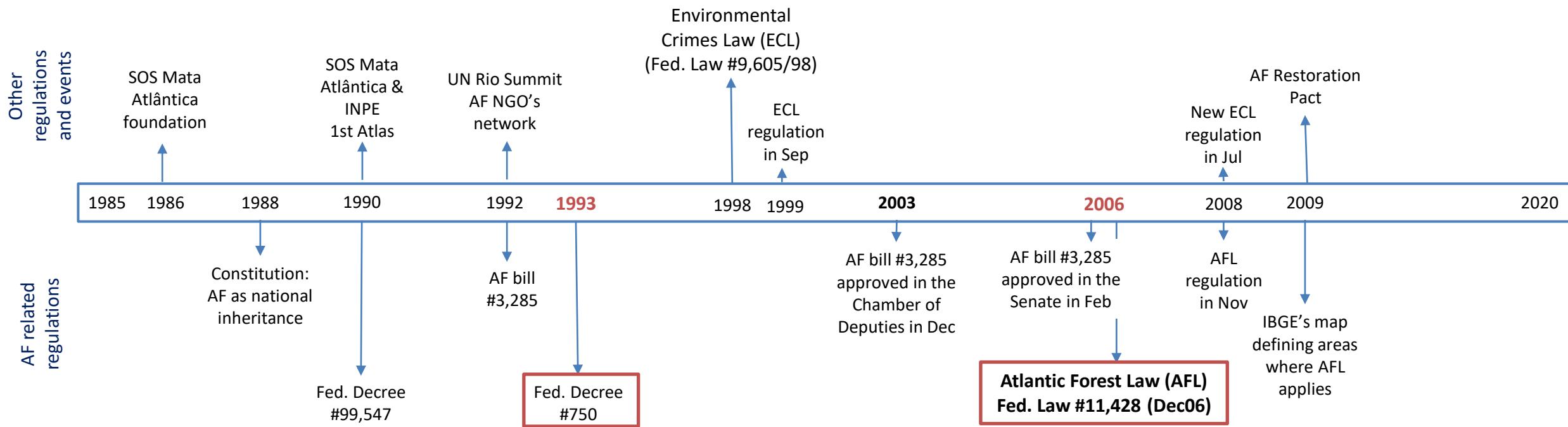
Atlantic Forest Law (AFL)

zero deforestation policy
for the biome (Dec. 2006)



Data source: Mapbiomas collection 6.0

Timeline



PART I

The Atlantic Forest Law: is it effective?

Outline

- Literature review
- Data and empirical strategy (Mapbiomas, DiD)
- Results (natural cover and net revegetation)
- Robustness check (nat. forest cover, anthropic land cover, subsample)
- Analysis of heterogeneity I (munic. with different levels of remnants of native veg. in the baseline: 0-25, 25-50, 50-75, 75-100)
- Analysis of heterogeneity II (different groups of states)
- Conclusions

Literature review

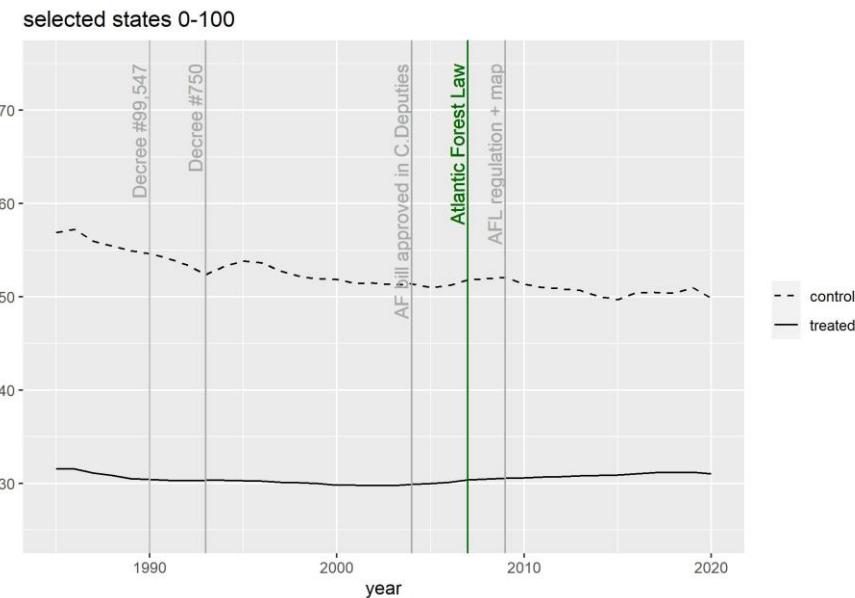
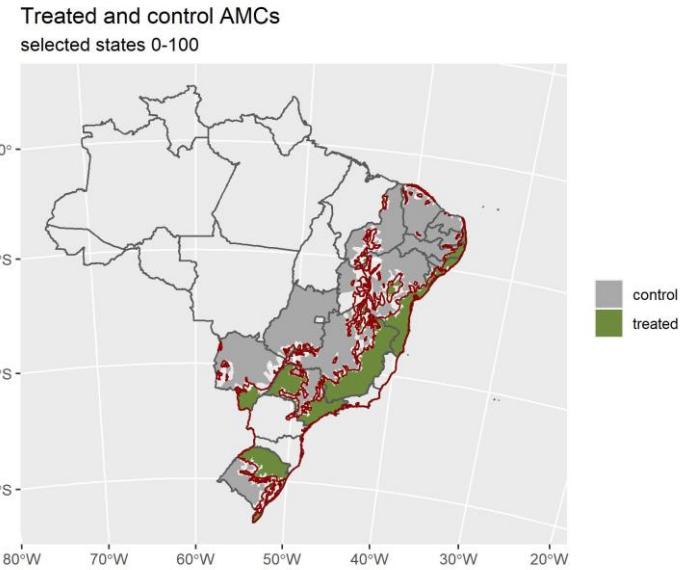
- Protected areas – conservation units or indigenous lands
 - Effect of protected areas from the World Database for Protected Areas (Joppa & Pfaff 2011, Proc. Royal Soc. B)
 - Effect of protected areas, different levels of protection (Ferraro et al., 2013, Environ Res. Letters; Pfaff et al. 2014, World Development)
 - Formalization of indigenous land rights (BenYishay et al. 2017, JEEM)
- Different instruments to protect native vegetation:
 - PPCDAm (command and control, new protected areas, targeting and law enforcement) (Assunção et al. 2015, Environ. Develop. Econ.; Burgess et al. 2018, NBER WP; Assunção & Rocha 2019, Environ. Develop. Econ.; Sills et al. 2015 PLoS ONE)
 - Protection of particular species (Chimeli & Boyd 2010, Land Economics; Ferraro et al. 2004, JEEM)
 - Payment for Ecosystem Services (Fiorini et al. 2020 Ecol. Econ)

Literature review

- Forest Code
 - Depends basically on landowner's decision
 - Quantification of native vegetation areas contrasting them to the minimum required by law (Biggs et al., 2019; Soares-filho et al., 2014)
 - Simulation contrasting scenario with/without enforcement (Soterroni et al., 2018; Verburg et al., 2014)
 - No studies assessing causal impact of the Forest Code
- Vegetation Management Act, Queensland AU (Simmons et al. 2018, Environ. Res. Letters)
- Atlantic Forest
 - Quantifying remaining native vegetation and its implications for conservation (Crouzeilles et al., 2019; Rezende et al., 2018; Ribeiro, Metzger, Martensen, Ponzoni, & Hirota, 2009)
 - Drivers of deforestation and conservation of the Atlantic Forest (Ruggiero et al. 2022 Ecol. Econ; Ruggiero et al. 2021, Cons. Letters, Freitas, Hawbaker, & Metzger, 2010 Forest Ecol. and Manag.)
- To our knowledge, this is the first study to analyze the impact of Atlantic Forest protection policy.

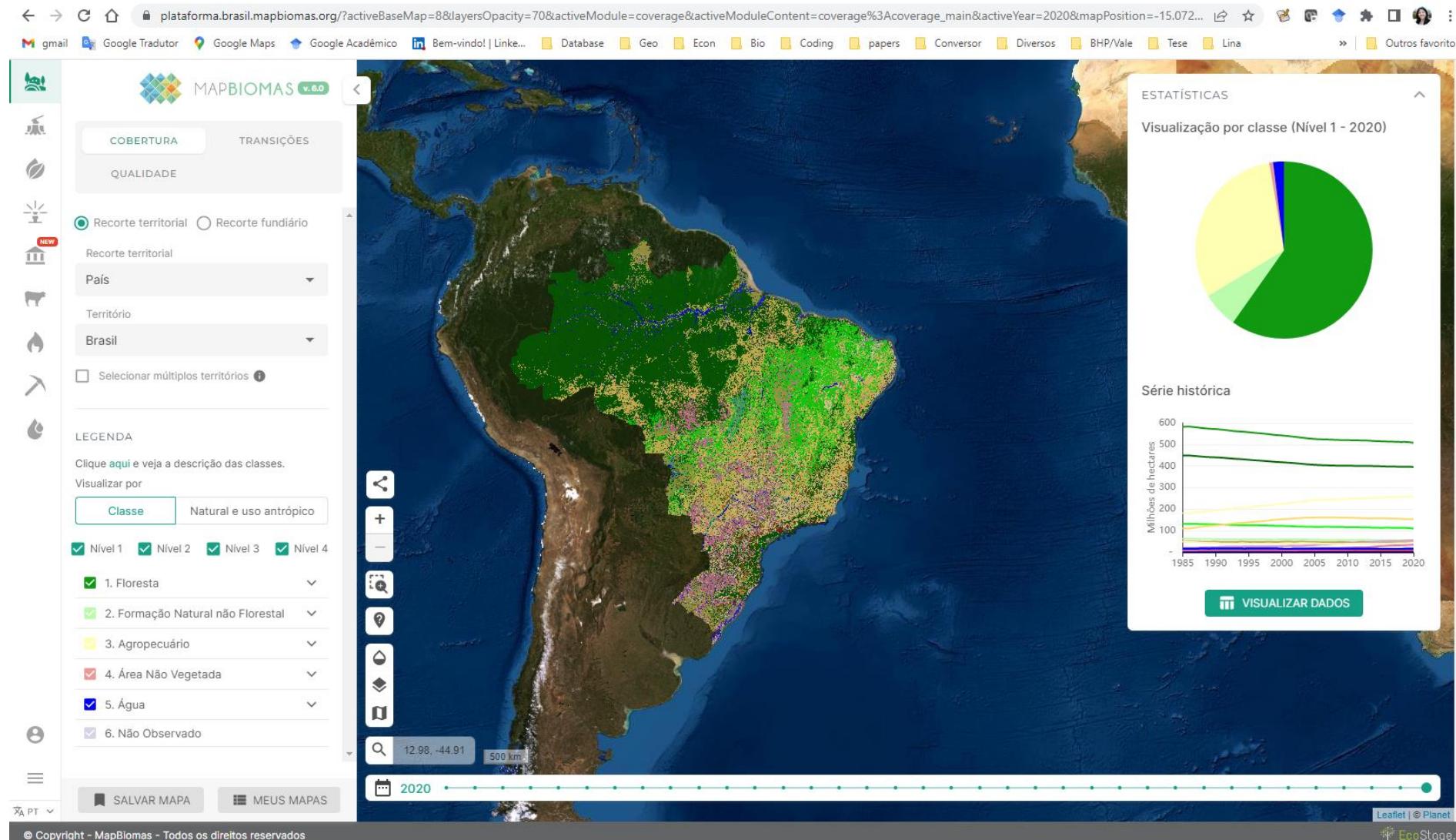
Data

- Panel data of AMCs^[1] (time consistent municipalities), 1993-2020
 - Only AMCs in states with treated *and* untreated (removes Legal Amazon)
 - AMCs in the boundaries (territories partially treated) were removed
- Treatment starts at 2007 (AFL: Dec2006; 1-3y of anticipation)
- Outcome: natural cover (% of AMC area) from Mapbiomas collection 6.0
 - 30m resolution Landsat images
- Covariates (X)
 - State dummies
 - Natural cover in the baseline (1985)
 - Other covar. in the baseline (ln GDP, ln GDP per capita, sectorial value-added, share of urban population, infant mortality rate - deaths per 1,000 live births, Conservation Units coverage)



[1] AMC: minimum comparable areas, aggregates some municipalities to have a time-consistent region

Mapbiomas



NEREUS

The University of São Paulo
Regional and Urban Economics Lab

Ussami & Chimeli, May 2022

10

 **FEAUSP**

Methodology

- Sant'Anna & Zhao (2020): Conditional Parallel Trends Assumption, double robust estimation method (combines regression outcome and inverse probability weighting), 2x2 canonical DiD
- Callaway & Sant'Anna (2020): multiple time periods, doubly robust

for all g and t such that

$g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2 + \delta, 3 + \delta, \dots, T\}$, $t \in \{2, \dots, T - \delta\}$ and $t \geq g - \delta$,
 $ATT(g, t)$ is nonparametrically identified by the DR estimand

$$ATT_{dr}^{nev}(g, t; \delta) = \mathbb{E} \left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]} \right) (Y_t - Y_{g-\delta-1} - m_{g,t,\delta}^{nev}(X)) \right]$$

where $m_{g,t,\delta}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-\delta-1}|X, C = 1]$.

- Averages of the $ATT(g, t)$:

$$\theta_S(g) = \frac{1}{T - g + 1} \sum_{t=2}^T \mathbf{1}\{g \leq t\} ATT(g, t).$$

- Simultaneous confidence interval
- Cluster robust (at amc)
- Based on “never treated” units

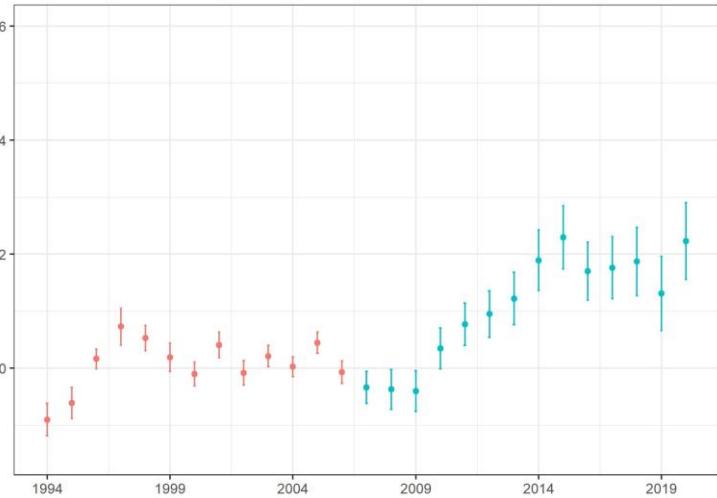
Methodology

- Concurrent incentives to protect other control regions/biomes through other instruments (PPCerrado) -> downward bias
- Different land cover removal pressure in treated and control areas (i.e., Conservation Units) -> downward bias
- Spillover effects (from AFL and from Amazon, same untreated municipalities)
 - No municipalities from the Legal Amazon
 - Robustness: sample restricted to municipalities with less than 25% of natural cover in the baseline
 - Flow outcomes: net revegetation, natural cover loss and natural cover recovery

Effects on natural cover, no anticipation

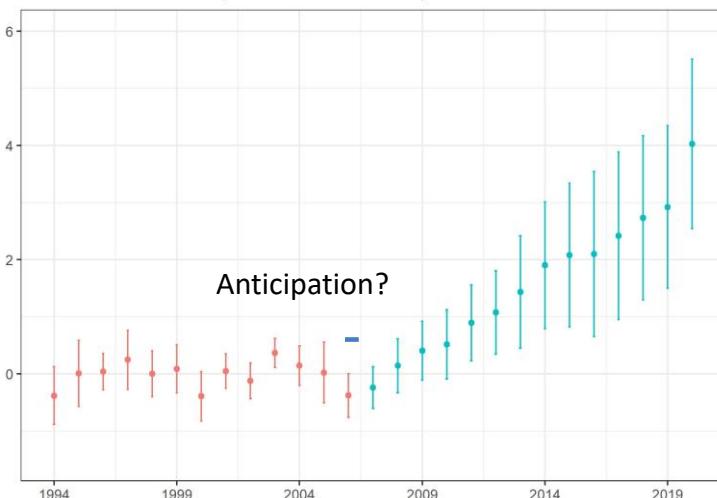
(11): Unconditional PTA

Effect on natural cover, selected states 0-100, 11cs0 2007



(28): State dummies + Baseline nat.cover + Other baseline covar

Effect on natural cover, selected states 0-100, 28cs0 2007



Average effect wih no anticipation

	Selected states, 0-100				
	Dependent variable: Nat.cover (% of AMC area)				
	(11)	(12)	(24)	(28)	
ATT		1.089*** (0.124)	2.457*** (0.141)	1.711*** (0.156)	1.601*** (0.311)
AMC cluster	✓	✓	✓	✓	
State dummy		✓	✓	✓	
Baseline nat.cover			✓	✓	
Baseline control variables			✓	✓	
Anticipation periods	0	0	0	0	
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12	
Qty. of treated AMCs	1661	1661	1661	1661	
Qty. of control AMCs	1461	1461	1461	1461	

Note:

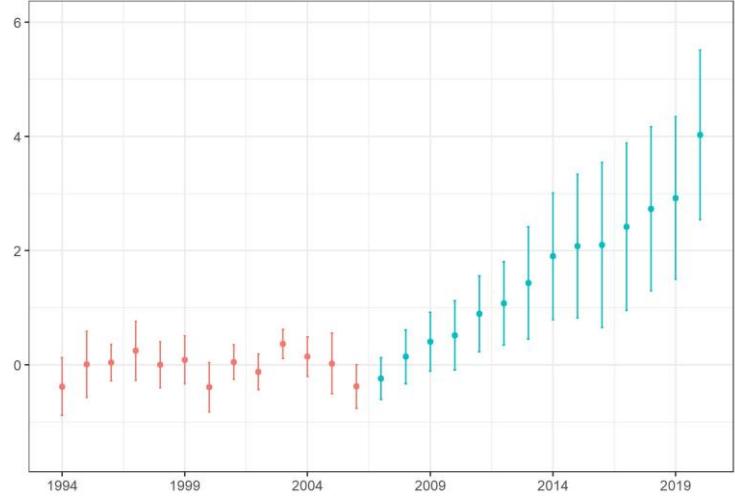
*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

- Positive and significant effects (~1.6pp)
- Suggestive evidence of anticipation related to the timing when the AF bill was being approved in the Senate;

Effects on natural cover

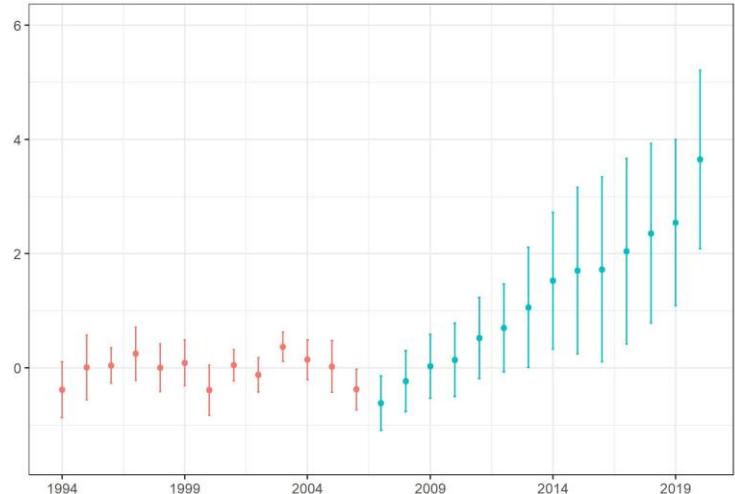
No anticipation

Effect on natural cover, selected states 0-100, 28cs0 2007



1 year anticipation

Effect on natural cover, selected states 0-100, 28cs1 2007



Average effect with different anticipation periods

	Selected states, 0-100				
	Dependent variable: <i>Nat.cover</i> (% of AMC area)				
	(1)	(2)	(3)	(4)	
ATT		1.601*** (0.311)	1.225*** (0.368)	1.250*** (0.447)	1.396*** (0.499)
AMC cluster	✓	✓	✓	✓	
State dummy	✓	✓	✓	✓	
Baseline nat.cover	✓	✓	✓	✓	
Baseline control variables	✓	✓	✓	✓	
Anticipation periods	0	1	2	3	
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12	
Qty. of treated AMCs	1661	1661	1661	1661	
Qty. of control AMCs	1461	1461	1461	1461	

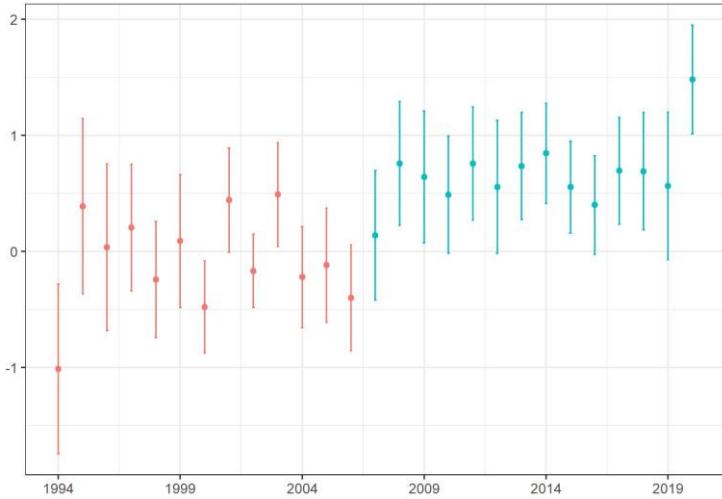
Note:

*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

Effects on net revegetation

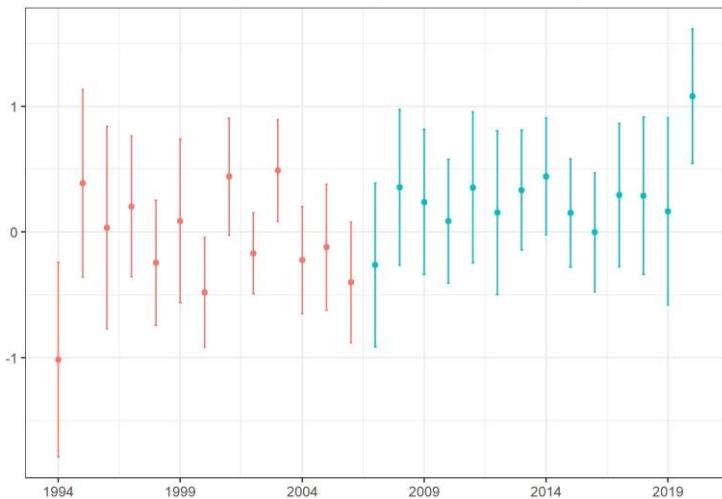
No anticipation

Effect on net nat. cover gain (stock diff), selected states 0-100, 28cs0 2007



1 year anticipation

Effect on net nat. cover gain (stock diff), selected states 0-100, 28cs1 2007



Average effect with different anticipation periods

	Selected states, 0-100			
	Dependent variable: Net natural cover recovery (% of AMC area)			
	(1)	(2)	(3)	(4)
ATT	0.664*** (0.128)	0.262* (0.153)	0.142 (0.125)	-0.082 (0.081)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline nat.cover	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	1	2	3
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

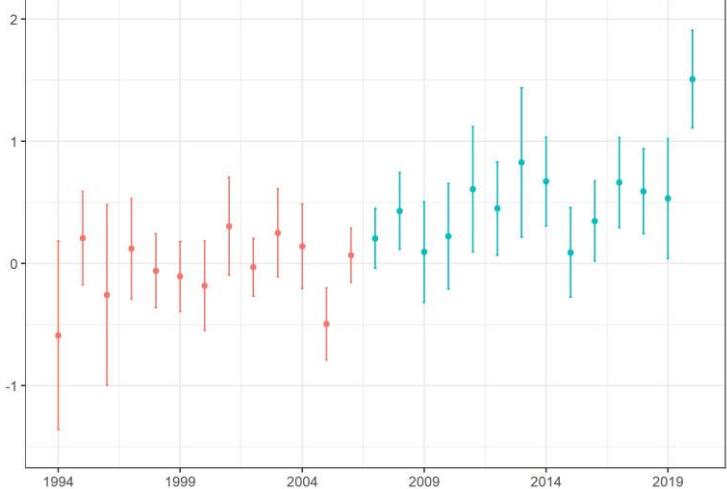
Note:

*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

Effects on norm. net revegetation

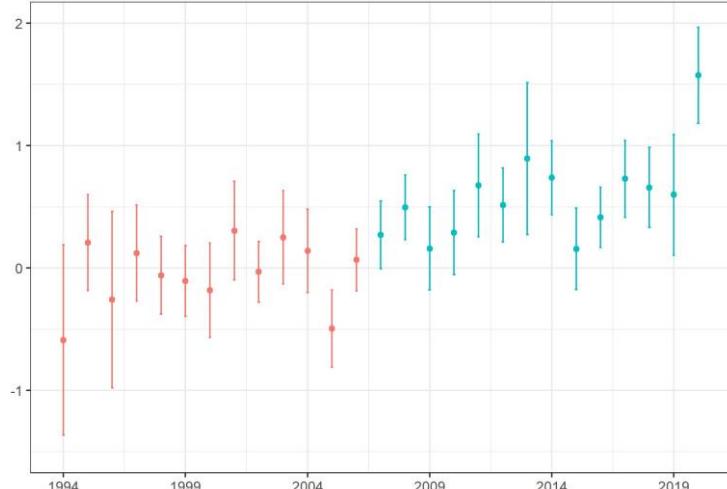
No anticipation

Effect on norm. net nat. cover gain 1, selected states 0-100, 28cs0 2007



1 year anticipation

Effect on norm. net nat. cover gain 1, selected states 0-100, 28cs1 2007



$$\text{Normalized Net Reveg}_{it} = \frac{\text{Net Reveg}_{it} - \bar{\text{Net Reveg}}_i}{\text{sd}(\text{Net Reveg}_i)}$$

Net Revegetation it in hectares
Mean and SD over the period 1993-2020

Average effect with different anticipation periods

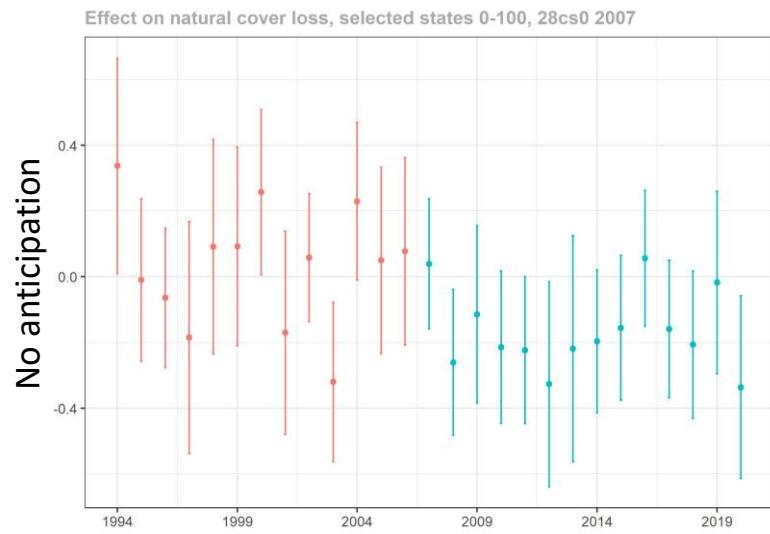
	Selected states, 0-100			
	Dependent variable: Normalized net natural cover recovery			
	(1)	(2)	(3)	(4)
ATT	0.516*** (0.101)	0.583*** (0.069)	0.088 (0.093)	0.228** (0.089)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline nat.cover	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	1	2	3
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

Note:

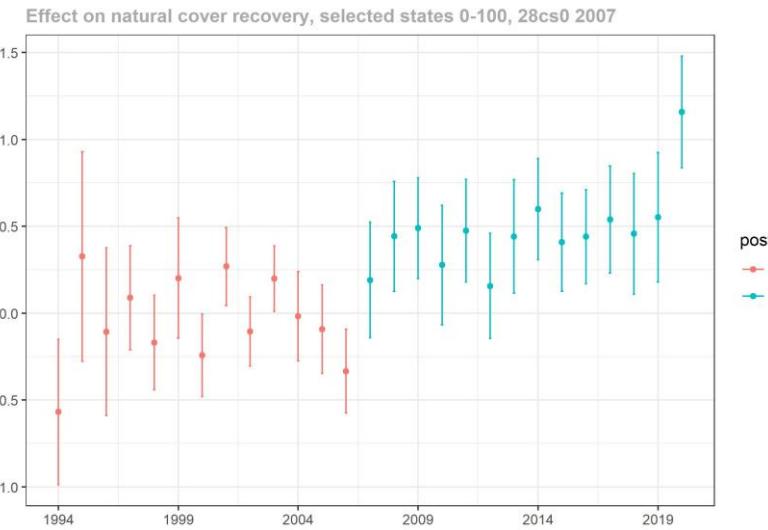
* p<0.1; ** p<0.05; *** p<0.01
Robust standard errors are in parenthesis

Effects on natural cover loss and recovery

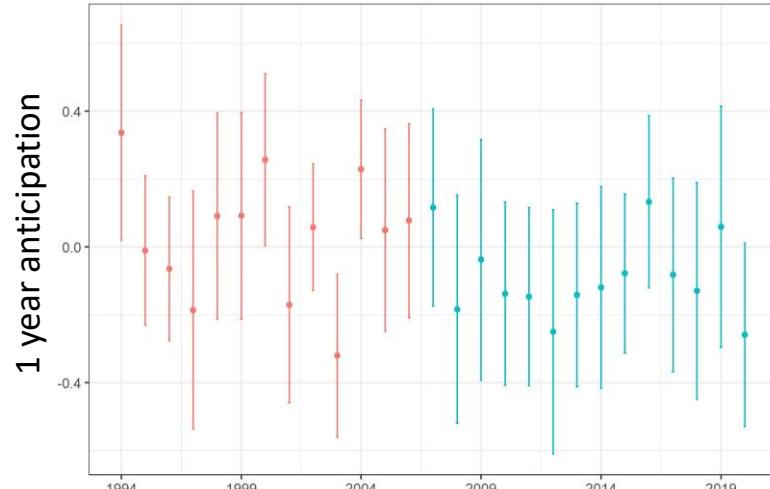
loss



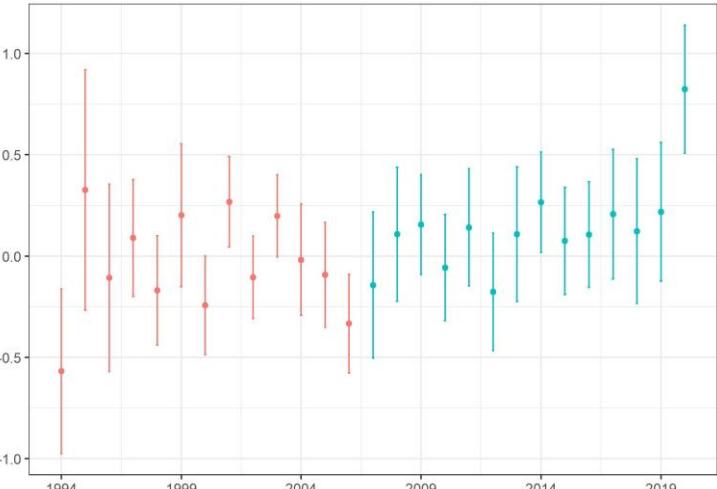
recovery



Effect on natural cover loss, selected states 0-100, 28cs1 2007



Effect on natural cover recovery, selected states 0-100, 28cs1 2007



Effects on natural cover loss and recovery

Average effect with different anticipation periods

Dependent variable	ATT	Selected states, 0-100			
		(1)	(2)	(3)	(4)
(a) Natural cover loss		-0.167*** (0.057)	-0.090 (0.083)	-0.040 (0.058)	0.189*** (0.052)
(b) Natural cover recovery	ATT	0.473*** (0.090)	0.140* (0.080)	0.047 (0.076)	0.029 (0.047)
AMC cluster		✓	✓	✓	✓
State dummies		✓	✓	✓	✓
Baseline nat. cover		✓	✓	✓	✓
Baseline control variables		✓	✓	✓	✓
Anticipation periods		0	1	2	3
Baseline nat. cover in treated AMCs (2006)		30.12	30.12	30.12	30.12
Qty. of treated AMCs		1661	1661	1661	1661
Qty. of control AMCs		1461	1461	1461	1461

Note:

* p < 0.1; ** p < 0.05; *** p < 0.01

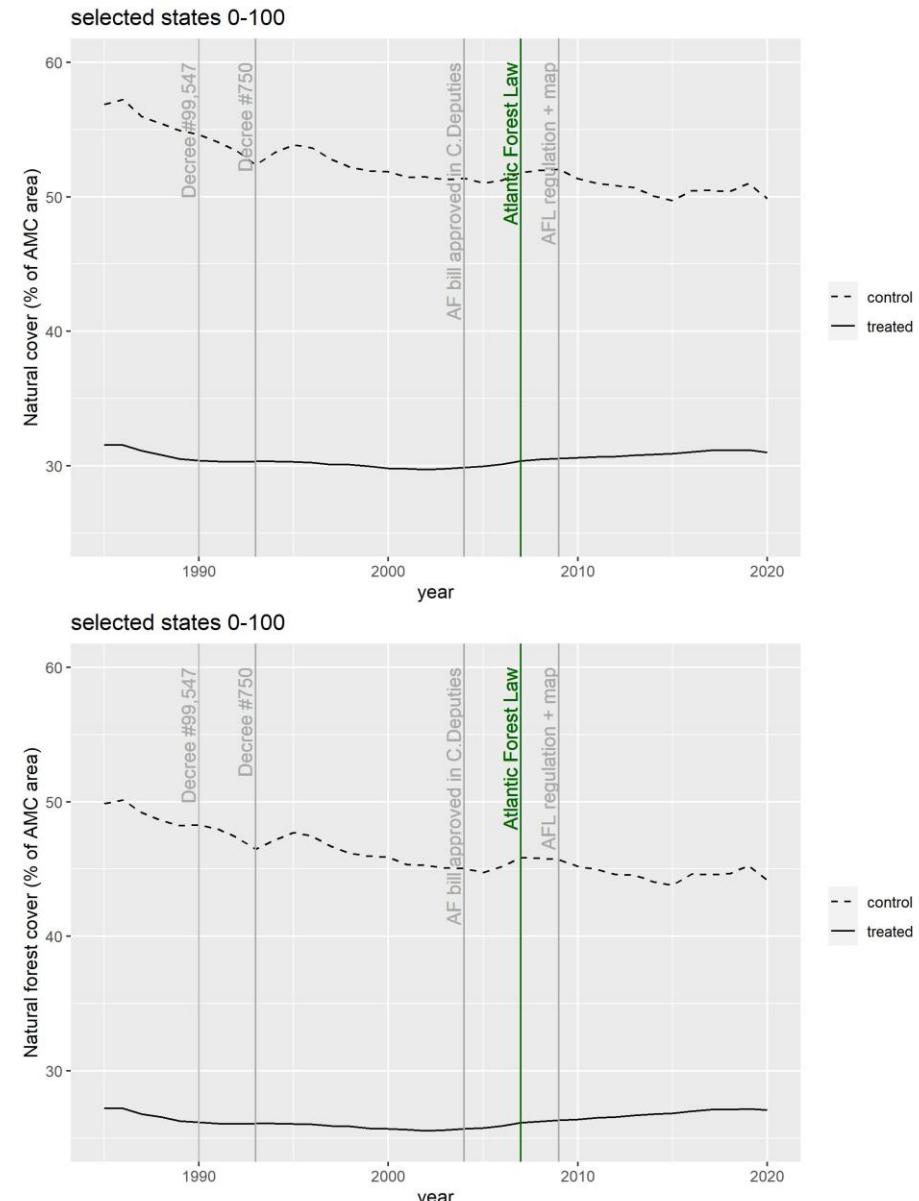
Robust standard errors are in parenthesis

Flow variables:

- Positive and significant effects for no anticipation case. Effects from natural cover loss and recovery.
- No effects when takes anticipation into account.

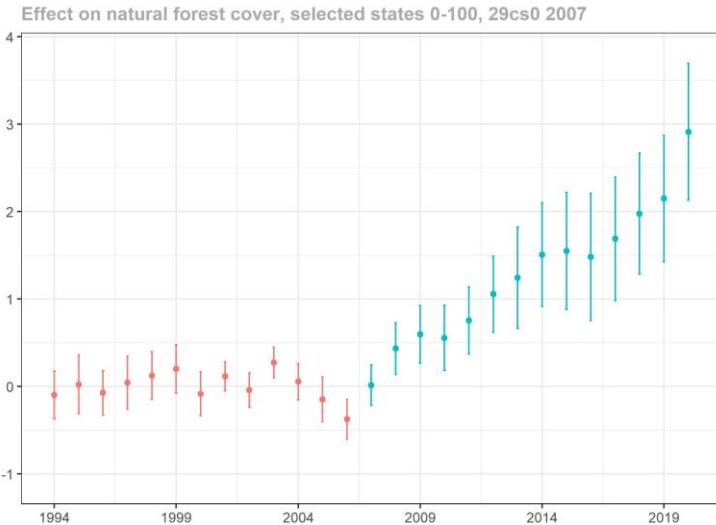
Robustness check using natural forest cover

- Natural cover includes different vegetation types. Mainly forests, but wetlands and grasslands are included.
- Different dynamics related to different vegetation composition (and their determinants)? (e.g. Pampa grasslands that may be used as natural pasture)
- Different pattern of classification error
- Natural forest cover as outcome for robustness check
- Positive and significant effects
- Lower effects (~1.3pp)

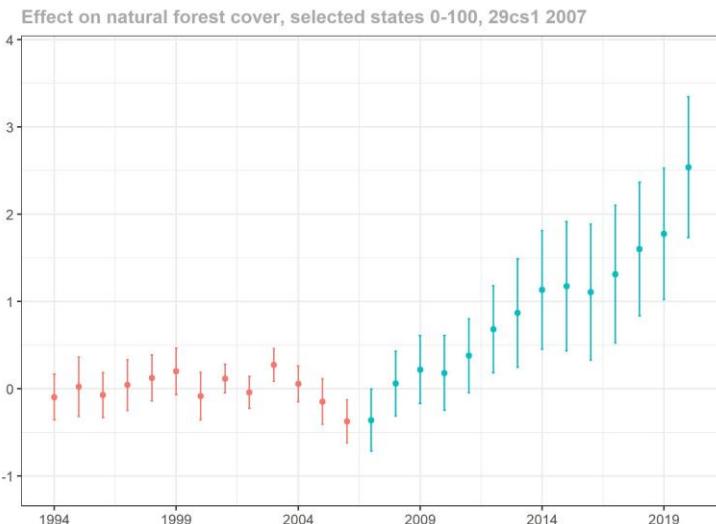


Robustness check using natural forest cover

No anticipation



1 year anticipation



Average effect with different anticipation periods

Selected states, 0-100				
Dependent variable: Nat.Forest (% of AMC area)				
	(1)	(2)	(3)	(4)
ATT	1.279*** (0.159)	0.905*** (0.170)	0.758*** (0.235)	0.814*** (0.258)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline nat.cover	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	1	2	3
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

Note:

*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

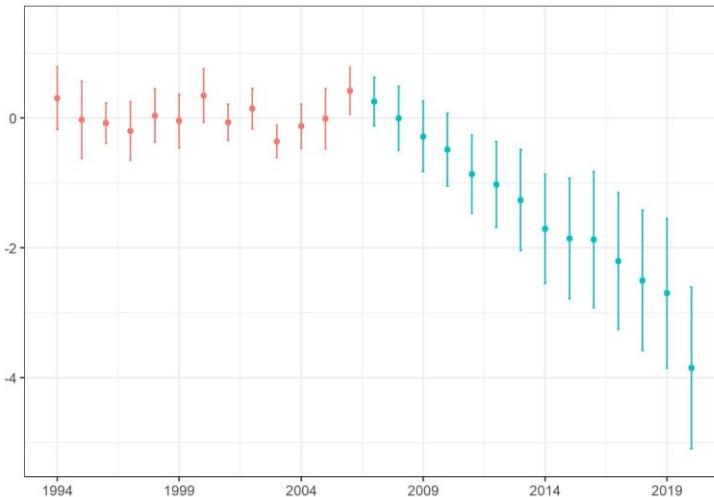
Robustness check using anthropic land cover

- Farming land cover and its disaggregated components (pasture, crops, silviculture, and mosaic of crops and pasture); urban land cover
- Farming cover
 - Negative and significant effects (as expected)
 - Effects coming mainly from pasture

Robustness check using anthropic land cover: farming cover

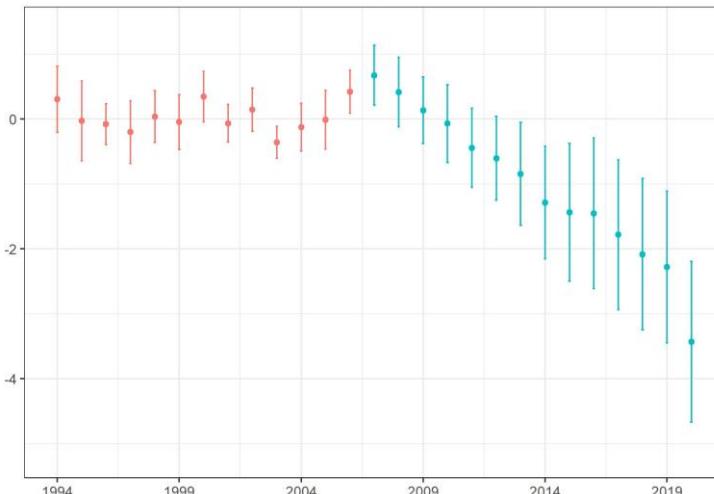
Farming cover, no anticipation

Effect on farming, selected states 0-100, 28cs0 2007



Farming cover, 1 year anticipation

Effect on farming, selected states 0-100, 28cs1 2007



Aggregate effect of Atlantic Forest Law on different types of farming cover

	Selected states, 0-100 Dependent variable:						
	Farming		Pasture	Crops	Silviculture	Mosaic pasture/crops	Urban
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ATT	-1.454*** (0.231)	-1.035*** (0.255)	-1.202 (0.762)	0.000 (0.390)	-0.436 (0.273)	0.183 (0.488)	-0.156 (0.136)
AMC cluster	✓	✓	✓	✓	✓	✓	✓
State dummies	✓	✓	✓	✓	✓	✓	✓
Baseline nat. cover	✓	✓	✓	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓	✓	✓	✓
Anticipation periods	0	1	0	0	0	0	0
Baseline nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461	1461	1461	1461

Note:

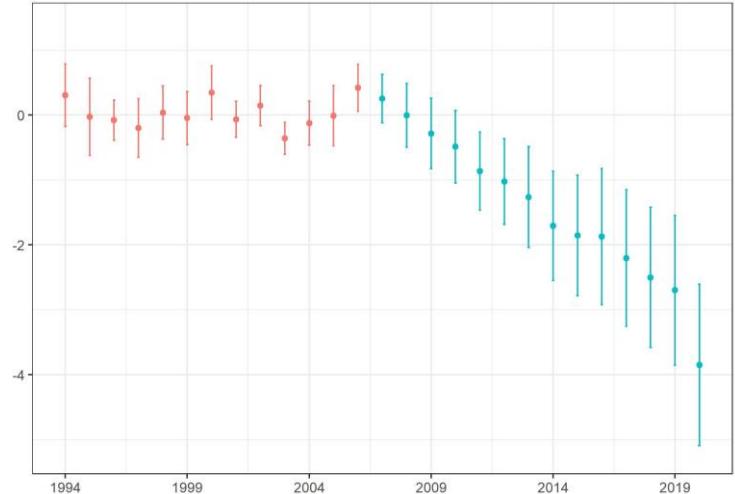
* p < 0.1; ** p < 0.05; *** p < 0.01

Robust standard errors are in parenthesis

Robustness check using anthropic land cover

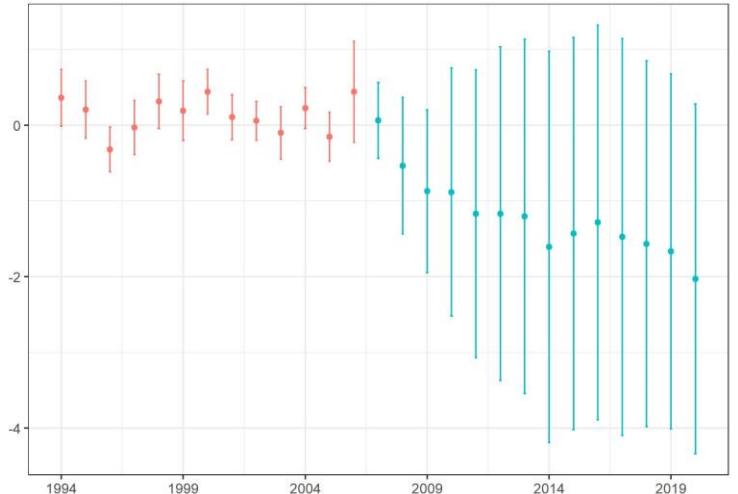
Farming cover

Effect on farming, selected states 0-100, 28cs0 2007



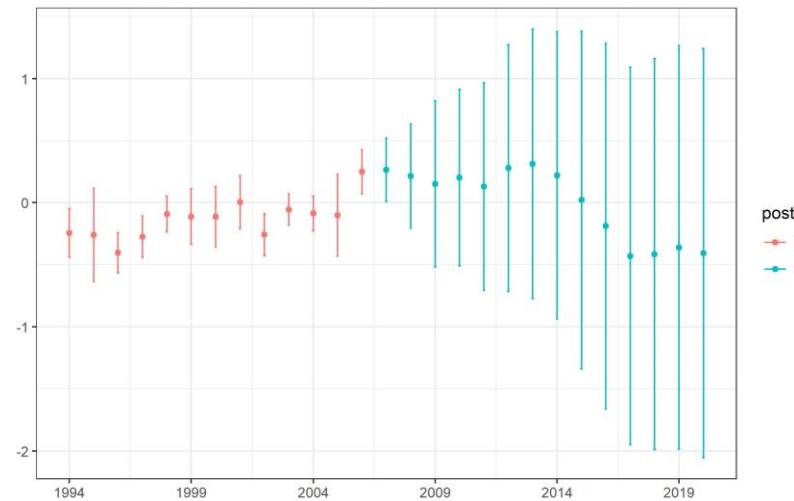
Pasture

Effect on pasture cover, selected states 0-100, 28cs0 2007



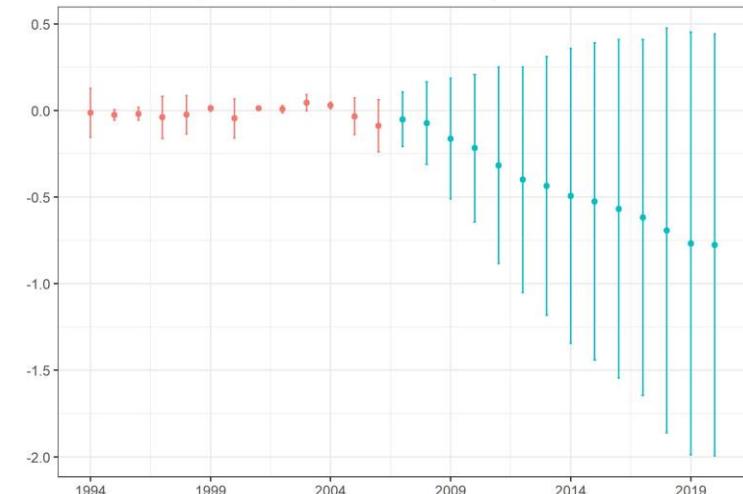
Crops

Effect on crops, selected states 0-100, 28cs0 2007



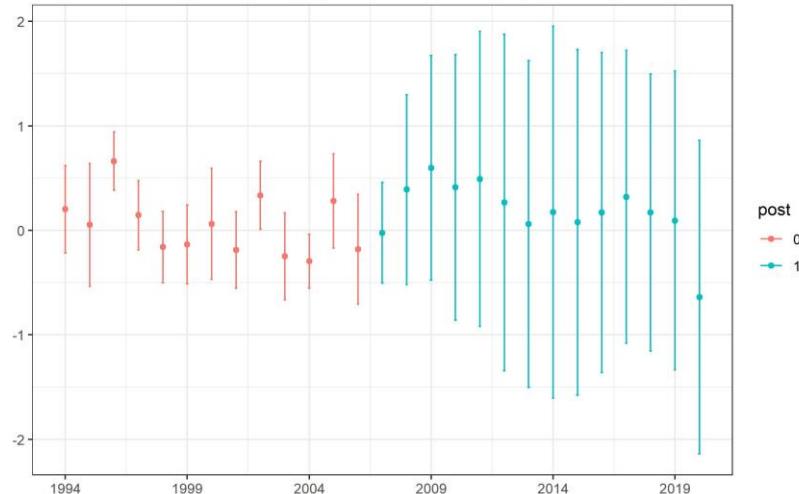
Forest plantation

Effect on forest plantation, selected states 0-100, 28cs0 2007



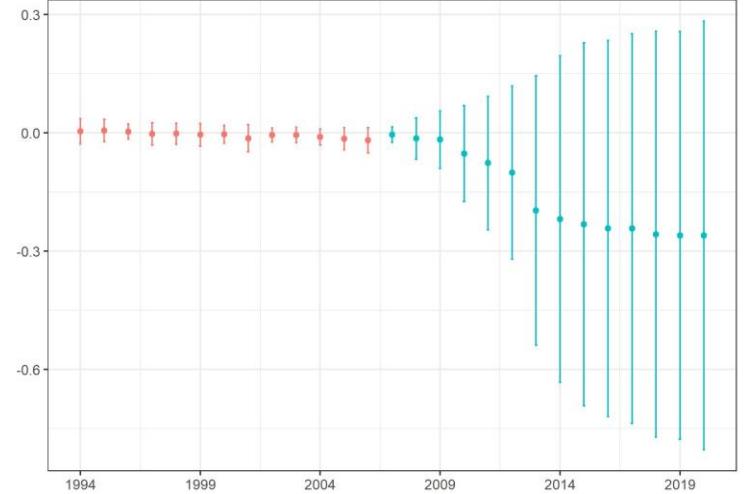
Mosaic of pasture/crops

Effect on mosaic (pasture/crops), selected states 0-100, 28cs0 2007



Urban

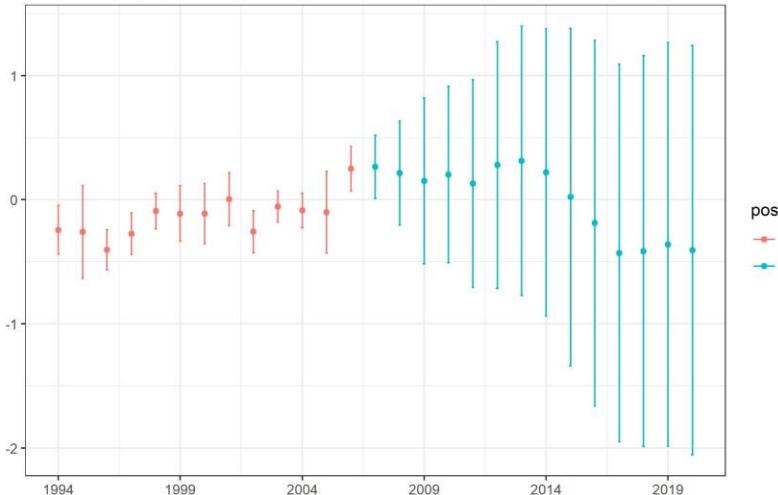
Effect on urban cover, selected states 0-100, 28cs0 2007



Robustness check using anthropic land cover: crops

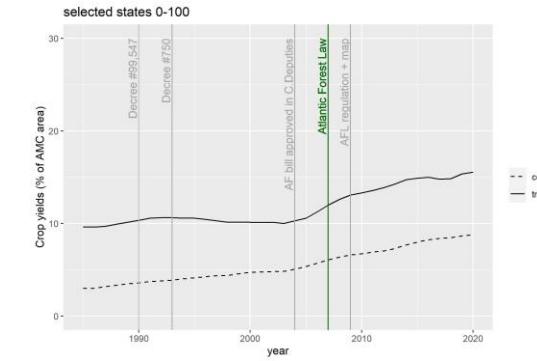
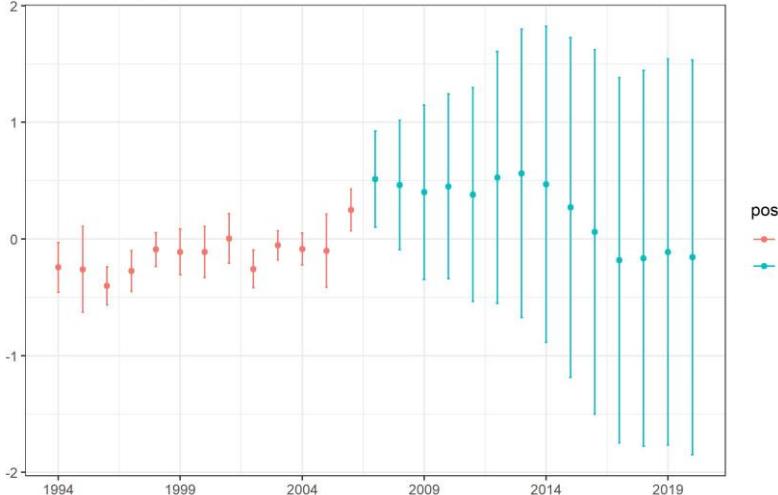
No anticipation

Effect on crops, selected states 0-100, 28cs0 2007



1 year anticipation

Effect on crops, selected states 0-100, 28cs1 2007



Average effect with different anticipation periods

	Selected states, 0-100			
	Dependent variable: Crops cover (% of AMC area)			
	(1)	(2)	(3)	(4)
ATT	0.000 (0.390)	0.249 (0.417)	0.148 (0.398)	0.062 (0.436)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline nat.cover	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	1	2	3
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

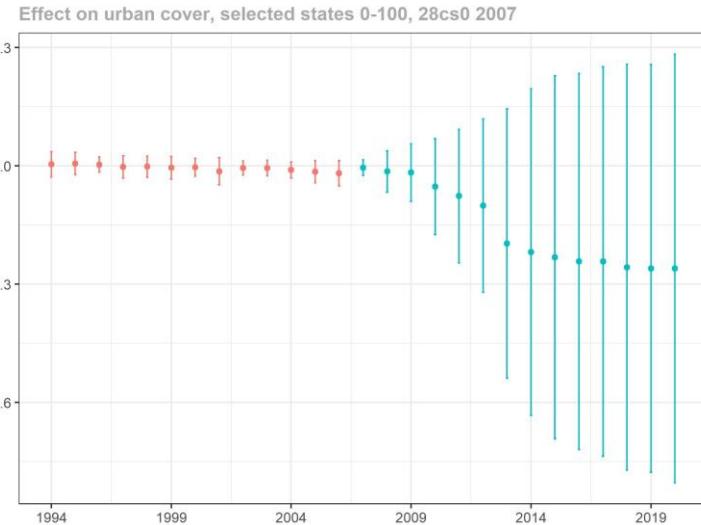
Note:

* p<0.1; ** p<0.05; *** p<0.01
Robust standard errors are in parenthesis

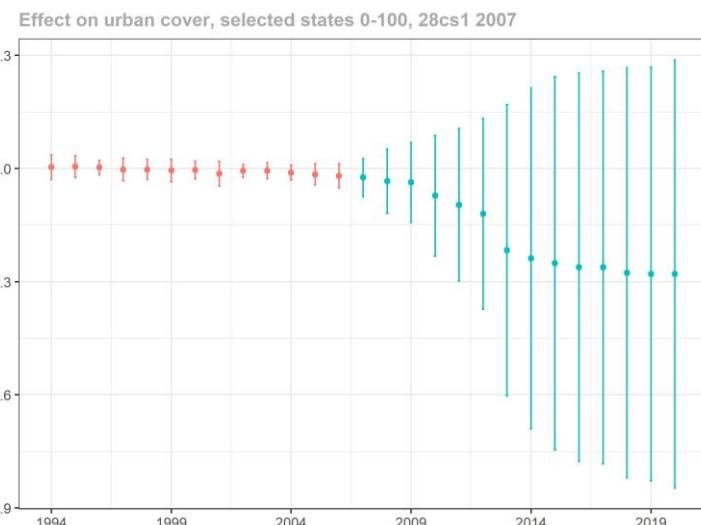
- Agricultural credit is probably not related to the observed effects

Robustness check using anthropic land cover: urban cover

No anticipation



1 year anticipation



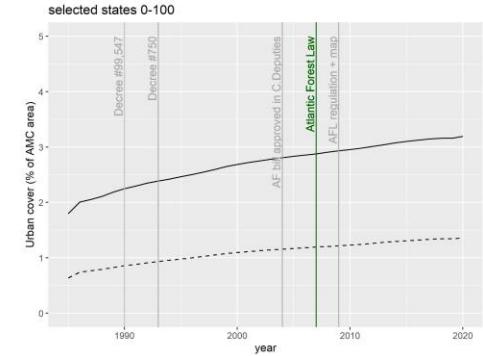
Average effect with different anticipation periods

Selected states, 0-100				
Dependent variable: Urban Cover (% of AMC area)				
	(1)	(2)	(3)	(4)
ATT	-0.156 (0.136)	-0.175 (0.147)	-0.190 (0.158)	-0.201 (0.165)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline nat.cover	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	1	2	3
Nat. cover in treated AMCs (2006)	30.12	30.12	30.12	30.12
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

Note:

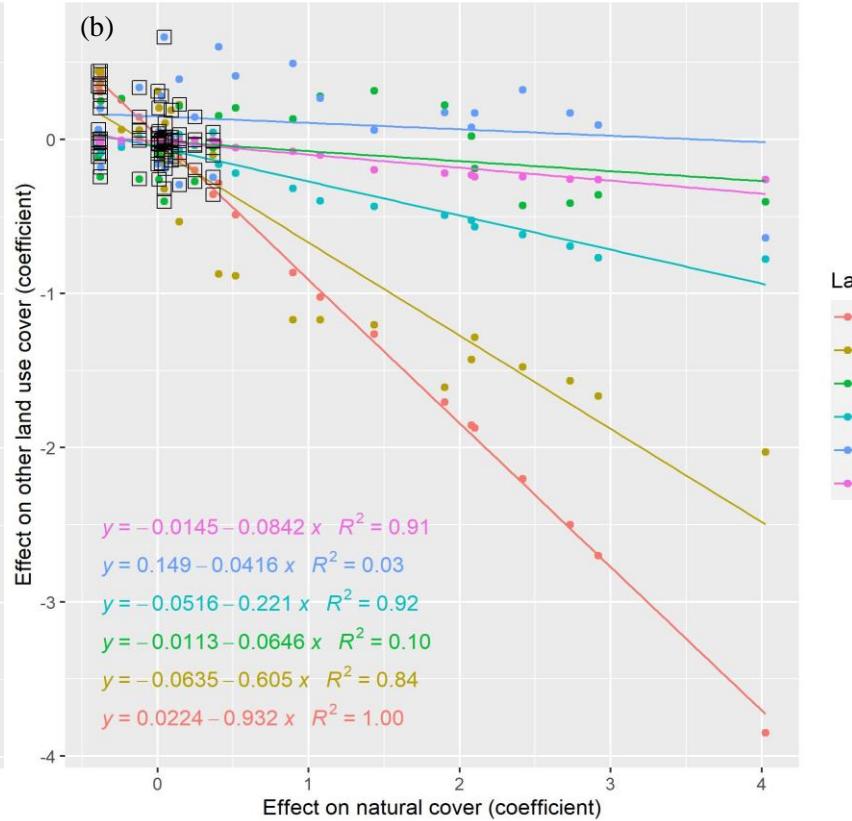
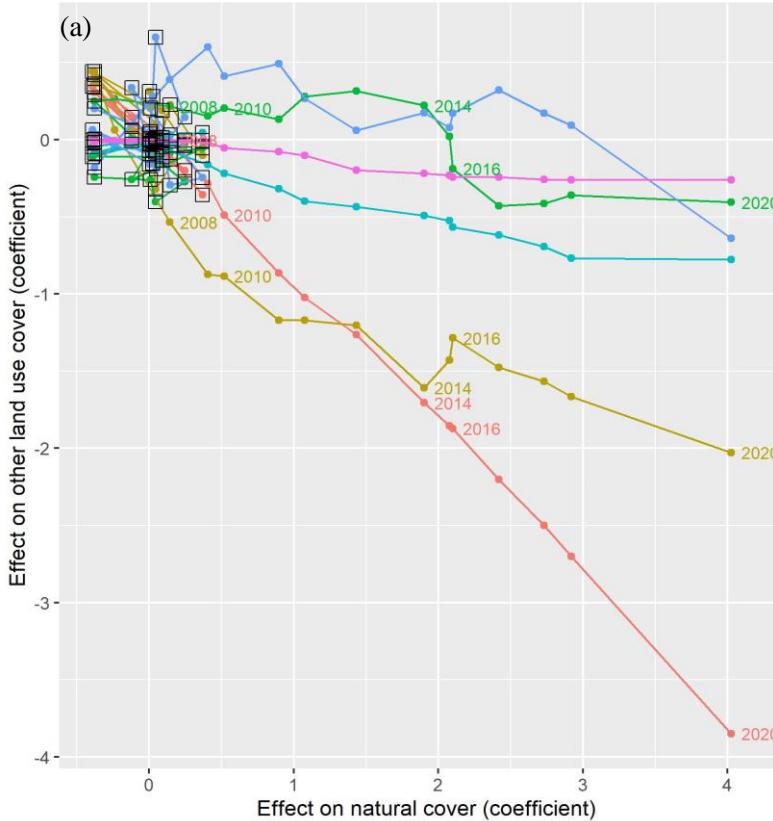
*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

- Urban areas are also associated with deforestation (SOS Mata Atlântica, 2022), but our results suggest a relative reduction in urban cover



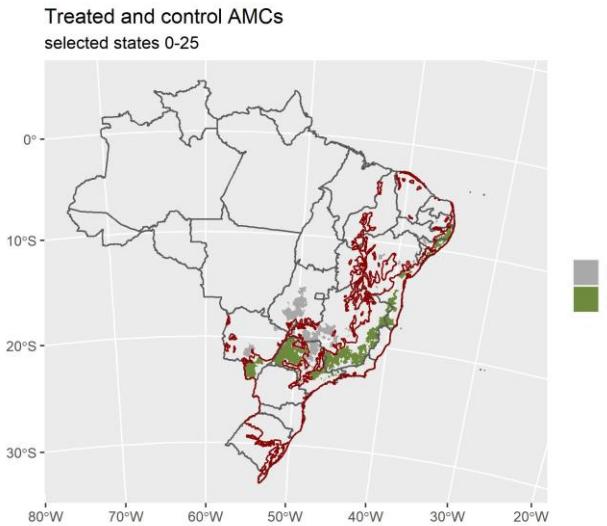
Robustness check using anthropic land cover

Selected states, 0-100
(Pre intervention years in red boxes)



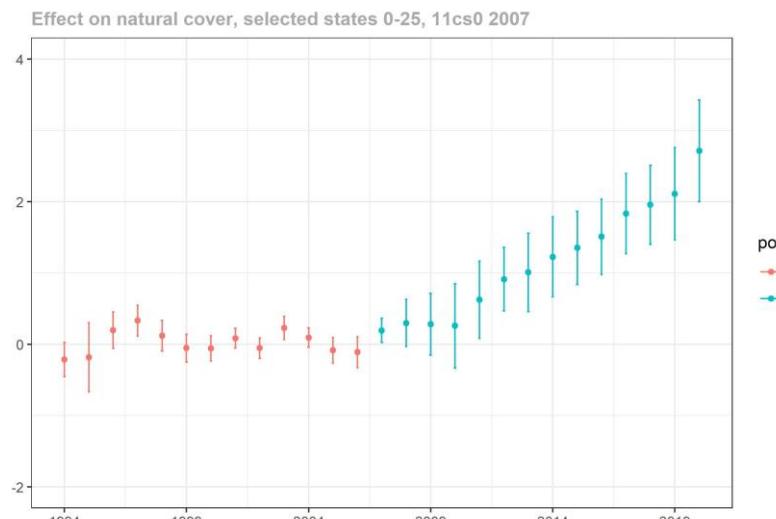
- Relative reduction in farming in treated municipalities occurred as the natural cover in these areas was increasing (almost of the same magnitude)
- Farming reduction came mainly from the decrease in the areas dedicated to pasture and silviculture

Robustness check using AMCs with < 25% of natural cover in the baseline

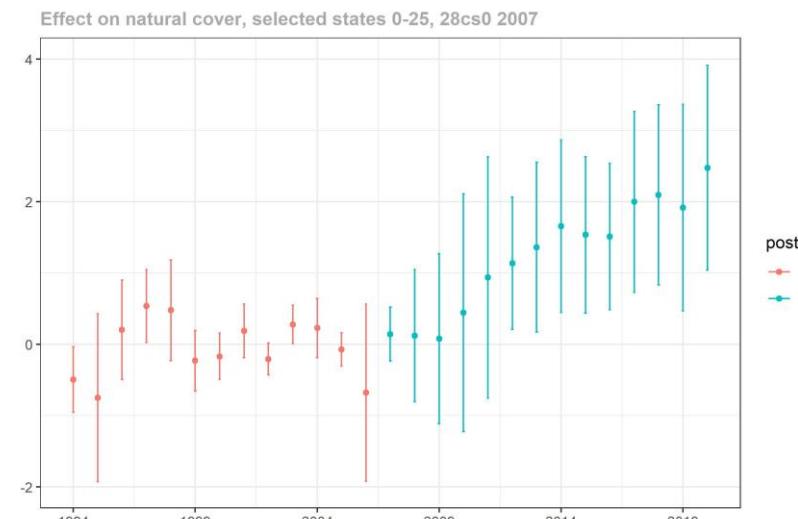


- Rural properties in these AMCs are equally constrained by the Forest Code (scarce areas for legal vegetation removal).
- Possible differences related to the exploration of logging in native forests (which might be different in different biomes) are reduced (tackles the possible bias coming from the common untreated area between AFL and protection policies in Amazon.)

(11): Unconditional PTA



(28): State dummies + Baseline nat.cover + Other baseline covar



Conclusions

- AFL had positive and significant effects on the natural cover, mainly at the expense of the farming cover
- Increases in natural cover occur through the avoidance of natural cover removal and the recovery process, with positive annual net revegetation after the law => progressive rejuvenation of native forest cover
- Effects on nat. cover and net revegetation start in the first year of treatment
- Anticipation of 1-3 years: effects on stock outcome persist with anticipation, but some effects from flow outcomes are lost
- Analysis of heterogeneity:
 - positive effects on the recovery process were stronger than the reduction in vegetation removal
 - most of the results come from the states of SP, MG and BA and removing municipalities from the reference specification neutralizes these results

Conclusions

- Ref. specification
 - Accum. increase of 4.0 pp up to 2020 in natural cover = relative increase of 2.6 Mha
 - 2007 to 2020:these municipalities experienced an increase in the natural cover stock of 0.2 Mha
- Nat. *forest* cover with 1 anticipation period
 - Accum. increase of 2.4 pp up to 2020 in natural forest cover = relative increase of 1.5 Mha
 - 2007 to 2020, these municipalities experienced an increase in the natural cover stock of 0.5 Mha
- Case of success?
 - draft bill in the Chamber of Deputies: to regulate native vegetation protection in Pantanal biome
 - state's characteristics (which includes physical, environmental, social and institutional characteristics) are important
 - engagement from the civil society
 - not enough to prevent old native vegetation clearance

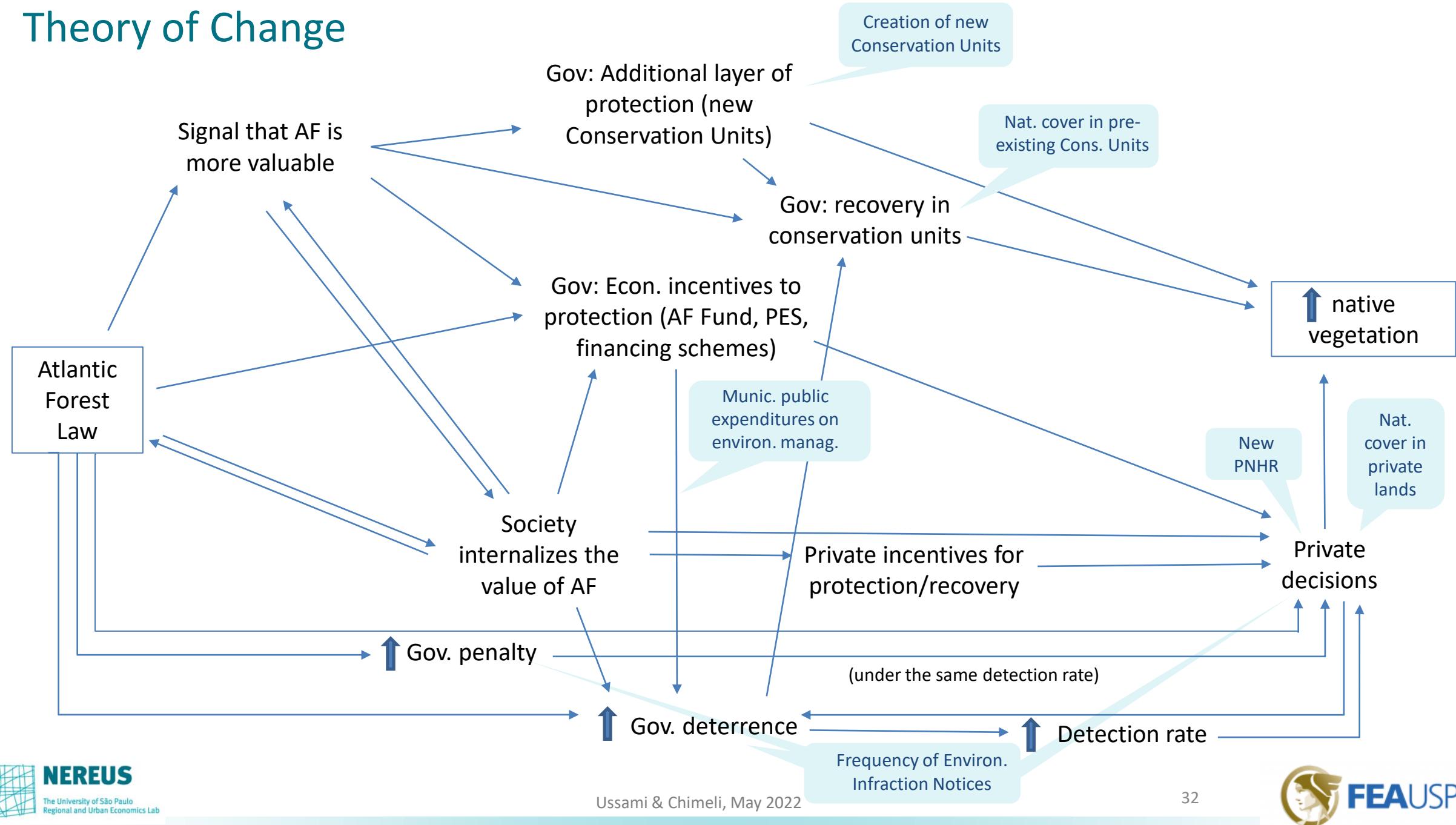
PART II

Exploring the mechanisms behind the Atlantic Forest conservation

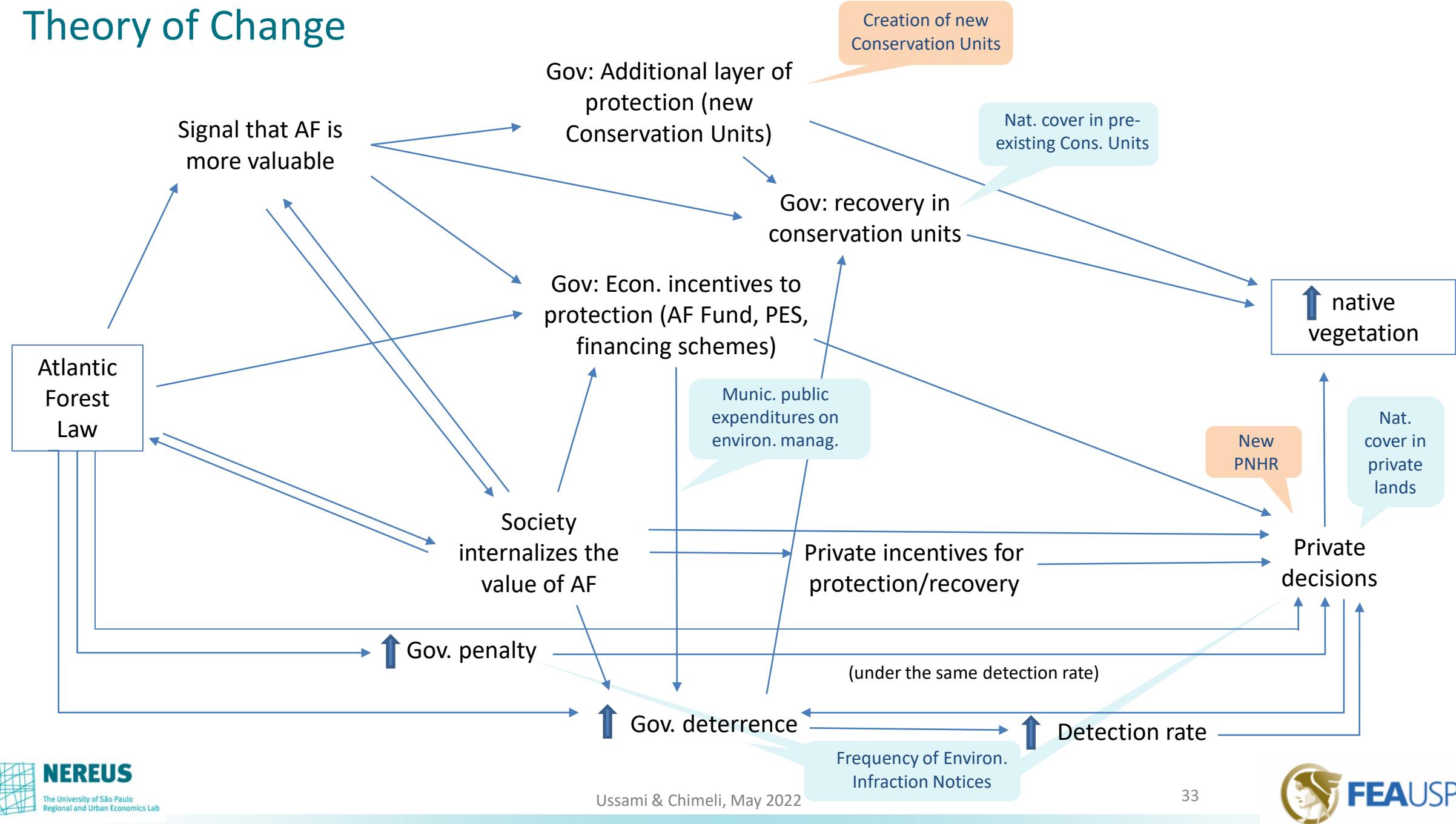
Outline

- Conservation Units:
 - Creation of new Conservation Units
 - Natural cover in pre-existing Conservation Units
- Private Lands
 - Natural cover in private lands
 - Public expenditures in environ. management by municipalities
 - Frequency of Environ. Infraction Notices

Theory of Change

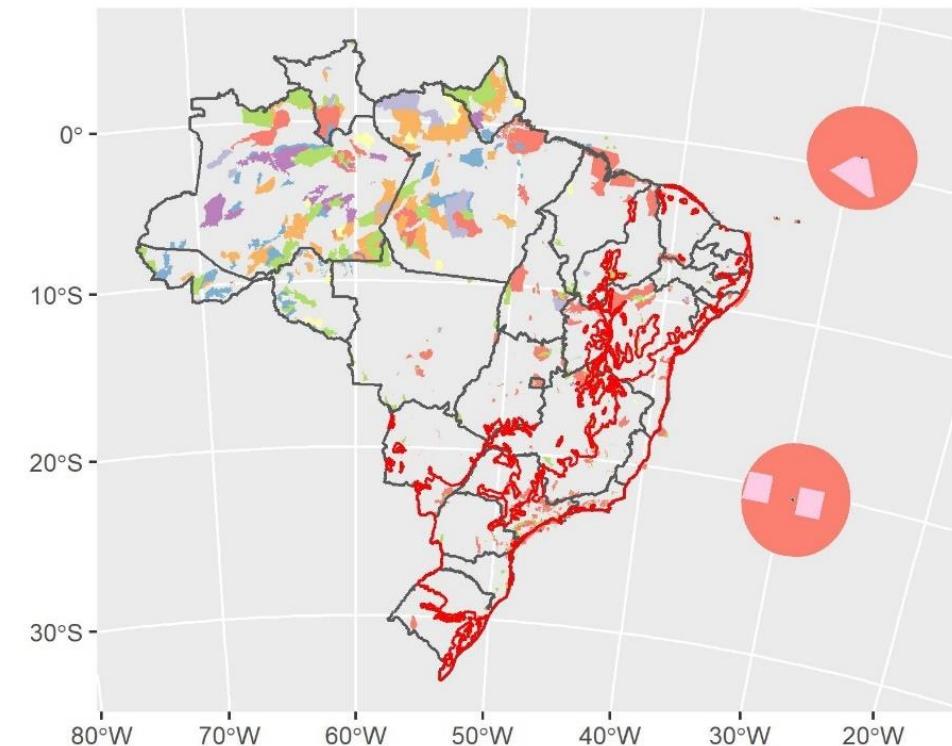
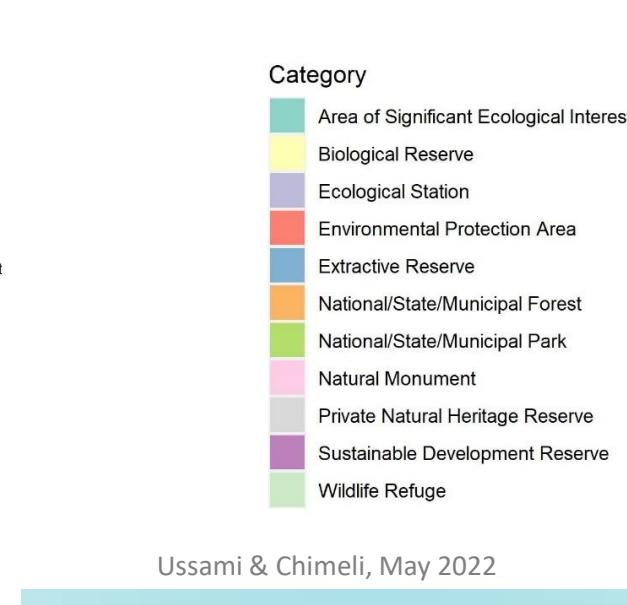
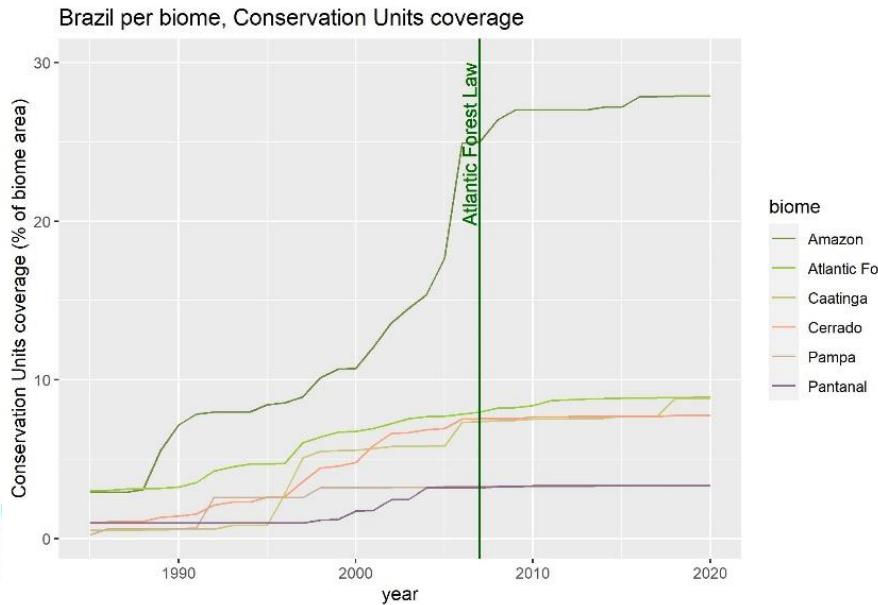


Theory of Change



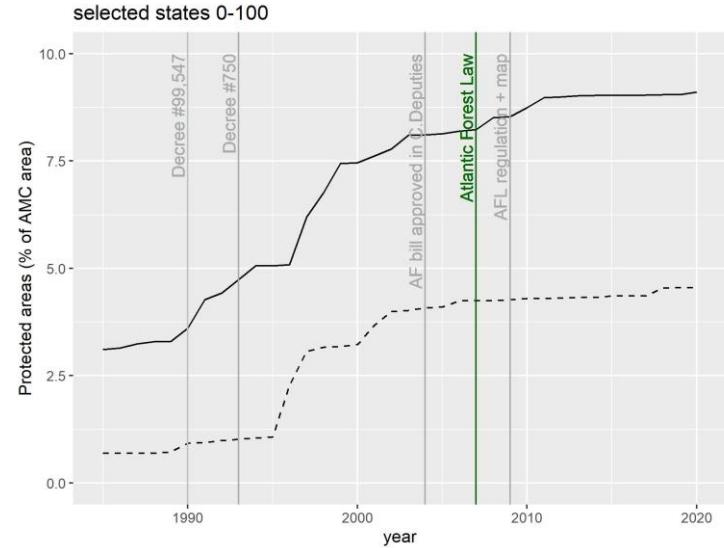
Creation of new Conservation Units

- Creation of new Conservation Units as environmental policies
 - Convention on Biological Diversity target: 17% by 2020
 - The most commonly used tool for biodiversity conservation in developing countries (Miteva et al. 2012)
 - Mechanism to fight deforestation (PPCDAm, PPCerrado)
 - Response to fiscal transfer (Ruggiero et al. 2022)
- Did the AFL affect the creation of new UC?



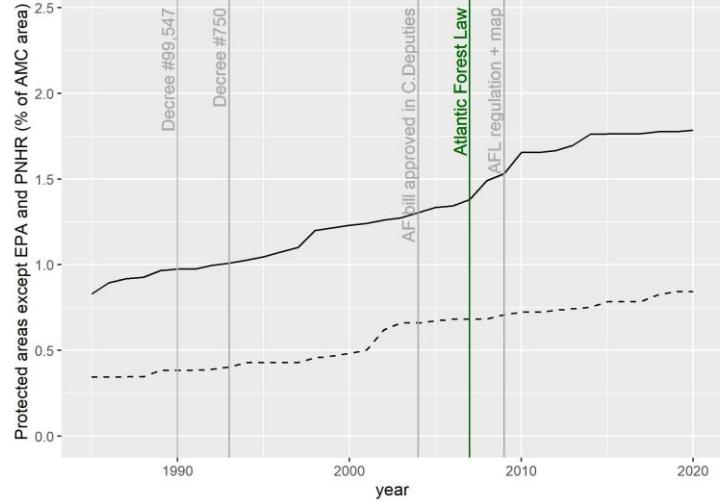
Creation of new Conservation Units

All categories of protected areas



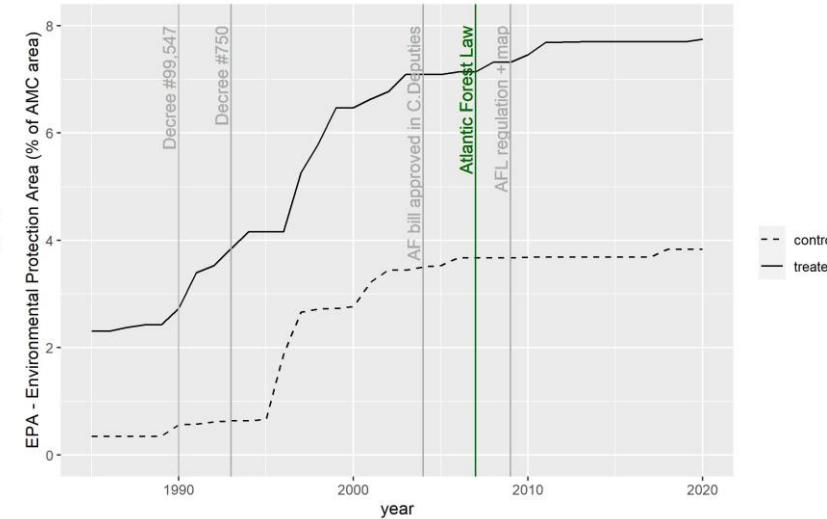
Except EPAs and PNRs

selected states 0-100



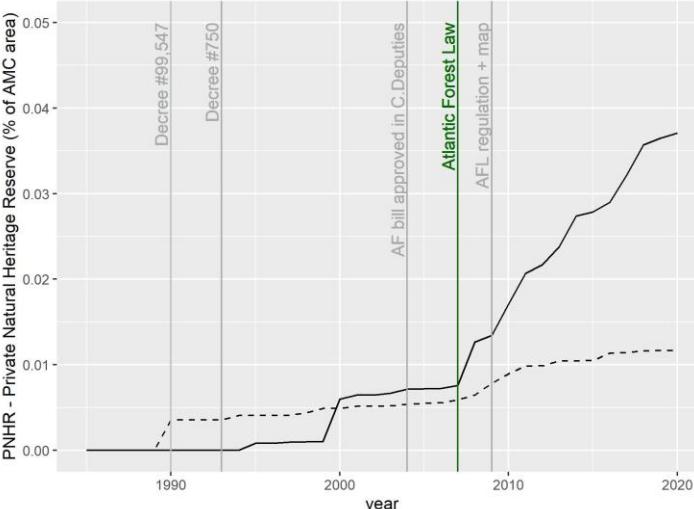
EPA only

selected states 0-100



PNHR only

selected states 0-100



- EPA's share in area protected by Conservation Units: highest

Creation of new Conservation Units (stock)

	Selected states, 0-100			
	Dependent variable: Conservation Units (% of AMC area)			
	(1) All	(2) Except EPA and PNHR	(3) EPA	(4) PNHR
ATT	0.524** (0.196)	0.321*** (0.078)	0.295** (0.177)	0.020*** (0.006)
AMC cluster	✓	✓	✓	✓
State dummies	✓	✓	✓	✓
Baseline Conservation Units	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	0	0	0
Baseline Conservation Units coverage in treated AMCs (2006)	8.19	1.34	7.14	0.007
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

Note:

* p < 0.1; ** p < 0.05; *** p < 0.01

Robust standard errors are in parenthesis

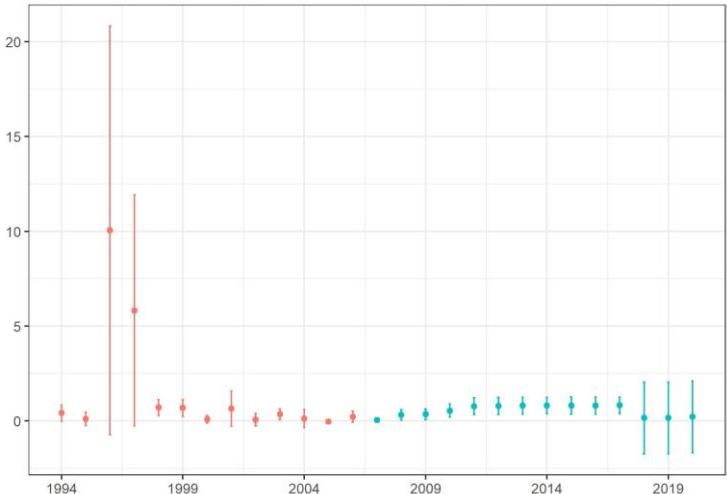
EPA: Area of Environmental Protection; PNHR: Private Natural Heritage Reserve

- After 2006, the % of AMCs protected by Conservation Units in the Atlantic Forest increased (relative to untreated areas).

Creation of new Conservation Units

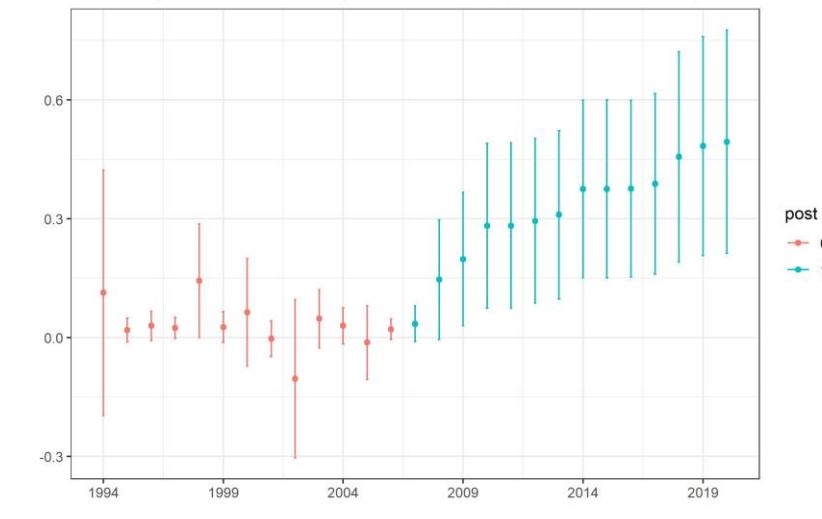
All categories of protected areas

Effect on protected areas, selected states 0-100, 28cs0 2007



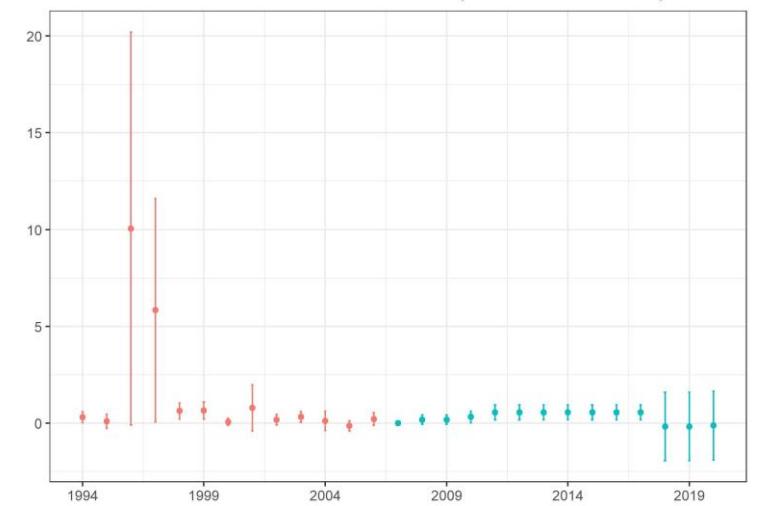
Except EPAs and PNHRs

Effect on protected areas except EPA and PNHR, selected states 0-100, 28cs0 2007



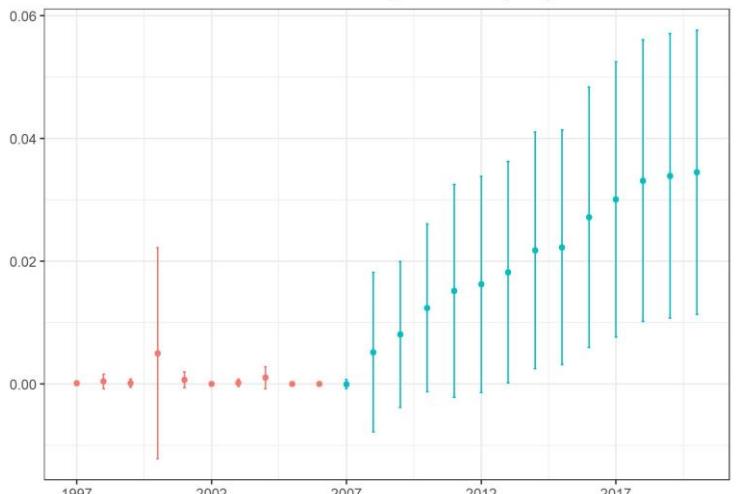
EPA only

Effect on EPA - Environmental Protection Area, selected states 0-100, 28cs0 2007



PNRH only

Effect on PNHR - Private Natural Heritage Reserves (area), selected states 0-100, 28cs0



- All categories and EPA: effects disappear after 2018

First RPPN were created in 1990. Data from before 1996 were ignored in order to keep common support for all covariates

37

Creation of new Conservation Units (robustness: except municipal)

	Selected states, 0-100			
	<i>Dependent variable: Conservation Units, except municipal (% of AMC area)</i>			
	(1) All	(2) Except EPA and PNHR	(3) EPA	(4) PNHR
ATT	0.521*** (0.188)	0.308*** (0.076)	0.293* (0.178)	0.020*** (0.007)
AMC cluster	✓	✓	✓	✓
State dummy	✓	✓	✓	✓
Baseline Conservation Units	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓
Anticipation periods	0	0	0	0
Protected lands in treated AMCs (2006)	8.05	1.34	6.99	0.007
Qty. of treated AMCs	1661	1661	1661	1661
Qty. of control AMCs	1461	1461	1461	1461

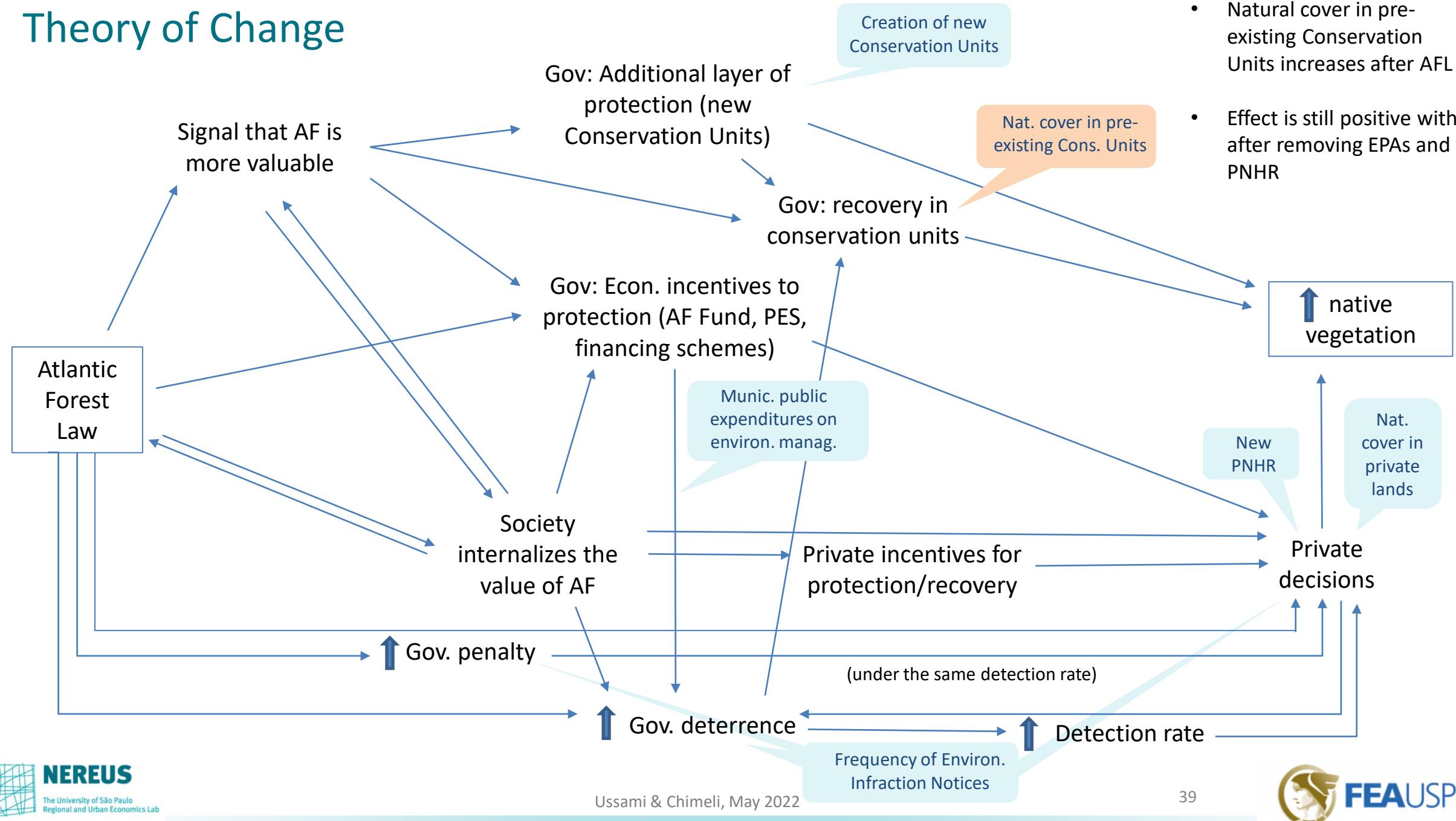
Note:

* p<0.1; ** p<0.05; *** p<0.01

Robust standard errors are in parenthesis

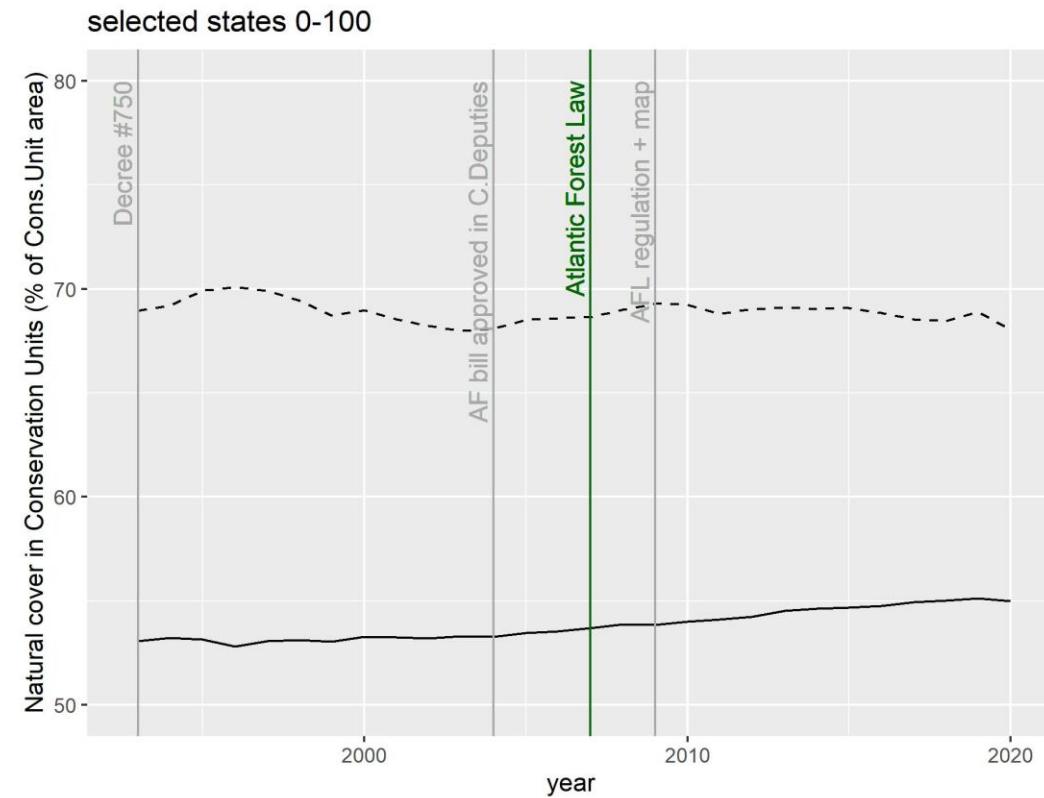
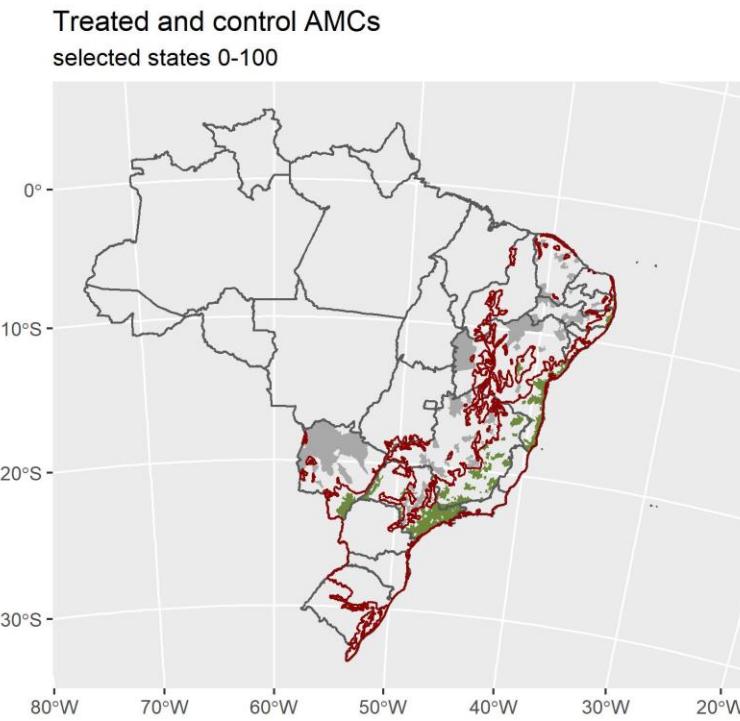
EPA: Area of Environmental Protection; PNHR: Private Natural Heritage Reserve

Theory of Change



Nat. cover in pre-existing protected areas

Outcome: natural cover inside Conservation Units (as % of Conservation Unit area inside AMC)



Nat. cover in pre-existing protected areas

	Selected states, 0-100			
	Dependent variable: Nat.cover (% of Cons. Unit area in AMC)			
	(11)	(12)	(34)	(35)
ATT	0.639*	1.053***	0.900**	1.672***
	(0.380)	(0.387)	(0.404)	(0.462)
AMC cluster	✓	✓	✓	✓
State dummy		✓	✓	✓
Baseline nat.cover in Cons.Units			✓	✓
Baseline control variables				✓
Anticipation periods	0	0	0	0
Nat. cover in Cons.Units in treated AMCs (2006)	53.55	53.55	53.55	53.55
Qty. of treated AMCs	416	416	416	416
Qty. of control AMCs	164	164	164	164

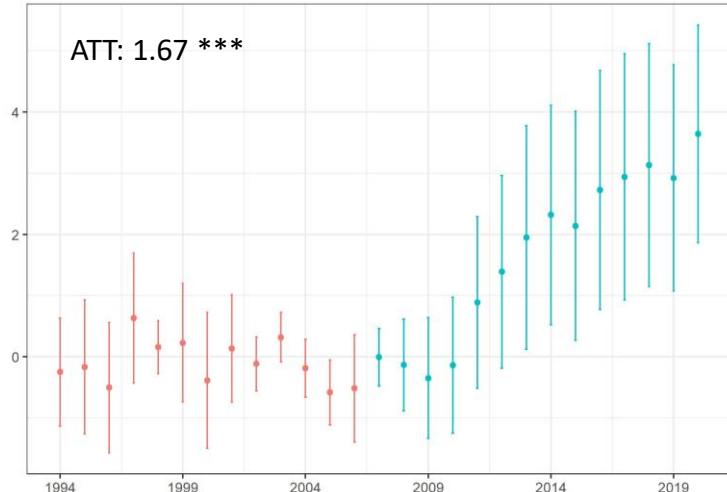
Note:

*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

- Increase in nat cover in pre-existing Cons Units, from 2011 on

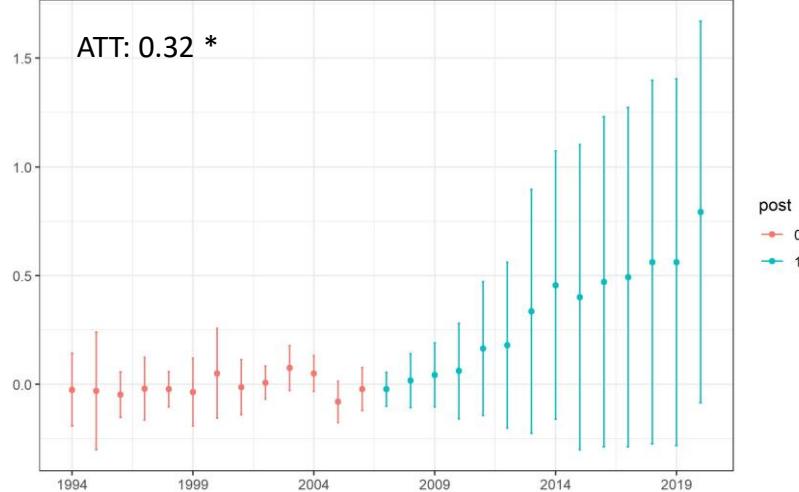
Nat cover in Cons. Units (as % Cons. Unit area in AMC)

Effect on natural cover in Conservation Units, selected states 0-100, 35cs0 2007



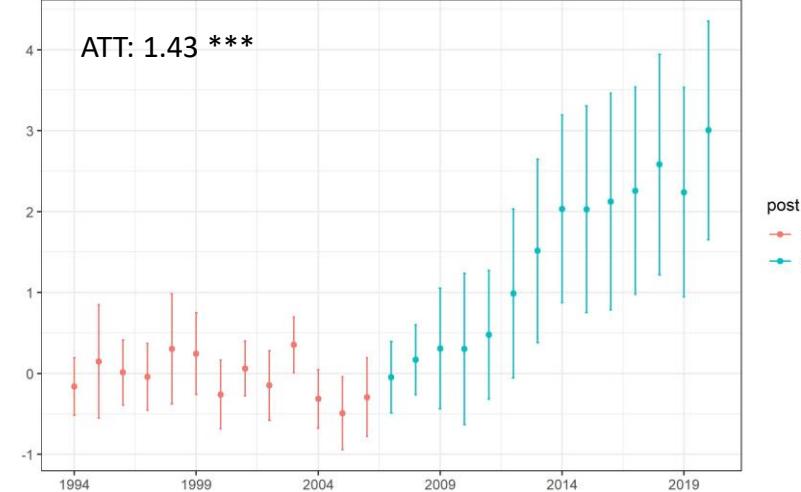
Nat cover in Cons. Units (as % AMC)

Effect on natural cover in Conservation Units (% AMC), selected states 0-100, 35cs0 2007

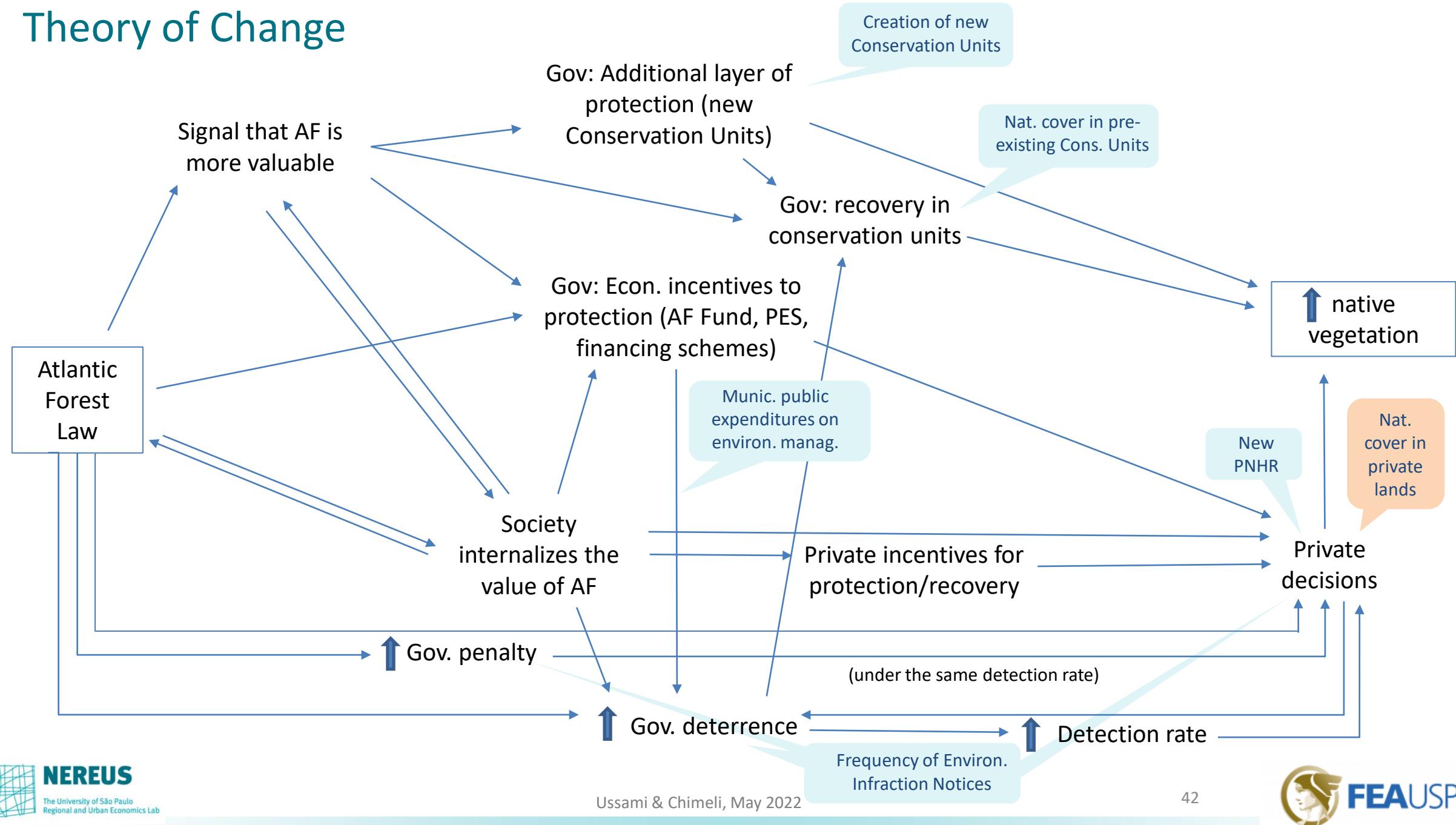


Nat cover in AMC (as % AMC)

Effect on natural cover, selected states 0-100, 28cs0 2007

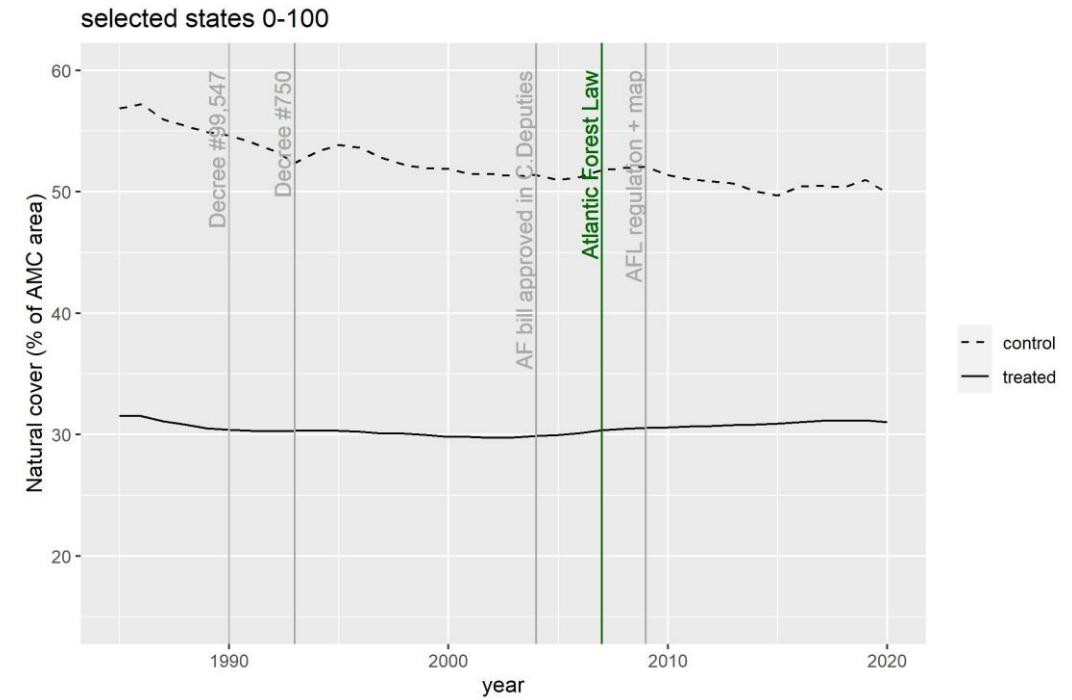
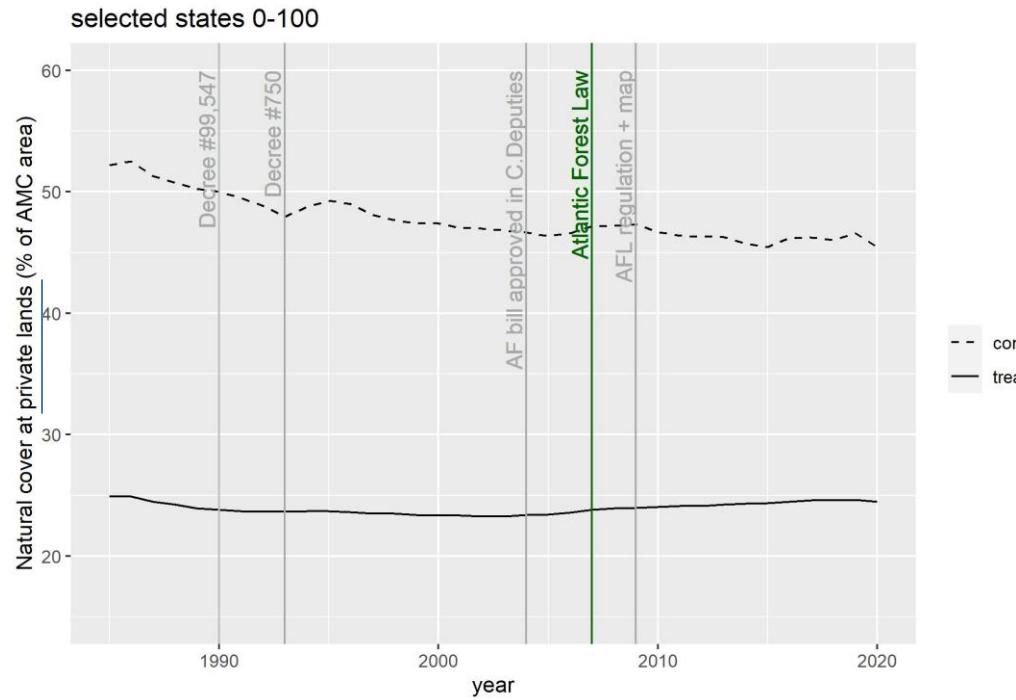


Theory of Change



Mechanisms: effects at private Lands

- “Private lands”: Excludes areas of Conservation Units/ Indigenous Lands/ Quilombo Lands [1]

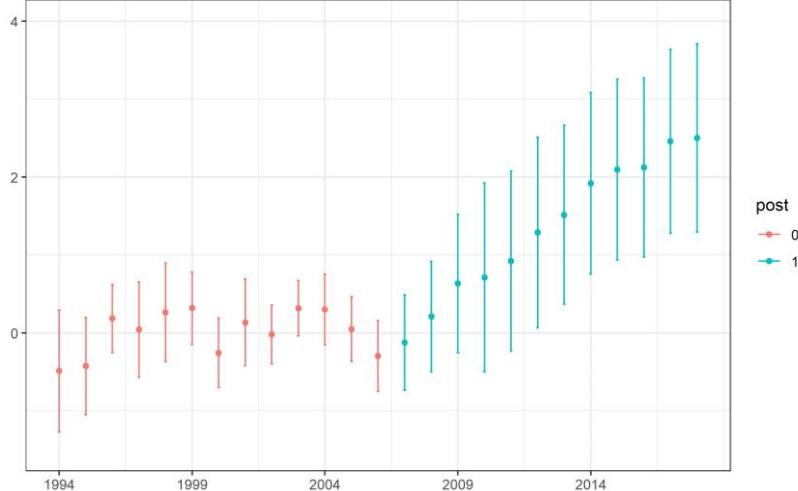


[1] settlements first established by escaped slaves

Mechanisms: effects at private Lands

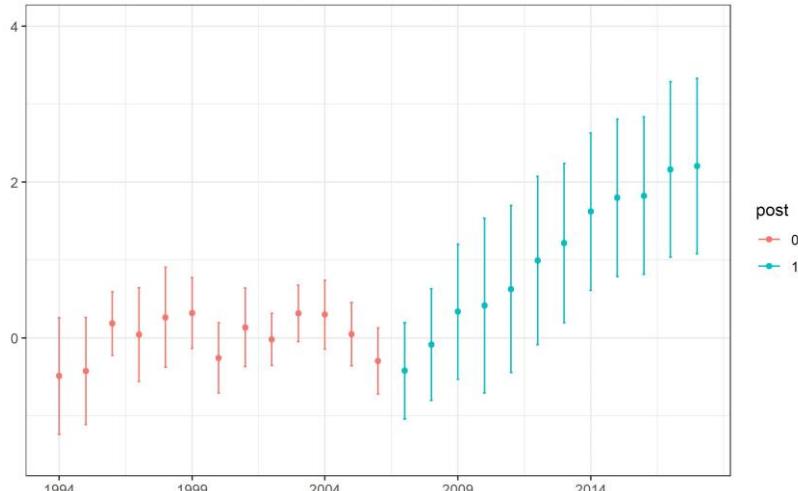
No anticipation

Effect on natural cover at non-protected lands, selected states 0-100, 25cs0 2007



1 year anticipation

Effect on natural cover at non-protected lands, selected states 0-100, 25cs1 2007



Dependent variable	ATT	Selected states			
		Private lands ^[1]		Reference	
		(1)	(2)	(3)	(4)
(a) Natural cover	ATT	1.317*** (0.203)	0.952*** (0.195)	1.601*** (0.311)	1.225*** (0.368)
(b) Net revegetation	ATT	0.590*** (0.109)	0.115 (0.180)	0.664*** (0.128)	0.262* (0.153)
AMC cluster		✓	✓	✓	✓
State dummies		✓	✓	✓	✓
Baseline nat. cover		✓	✓	✓	✓
Baseline control variables		✓	✓	✓	✓
Anticipation periods		0	1	0	1
Baseline nat. cover in treated AMCs (2006)		23.58	23.58	30.12	30.12
Qty. of treated AMCs		1661	1661	1661	1661
Qty. of control AMCs		1461	1461	1461	1461

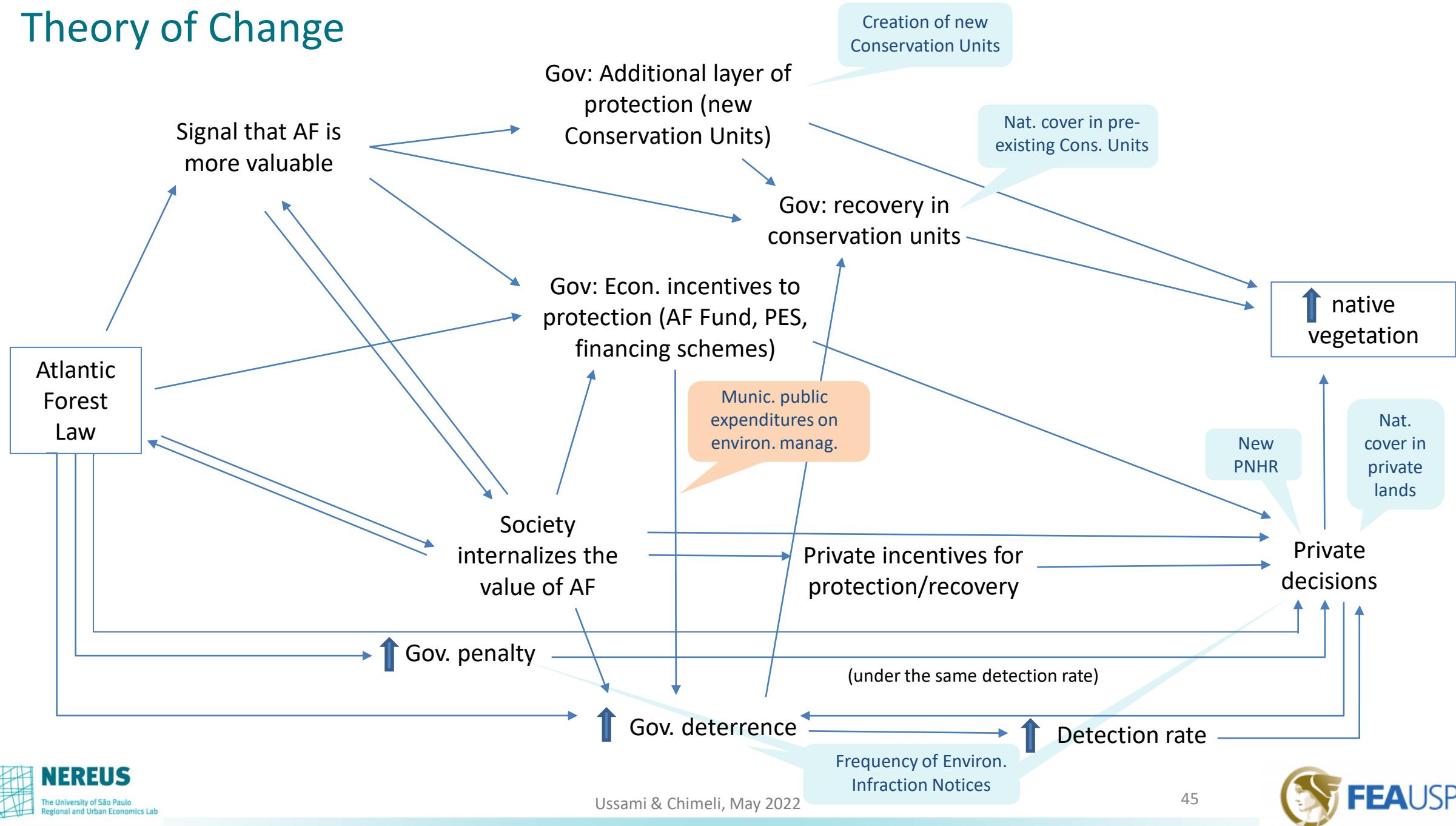
Note:

* p < 0.1; ** p < 0.05; *** p < 0.01

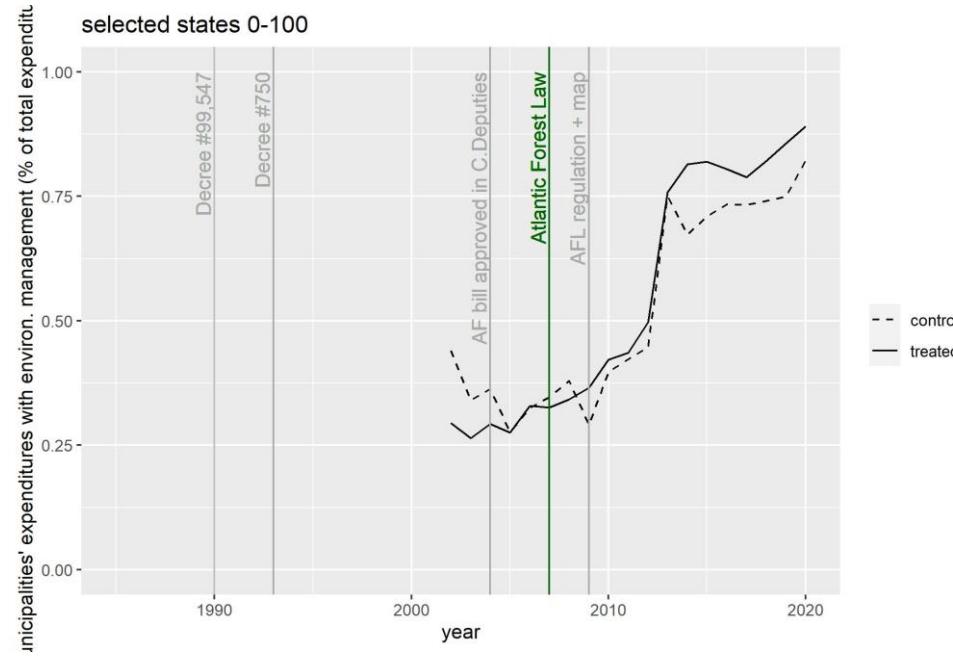
Robust standard errors are in parenthesis

^[1] excludes Conservation Units, Indigenous Lands, Quilombola Lands, and Rural Settlements

Theory of Change

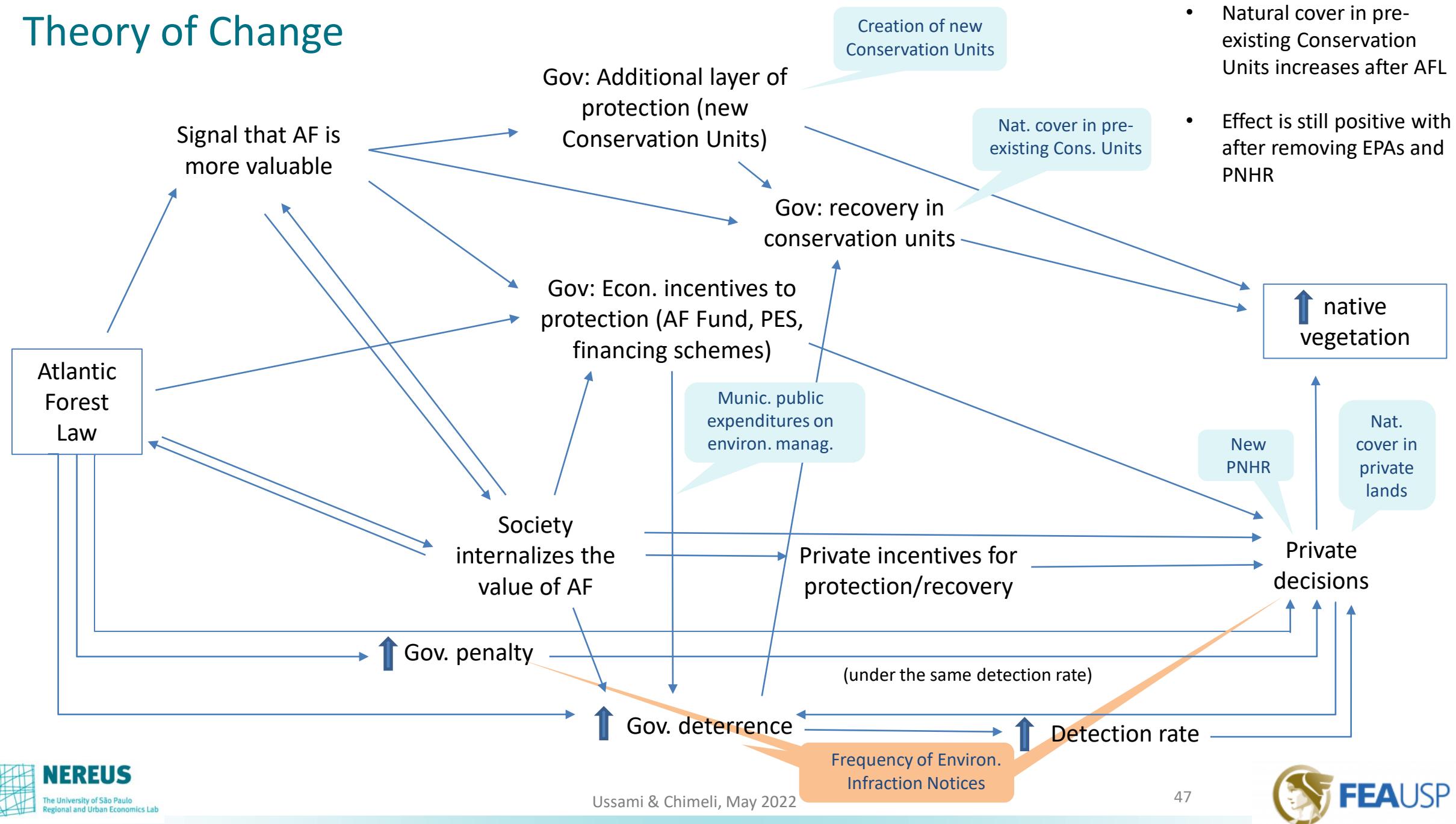


Munic. public expenditures on environ. management



- Expenditure data by function: available from 2002 on
- 0.5% is similar to state's expenditures
- No effect

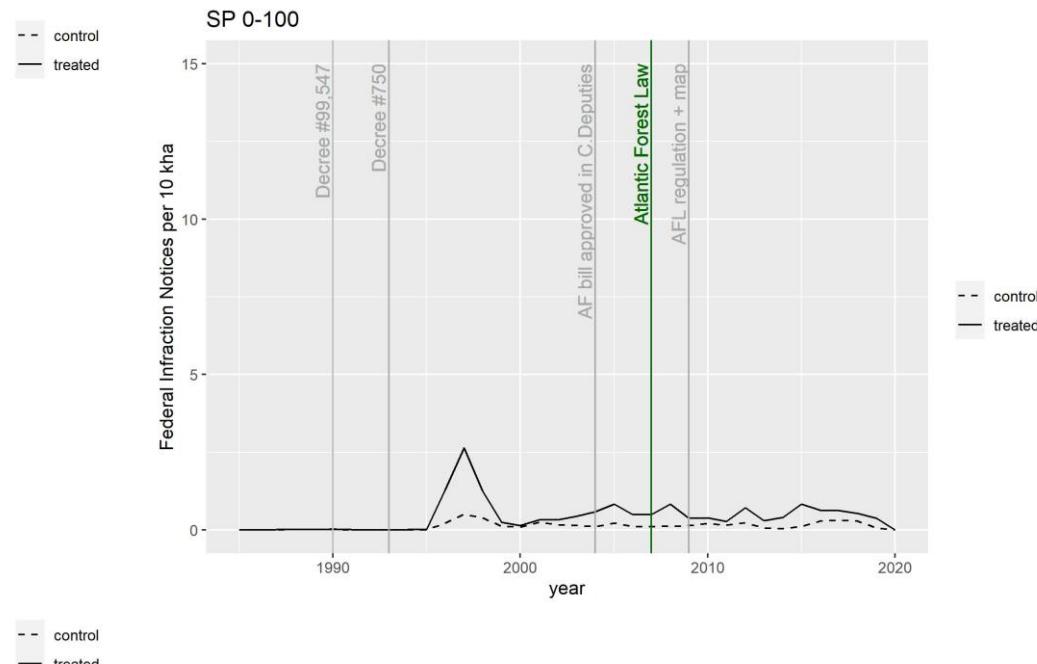
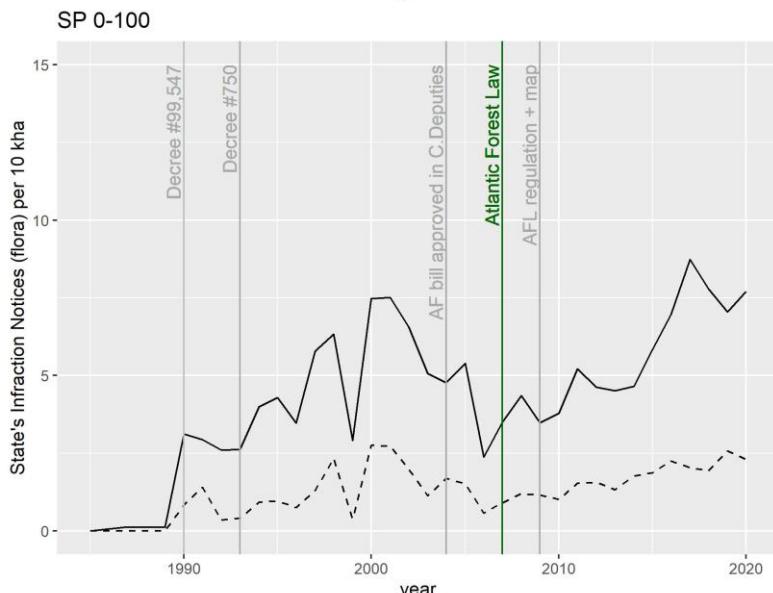
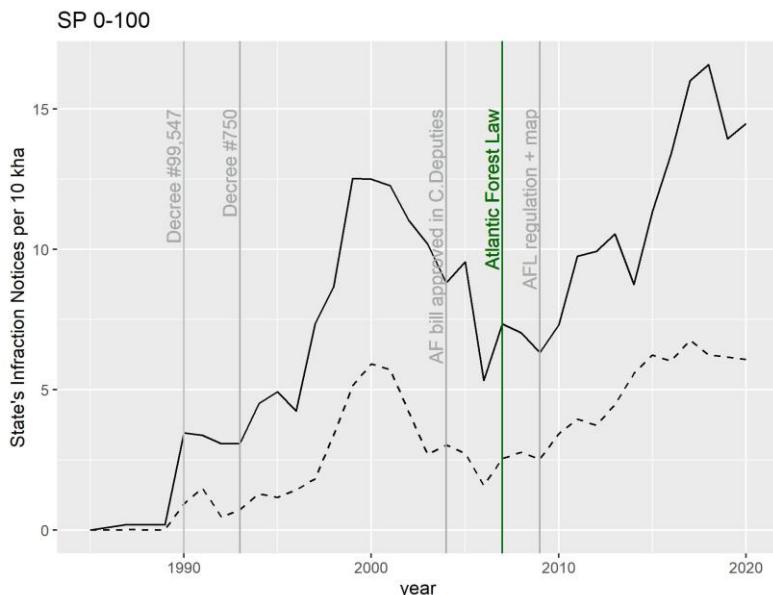
Theory of Change



SP, Environ. Infraction Notices per 10kha

Licensing and inspection of vegetation suppression and management is generally of state entities' duty

Shared mandate among federal entities: partially regulated by the Complementary Law nº 140/2011



Ussami & Chimeli, May 2022

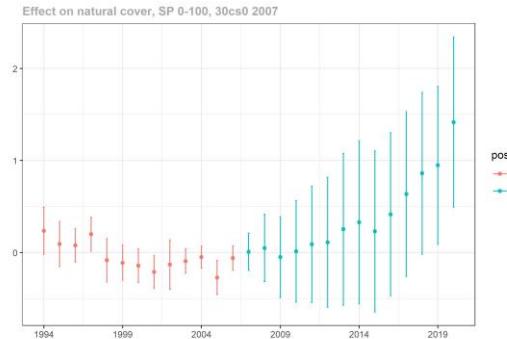
SP, Environ. Infraction Notices per 10kha

	State of Sao Paulo Dependent variable				
	State's EIN		Flora related State's EIN	State's EIN	Federal EIN
	(1)	(2)	(3)	(4)	(5)
ATT	2.256*** (0.838)	-0.409 (0.830)	2.113*** (0.383)	0.502 (0.452)	-0.076 (0.096)
AMC cluster	✓	✓	✓	✓	✓
State dummies					
Baseline nat. cover	✓	✓	✓	✓	✓
Baseline control variables	✓	✓	✓	✓	✓
Anticipation periods	0	1	0	1	0
Baseline EIN frequency (per 10 kha) in treated AMCs (2006)	5.32	5.32	2.39	2.39	0.50
Qty. of treated AMCs	399	399	399	399	399
Qty. of control AMCs	111	111	111	111	111

Note:

* p < 0.1; ** p < 0.05; *** p < 0.01

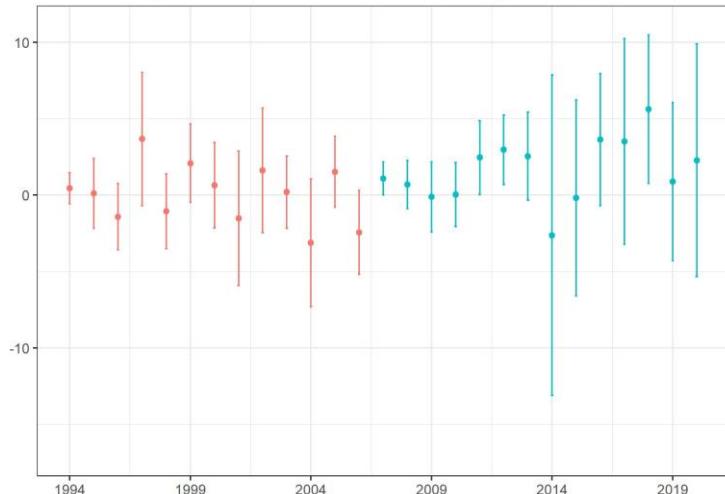
Robust standard errors are in parenthesis



- effects on the state's EIN related to flora increased after 2015 (nat cover from SP also increased after 2015)

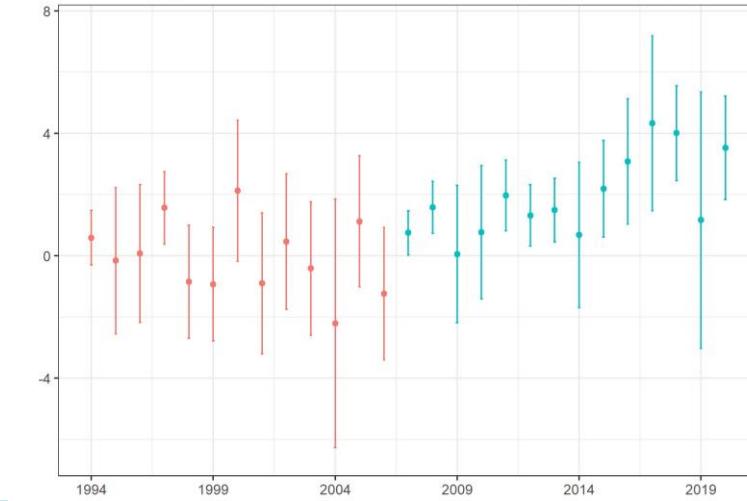
State's EIN

Estimated impact of AFL on state infraction notices per 10kha, SP 0-100, 20cs0 2007



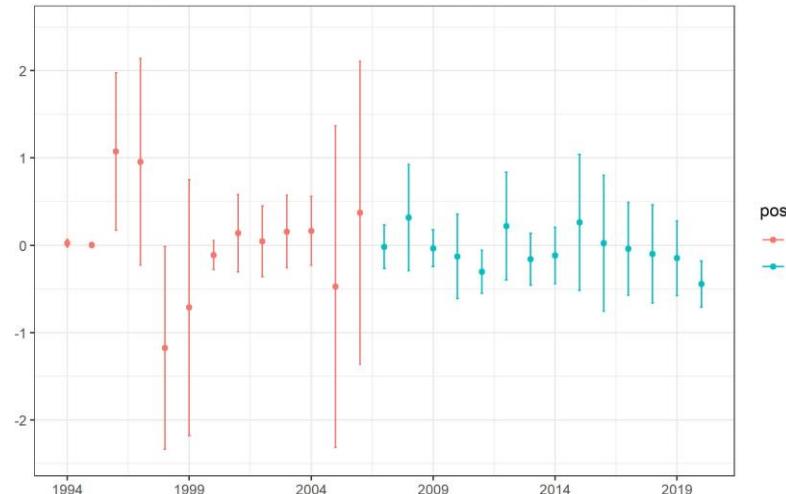
State's EIN (flora)

Estimated impact of AFL on state infraction notices (flora) per 10kha, SP 0-100, 20cs0 201



Federal EIN

Estimated impact of AFL on fed.env.infr.notices per 10kha, SP 0-100, 20cs0 2007



BR, Environ. Infraction Notices per 10kha

	<i>Dependent variable: Fed. environ. infraction notices per 10kha</i>				
	0-25 (1)	25-50 (2)	50-75 (3)	75-100 (4)	0-100 (5)
ATT	-0.211*** (0.072)	-0.196 (0.257)	0.073 (0.966)	-0.602*** (0.221)	0.178 (0.199)
AMC cluster	✓	✓	✓	✓	✓
State dummy	✓	✓	✓	✓	✓
Baseline nat.forest cover	✓	✓	✓	✓	✓
Baseline control variables ^[1]	✓	✓	✓	✓	✓
Anticipation periods	0	0	0	0	0
Baseline Fed. EIN per 10kha in treated AMCs (2006)	0.78	1.46	0.94	1.10	0.98
Qty. of treated AMCs	744	481	191	75	1661
Qty. of control AMCs	231	277	249	320	1461

Notes:

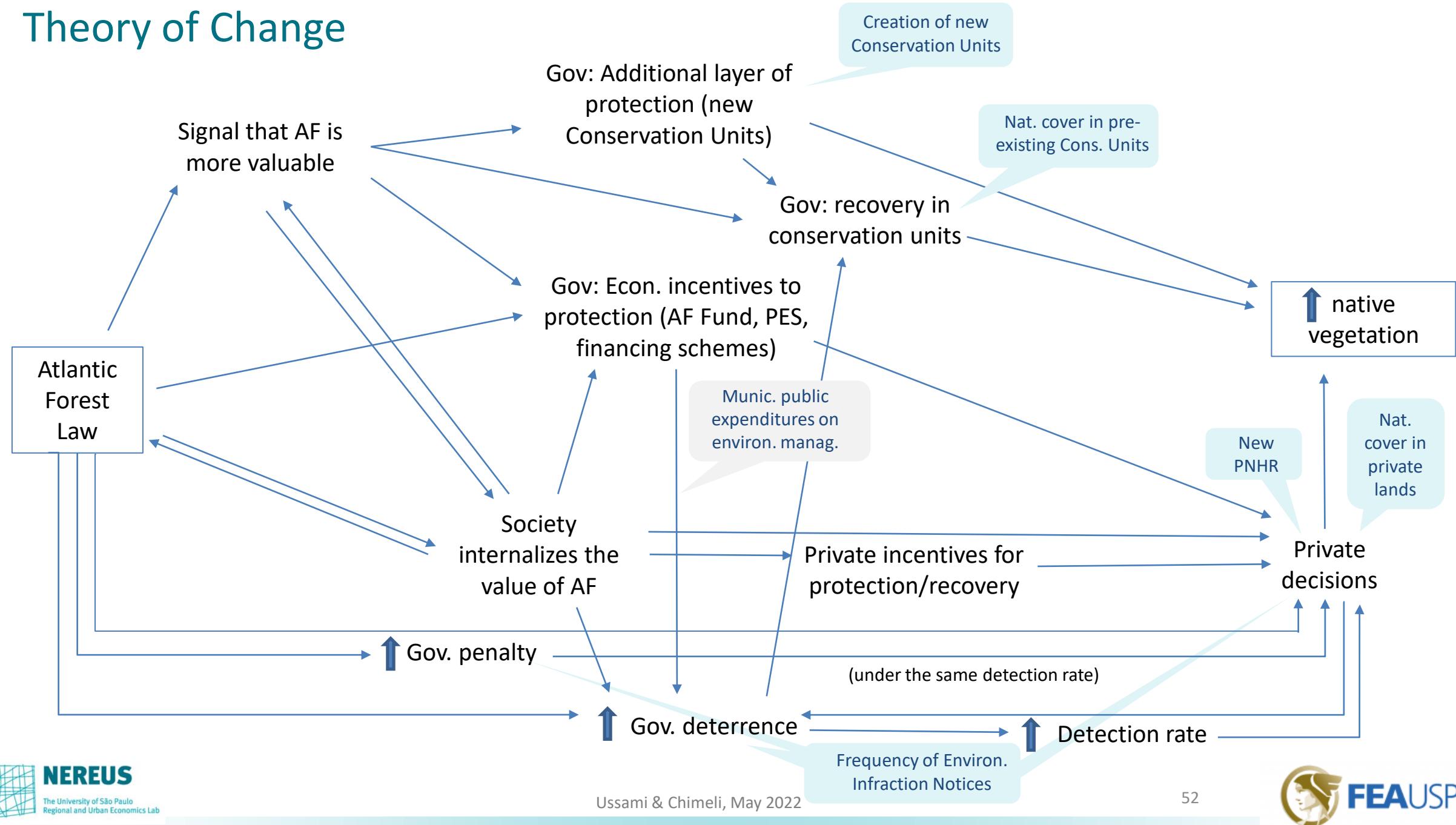
*p<0.1; **p<0.05; ***p<0.01
Robust standard errors are in parenthesis

- Reduction of the frequency in 0-25 and 75-100 groups

Conclusions

- Conservation Units coverage:
 - positively affected by AFL, in every group of Conservation Unit categories that we analyzed
 - positive effects also in private Conservation Units (PNHR)
- Pre-existing Conservation Units
 - increase in nat cover in pre-existing Cons Units, from 2011 on
- Private lands
 - Positive effects
- Environ. Infraction Notices:
 - SP: increase in EIN (flora) after 2015
 - Federal EIN: reduction in some samples
 - Future studies are required to address the reverse causality endogeneity
- Effectiveness of the AFL based on a set of different initiatives from different stakeholders

Theory of Change



Muito obrigada!

EAE 325 - Controle Sintético

Profa. Paula Pereda (FEA-USP)

June 21, 2022

Breve Apresentação

- ▶ Professora: [Link do site pessoal](#)
 - ▶ Exemplos de avaliações de políticas públicas
 - ▶ Extensão da licença paternidade em 15 dias
 - ▶ Políticas de promoção de alimentação saudável e sustentável (tributação, rotulagem)
 - ▶ Políticas ambientais no setor de transporte marítimo
 - ▶ Políticas energéticas (subsídio gasolina, carbon tax, painéis solares)

Método de Controle Sintético

- ▶ Referências: Abadie e Gardeazabal (2003), Abadie, Diamond e Hainmueller (2010,2011,2015). Revisão recente em Abadie (2020) [aqui](#)
- ▶ Vantagens:
 1. Análise de Políticas Agregadas.
 2. Estudo de uma intervenção (estudo de caso).
 3. Dados agregados: (i) mais disponíveis; (ii) menor incerteza (medidas agregadas tem menos erro).
 4. Reduz-se a discricionariedade em escolher grupos de controles (procedimento *data-driven*).
 5. Permite que heterogeneidades individuais variem no tempo (modelo de fatores).

Método de Controle Sintético: Motivação

- ▶ Abadie e Gardeazabal (2003) investigaram o efeito da instabilidade política na prosperidade econômica.
- ▶ Causalidade reversa: flutuações econômicas levarem a conflitos políticos
- ▶ Comparação cross-country: complicada por conta da heterogeneidade dos conflitos (Alesina et al)
- ▶ Abadie e Gardeazabal (2003) investigam o País Basco
 - ▶ Comunidade autônoma do norte da Espanha rica no passado.
 - ▶ O ETA, grupo sem motivação econômica, inicia suas atividades em 1970 (O pico de atividades ocorre nos anos 80)
 - ▶ Empresas são vítimas de violência e extorsão.

Método de Controle Sintético: Motivação

- ▶ Dificuldades na estimação do efeito
 - ▶ Δt : Houve uma recessão na Espanha no final dos anos 70, começo dos 80 (confounder)
 - ▶ Δi : O País Basco e a Espanha tem características distintas que podem influenciar o resultado econômico.
- ▶ Portanto, Comparações captura efeito do terrorismo + efeito das diferenças determinantes de crescimento pré-terrorismo.
- ▶ AG 2003: Usam uma combinação de regiões da Espanha para construir um contrafactual (controle sintético) que se assemelhe (em características) ao País Basco nos anos 1960.
- ▶ Ideia: Evolução econômica do controle sintético seria semelhante à evolução econômica do País Basco sem terrorismo.
- ▶ É uma comparação de caso (case study)

Método de Controle Sintético: Modelo

- ▶ Conta-se com $J + 1$ unidades observadas em T períodos.
 - ▶ A primeira unidade sofre intervenção
 - ▶ **Donor pool**, potenciais controles, é o conjunto das unidades 2 a $J + 1$.
- ▶ A intervenção ocorre a partir de T_0 tal que $1 \leq T_0 \leq T$ e seus efeitos podem se propagar de $T_0 + 1$ a T .
- ▶ Y_{it}^N : resultados potenciais para a região i no período t **sem intervenção**
- ▶ Y_{it}^I os resultados potenciais da unidade i nos períodos t posteriores à exposição de i à intervenção.
- ▶ A intervenção não afeta Y antes da implementação e os Y não-tratados não são impactados pela intervenção (não há efeitos antecipação e contaminação)

Método de Controle Sintético: Modelo

- ▶ Sejam:
 - ▶ $\alpha_{it} = Y_{it}^I - Y_{it}^N$ o efeito da intervenção na unidade i em t
 - ▶ $D_{it} = 1$ se $i = 1$ e $t > T_0$ e 0 caso contrário.
- ▶ Os resultados observados são

$$Y_{it} = Y_{it}^N + \alpha_{it} D_{it} \quad (1)$$

- ▶ e queremos estimar $\alpha_{1t} = Y_{1t}^I - Y_{1t}^N$.
- ▶ Y_{1t}^N é missing após a intervenção (Problema Fundamental da Inferência Causal)

Método de Controle Sintético: Modelo

- ▶ Suponha que a equação de Y_{it}^N seja dada pelo modelo de fatores

$$Y_{it}^N = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{it} \quad (2)$$

- ▶ em que
 - ▶ δ_t é um fator comum no tempo não observado (comum entre as unidades);
 - ▶ Z_i é o vetor de covariadas observadas (não-afetadas pelo tratamento);
 - ▶ λ_t é um vetor de fatores comuns não-observados;
 - ▶ μ_i é um vetor de fatores específicos; e
 - ▶ ε_{it} representa choques transitórios.
- ▶ Obs.: $\lambda_t \mu_i$ (generalização do DID ∵ vantagem do CS) permite que efeitos fixos variem no tempo (mudanças temporárias). Permite respostas heterôgeneas para fatores não-observados múltiplos.

Método de Controle Sintético

- ▶ Considere um vetor de pesos $w = (w_2, \dots, w_{J+1})$ em que $w_j \geq 0, \forall j$ e que $\sum_{j=2}^{J+1} w_j = 1$. Cada w representa um controle sintético (média ponderada das regiões) e o resultado potencial de cada controle sintético é:

$$\sum_{j=2}^{J+1} w_j Y_{jt}^N = \delta_t + \theta_t \sum_{j=2}^{J+1} w_j Z_j + \lambda_t \sum_{j=2}^{J+1} w_j \mu_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt}$$

- ▶ Assim, tem-se que, subtraindo as equações (para $i = 1$),

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N = \theta_t \left[Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right] + \lambda_t \left[\mu_1 - \sum_{j=2}^{J+1} w_j \mu_j \right] +$$

$$+ \sum_{j=2}^{J+1} w_j (\varepsilon_{1t} - \varepsilon_{jt})$$

Método de Controle Sintético: Modelo

- ▶ Assuma que ε_{it} sejam independentes entre i e t .
- ▶ Mesmo assim, fatores não-observados $u_{it} = \lambda_t \mu_i + \varepsilon_{it}$ podem ser correlacionados entre i e t por conta de $\lambda_t \mu_i$.
- ▶ Por conta disso, não se pode estimar por MQO e é preciso eliminar o problema dos fatores não-observados.
- ▶ Desta forma, é preciso supor que ε_{it} sejam independentes de $\{Z_i, u_i\}_{i=1}^{J+1}$.
- ▶ Denotemos por P os seguintes vetores pré-tratamento:
 - ▶ Y_i^P — vetor $T_0 \times 1$ com o t -ésimo elemento Y_{it} ;
 - ▶ ε_i^P — vetor $T_0 \times 1$ com o t -ésimo elemento ε_{it} ;
 - ▶ θ^P — vetor $T_0 \times r$ com t -ésima linha θ_t ; e
 - ▶ λ^P — vetor $T_0 \times f$ com t -ésima linha λ_t .

Método de Controle Sintético: Modelo

- Obtemos da equação aplicada ao período pré-tratamento,

$$Y_1^P - \sum_{j=2}^{J+1} w_j Y_j^P = \theta^P \left[Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right] + \lambda^P \left[\mu_1 - \sum_{j=2}^{J+1} w_j \mu_j \right] +$$

$$\left(\varepsilon_1^P - \sum_{j=2}^{J+1} w_j \varepsilon_j^P \right)$$

- Assuma que $\lambda^P T \lambda^P$ seja não-singular e, portanto, inversível
- Subtraindo a equação multiplicada por $\lambda_t \left(\lambda^P T \lambda^P \right)^{-1} \lambda^P T$ da equação anterior, obtemos:

Método de Controle Sintético

$$\begin{aligned} & \left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right) - \lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \left(Y_1^P - \sum_{j=2}^{J+1} w_j Y_j^P \right) = \\ & = \left(\theta_t - \left(\lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \right) \theta^P \right) \left[Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right] + \\ & + \left(\lambda_t - \lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \lambda^P \right) \left[\mu_1 - \sum_{j=2}^{J+1} w_j \mu_j \right] + \\ & + \sum_{j=2}^{J+1} w_j (\varepsilon_{1t} - \varepsilon_{jt}) - \lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \left(\varepsilon_1^P - \sum_{j=2}^{J+1} w_j \varepsilon_j^P \right) \end{aligned}$$

Método de Controle Sintético

e, portanto,

$$\begin{aligned} \left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right) &= \lambda_t \left(\lambda^{PT} \lambda^P \right)^{-1} \lambda^{PT} \left(Y_1^P - \sum_{j=2}^{J+1} w_j Y_j^P \right) + \\ &+ \left(\theta_t - \lambda_t \left(\lambda^{PT} \lambda^P \right)^{-1} \lambda^{PT} \theta^P \right) \left[Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right] + \\ &+ \sum_{j=2}^{J+1} w_j (\varepsilon_{1t} - \varepsilon_{jt}) - \lambda_t \left(\lambda^{PT} \lambda^P \right)^{-1} \lambda^{PT} \left(\sum_{j=2}^{J+1} w_j (\varepsilon_1^P - \varepsilon_j^P) \right) \end{aligned}$$

Método de Controle Sintético

Suponha agora que existe $W^* = (w_2^*, \dots, w_{J+1}^*)$ tal que

$$\begin{aligned}\sum_{j=2}^{J+1} w_j^* Y_{j1} &= Y_{11}, \quad \sum_{j=2}^{J+1} w_j^* Y_{j2} = Y_{12}, \\ &\vdots \\ \sum_{j=2}^{J+1} w_j^* Y_{jT_0} &= Y_{1T_0}, \\ \sum_{j=2}^{J+1} w_j^* Z_j &= Z_1.\end{aligned}\tag{3}$$

Método de Controle Sintético

- ▶ Assim, o ajuste do contrafactual pode ser reduzido a

$$Y_{1t}^N - \sum w_j^* Y_{jt}^N = R_{1t} + R_{2t} + R_{3t}$$

- ▶ em que

$$R_{1t} = \lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \sum w_j^* \varepsilon_j^P,$$

$$R_{2t} = -\lambda_t \left(\lambda^{P^T} \lambda^P \right)^{-1} \lambda^{P^T} \varepsilon_1^P,$$

$$R_{3t} = \left(\varepsilon_1^P - \sum_{j=2}^{J+1} w_j \varepsilon_j^P \right)$$

Método de Controle Sintético

- ▶ Os autores mostram que $E(R_{2t}) = E(R_{3t}) = 0, \forall t > T_0$.
- ▶ Já para R_{1t} , mostram que:

$$E(R_{1t}) \leq C(p)^{\frac{1}{p}} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{\frac{1}{p}} \operatorname{Max} \left\{ \frac{(\bar{m}_p)^{\frac{1}{p}}}{T_0^{1-\frac{1}{p}}}, \frac{\hat{\sigma}}{\sqrt{T_0}} \right\}$$

- ▶ em que
 - ▶ $C(p)$ é o p -ésimo momento central (menos 1) de uma variável aleatória Poisson com $\lambda = 1$,
 - ▶ $\left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right)$ é o limite inferior dos autovalores de $\frac{1}{M} \sum \lambda_t^T \lambda_t$
 - ▶ \bar{m}_p é a média do p -ésimo momento de ε_{jt} , e
 - ▶ $\hat{\sigma}$ é o $\operatorname{Max} \sigma_j^2$, em que $\sigma_j^2 = E(\varepsilon_{jt}^2)$.

Método de Controle Sintético: Modelo

- ▶ Importante: comportamento de $E(R_{1t})$ (viés do estimador limitado por ele).
- ▶ Os autores mostram que $E(R_{1t}) \rightarrow 0$ conforme T_0 aumenta (relativo à escala dos choques).
- ▶ Intuição: conforme T_0 cresce, o matching nos resultados pré-intervenção ajuda a controlar fatores não-observados e heterogeneidade dos efeitos observados sobre Y .
- ▶ I.e., unidades parecidas nos determinantes observados e não-observados de Y devem produzir trajetórias similares às de Y em instantes posteriores a T_0 . Portanto, a diferença entre as trajetórias em $t > T_0$ é o efeito da intervenção (gap, α_{1t}).

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, t \in \{T_0 + 1, \dots, T\}$$

Método de Controle Sintético: Modelo

- ▶ As condições para W^* estabelecidas valem na igualdade se $(Y_{11}, \dots, Y_{1T_0}, Z'_1)$ pertence ao cone convexo gerado por $\{(Y_{21}, \dots, Y_{2T_0}, Z'_2), \dots, (Y_{(J+1)1}, \dots, Y_{(J+1)T_0}, Z'_{J+1})\}$.
- ▶ O suporte deve ser parecido (facilita comparação/pareamento).
- ▶ W^* satisfaz as condições aproximadamente (não na igualdade).
- ▶ Viés de interpolação: entre regiões com características diferentes (minimiza-se restringindo donor pool a regiões com características similares).

Método de Controle Sintético: Implementação

- A implementação reside na escolha de W^* . Condições iniciais:

$$\begin{cases} w_j \geq 0, \forall j \in \{2, \dots, J+1\} \\ \sum_{j=2}^{J+1} w_j = 1 \end{cases}$$

- Seja também K um vetor $T_0 \times 1$ tal que $K = (K_1, \dots, K_{T_0})$ representa uma média ponderada dos resultados pré-intervenção para todas as unidades i , i.e., $\bar{Y}_i^K = \sum_{s=1}^{T_0} K_s Y_{is}$

Método de Controle Sintético: Implementação

- ▶ Sejam M combinações K_1, \dots, K_M , $X_1 = (Z'_1, \bar{Y}_1^{K_1}, \dots, \bar{Y}_1^{K_M})'$ um vetor $M + 1 \times 1$ de características pré-intervenção para a região exposta e X_0 a matriz $M + 1 \times J$ das regiões não-afetadas tais que a coluna j de X_0 é $(Z'_j, \bar{Y}_j^{K_1}, \dots, \bar{Y}_j^{K_M})'$ ¹².
- ▶ Assim, X_0 e X_1 são previsores dos resultados pós-intervenção.
- ▶ O vetor W^* é escolhido de tal sorte a minimizar a distância entre X_1 e X_0 ponderado:

$$W^* = \underset{\{W\}}{\operatorname{argmin}} \|X_1 - X_0 W\| \text{ s.t. } \begin{cases} w_j \geq 0, \forall j \in \{2, \dots, J+1\} \\ \sum_{j=2}^{J+1} w_j = 1 \end{cases}$$

Método de Controle Sintético: Implementação

- ▶ Funções de distância:

$$\|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

- ▶ em que V é uma matriz $(M+1) \times (M+1)$ simétrica positiva (ou diagonal). Importância relativa para a variável K .
- ▶ Escolha ótima de V :
 - ▶ Minimiza o Erro Quadrático Médio
 - ▶ Algoritmos $W(V)$ (ajuste dos dados)

Projeção de X_1 no conjunto convexo de X_0 (Abadie, 2020)

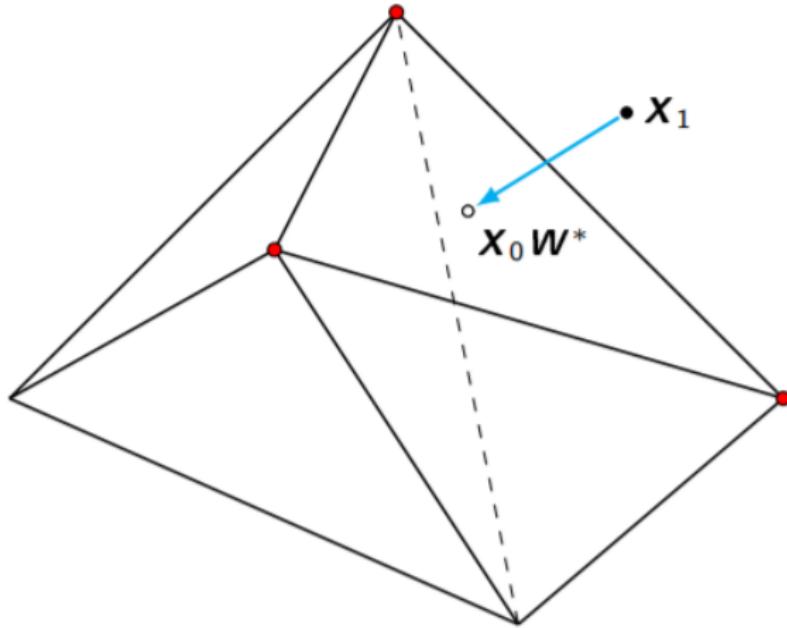


Figure 2: Projecting X_1 on the convex hull of X_0

Método de Controle Sintético vs DID

- ▶ CS generaliza DID com o modelo de fatores: No DID tradicional, $\lambda_t = \lambda, \forall t$.

$$Y_{it}^N = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{it}$$

- ▶ O DID tradicional também permite fatores não-observados (*confounders*), mas restringe a efeitos constantes no tempo e, portanto, são eliminados pela diferença temporal.
- ▶ A diferença temporal não elimina μ_i , mas sim o que o CS elimina.
- ▶ CS pode gerar estimativas mais úteis em alguns contextos do que DID.
- ▶ CS é um método bastante transparente (na escolha de controles e no ajuste do modelo)

Controle Sintético: Em resumo

- ▶ Hipóteses de Identificação:
 1. CS1: Modelo de fatores para o contrafactual (flexível):
$$y_{it}^0 = \lambda_t + x_i\theta_t + \delta_t\mu_i + \varepsilon_{it}$$
 2. CS2: Existem w_2^*, \dots, w_{J+1}^* , tal que $w_j^* \geq 0 \forall j = 2, \dots, J+1$ e
$$\sum_{j=2}^{J+1} w_j^* = 1:$$
$$\sum_{j=2}^{J+1} w_j^* y_{jt} = y_{1t} \text{ e } \sum_{j=2}^{J+1} w_j^* x_j = x_1 \forall t \leq T_0.$$
- ▶ Sob CS1 e CS2, o estimador $\hat{\alpha}_{1t} = y_{1t} - \sum_{j=2}^{J+1} w_j^* y_{jt}$ é consistente/não viesado para α_{1t} ($\text{Viés} \rightarrow 0$, quando $T_0 \rightarrow \infty$, ou quando T_0 é grande relativo à escala dos choques transitórios ε_{it}).
- ▶ Para validade de CS2: as variáveis do tratamento devem pertencer ao conjunto convexo das variáveis do grupo de controle (CS2 vale aproximadamente, inspeção gráfica).

Método de Controle Sintético: Inferência

- ▶ **Usual: testes com placebos (falsification tests)**
- ▶ Distribuição da estatística do teste: Computa-se usando permutações aleatórias das unidades amostrais sem intervenção (ou dos períodos sem intervenção).
 - ▶ Aplica-se o método às unidades do donos pool e plota a distribuição de resultados ($H_0: \alpha = 0$).
- ▶ Bertrand e Duflo 2004, Imbens e Wooldridge 2009 adicionam que o matching antes do controle sintético pode garantir a comparabilidade dos controles admissíveis
- ▶ Novos testes usando permutação: Chernozhukov et al (2020)
- ▶ Synthetic DID (SDID): Arkhangelsky et al (2019) => Duplo-robusto (DID com pesos com base em CS para controles)

Análise de Robustez: CS

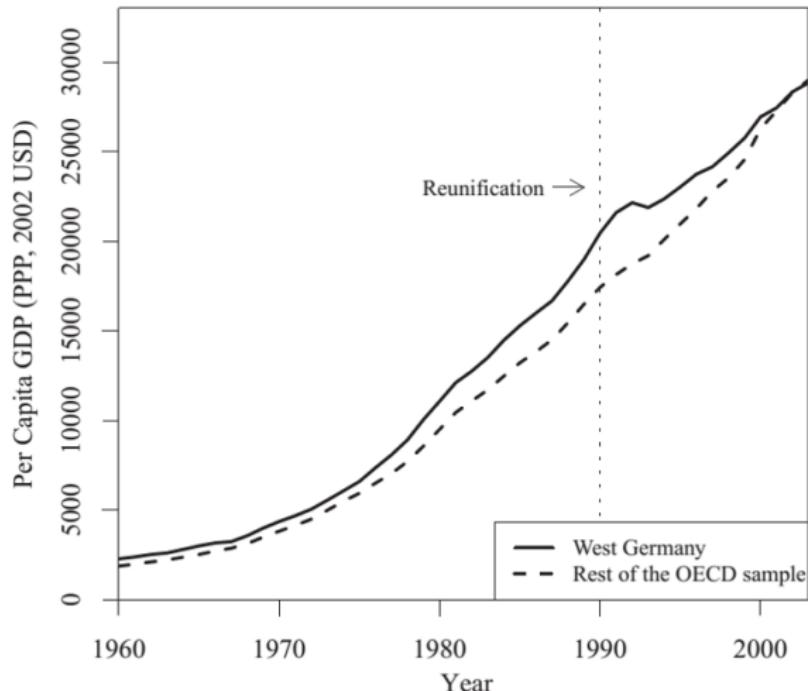
- ▶ **Teste de Falsificação:** Antecipar a data do tratamento ($t < T_0$). Verificar efeito antecipação.
- ▶ **Teste de Permutação** (robustez e inferência): Aleatorizar observações tratadas para testar efeito do tratamento.
- ▶ **Comparação "leave one out":** Excluir unidades de controle que receberam pesos positivos do *donor pool* e refazer análise. Verificar se uma unidade em particular é quem estima resultado (em geral, há poucas unidades com peso positivo).
- ▶ **Testar outros *donor pools*.**
- ▶ Variar covariadas (X) incluídas.

Controle Sintético: Aplicações

- ▶ Alemanha - Reunificação (Abadie et al 2015)
 - ▶ Foi avaliado o impacto econômico na Alemanha Ocidental da reunificação em out/1990 após 45 anos de separação.
 - ▶ O *donor pool* foi composto de países da OCDE.
 - ▶ O período analisado é de 1960 a 2003 (30 anos de dados pré-intervenção).
 - ▶ Os dados agregados disponíveis conformavam um painel cross-country anual de 16 países³ com GDP real per capita (PPP 2002 USD), Y , e o vetor de covariadas: % indústria, taxa de investimento, escolaridade, abertura comercial.
 - ▶ Placebo: reunificação em 1975.
 - ▶ O cálculo do efeito a partir de 1992 foi uma redução de 8% do GDP.
 - ▶ Como teste de robustez, analisou-se a sensibilidade do resultado a mudanças de W^* .
 - ▶ Crítica: efeito spillover da diminuição do GDP da Alemanha.

Controle Sintético: Aplicações

FIGURE 1 Trends in per Capita GDP: West Germany versus Rest of the OECD Sample



Controle Sintético: Aplicações

TABLE 1 Synthetic and Regression Weights for West Germany

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.09	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	United Kingdom	0	0.06
Japan	0.16	0.19	United States	0.22	0.13

Notes: The synthetic weight is the country weight assigned by the synthetic control method. The regression weight is the weight assigned by linear regression. See text for details.

Controle Sintético: Aplicações

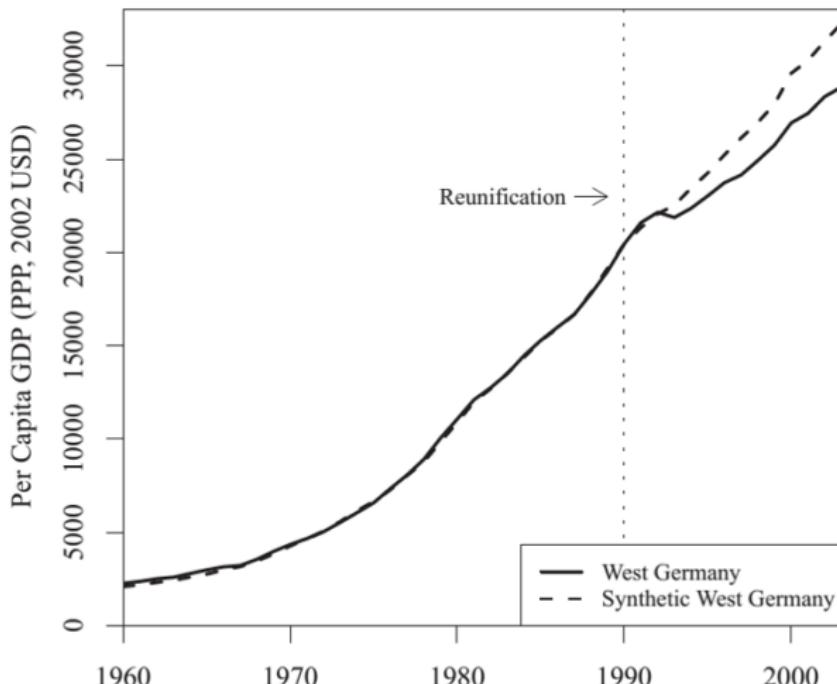
**TABLE 2 Economic Growth Predictor Means
before German Reunification**

	West Germany	Synthetic West Germany	OECD Sample
GDP per capita	15808.9	15802.2	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

Notes: GDP per capita, inflation rate, trade openness, and industry share are averaged for the 1981–90 period. Investment rate and schooling are averaged for the 1980–85 period. The last column reports a population-weighted average for the 16 OECD countries in the donor pool.

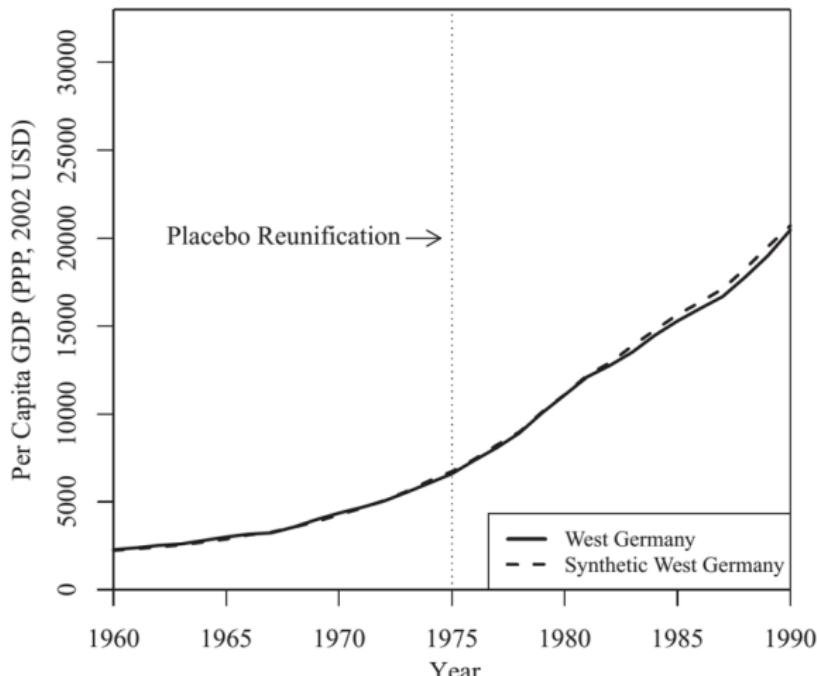
Controle Sintético: Aplicações

FIGURE 2 Trends in per Capita GDP: West Germany versus Synthetic West Germany



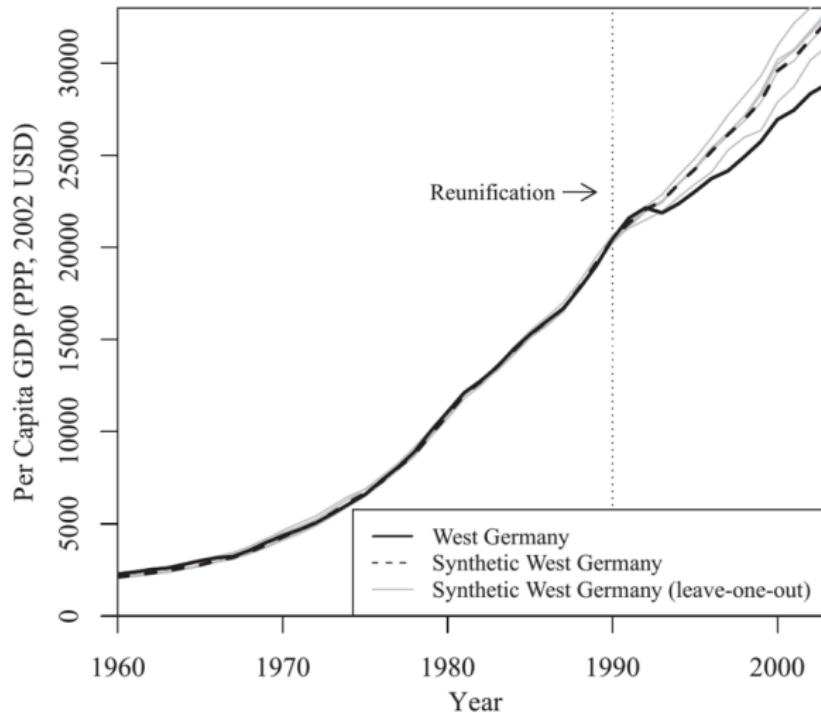
Controle Sintético: Aplicações

FIGURE 4 Placebo Reunification 1975—Trends in per Capita GDP: West Germany versus Synthetic West Germany



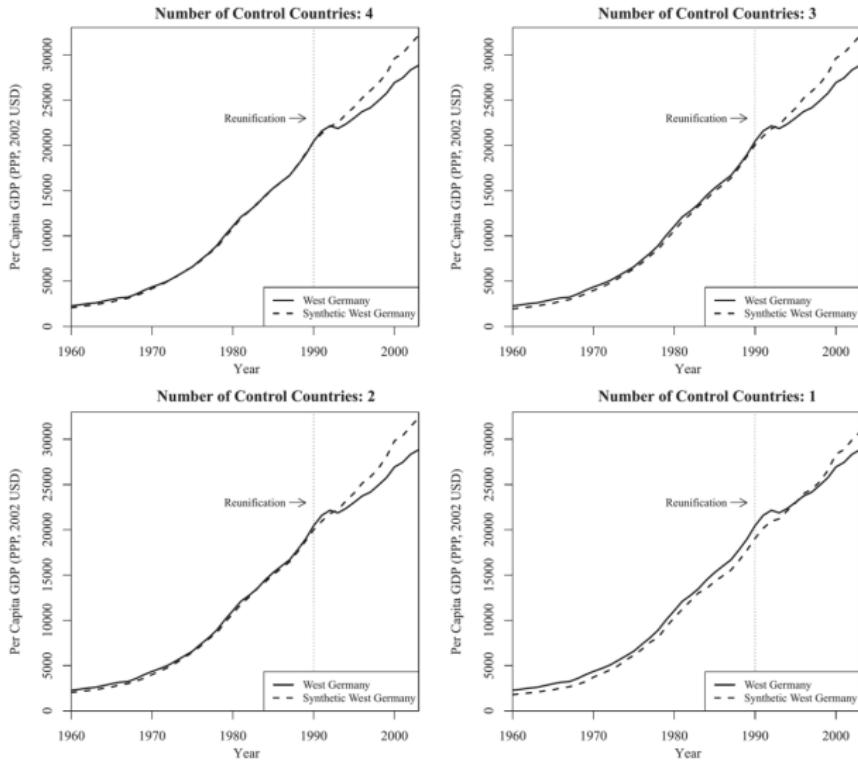
Controle Sintético: Aplicações

FIGURE 6 Leave-One-Out Distribution of the Synthetic Control for West Germany



Controle Sintético: Aplicações

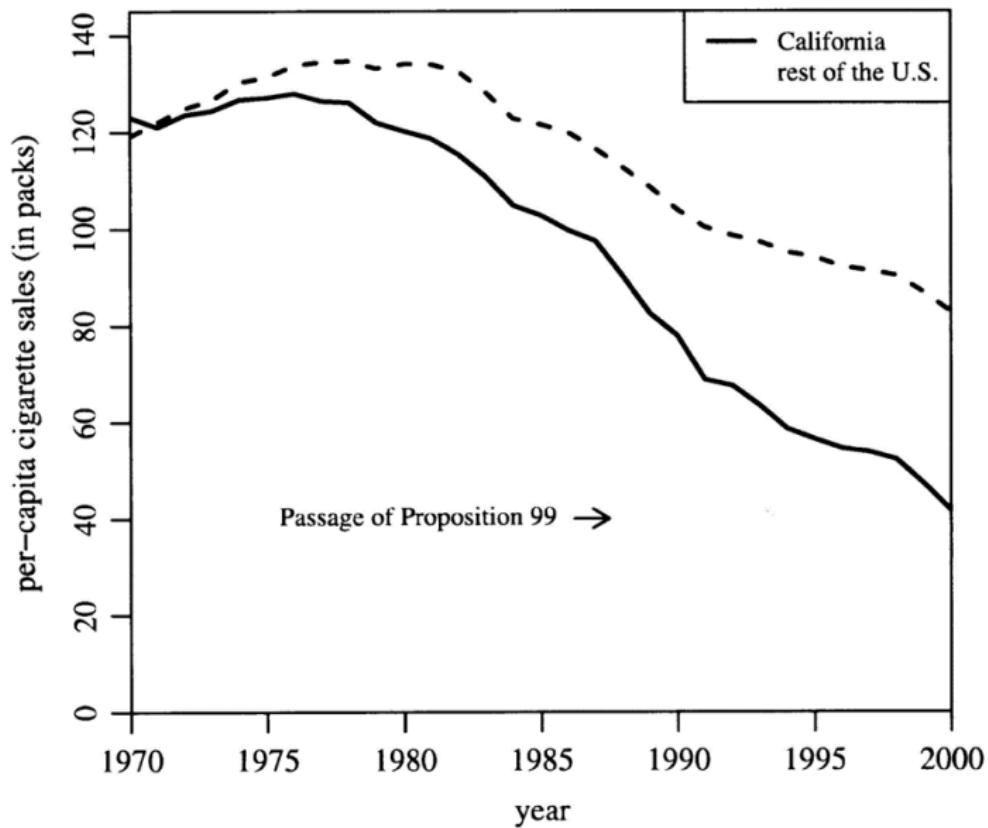
FIGURE 7 Per Capita GDP Gaps between West Germany and Sparse Synthetic Controls



Controle Sintético: Aplicações

- ▶ Califórnia - Legislação Antitabagista (Abadie et al 2010)
 - ▶ Em janeiro de 1989, aumentou-se o imposto incidente no cigarro e direcionou-se o dinheiro para saúde e educação antifumo.
 - ▶ O *donor pool* é composto de outros estados americanos.
 - ▶ Excluem-se do *donor pool* estados com legislação antitabagista.
 - ▶ O período analisado é 1970 a 2000.
 - ▶ Os dados agregados disponíveis são Y consumo per capita de cigarro e seu preço, logaritmo da renda, porcentagem da população de 15 a 24 anos e consumo de cerveja.
 - ▶ Placebo: aplicar CS sobre unidades controle iterativamente.
 - ▶ O resultado encontrado foi um declínio de 25% (gap estimado).
 - ▶ Como teste de robustez, foram adicionadas outras covariadas como desemprego, pobreza, demografia etc. que não afetaram o resultado.

Controle Sintético: Aplicações

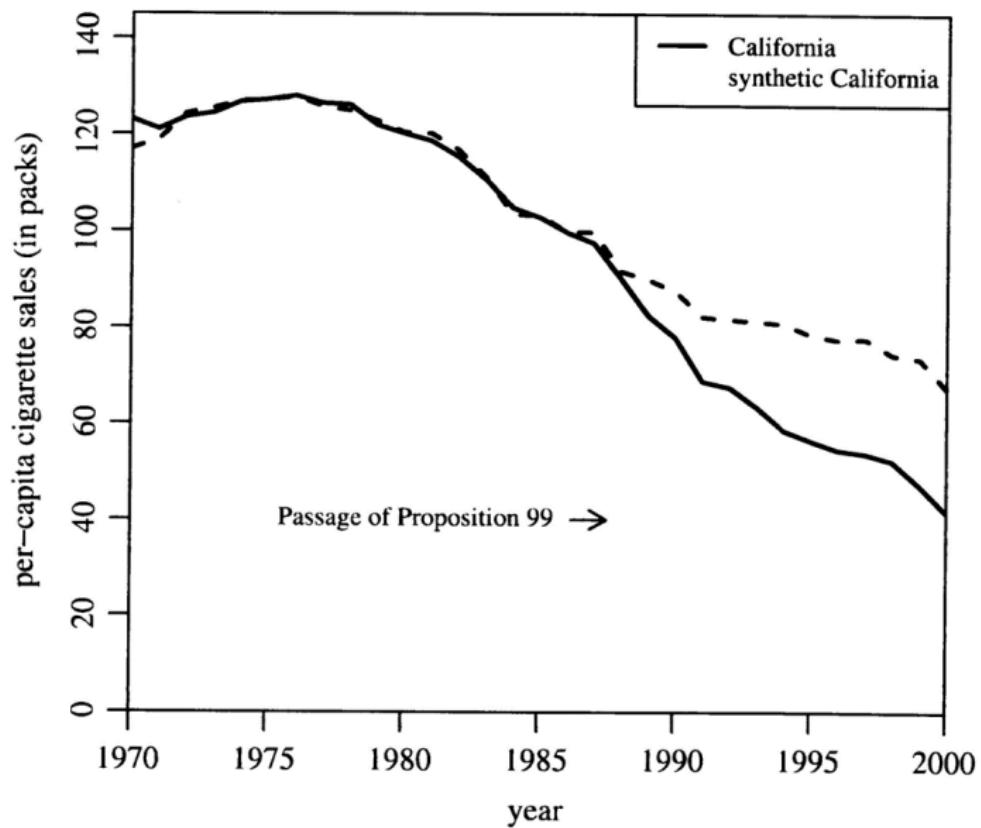


Controle Sintético: Aplicações

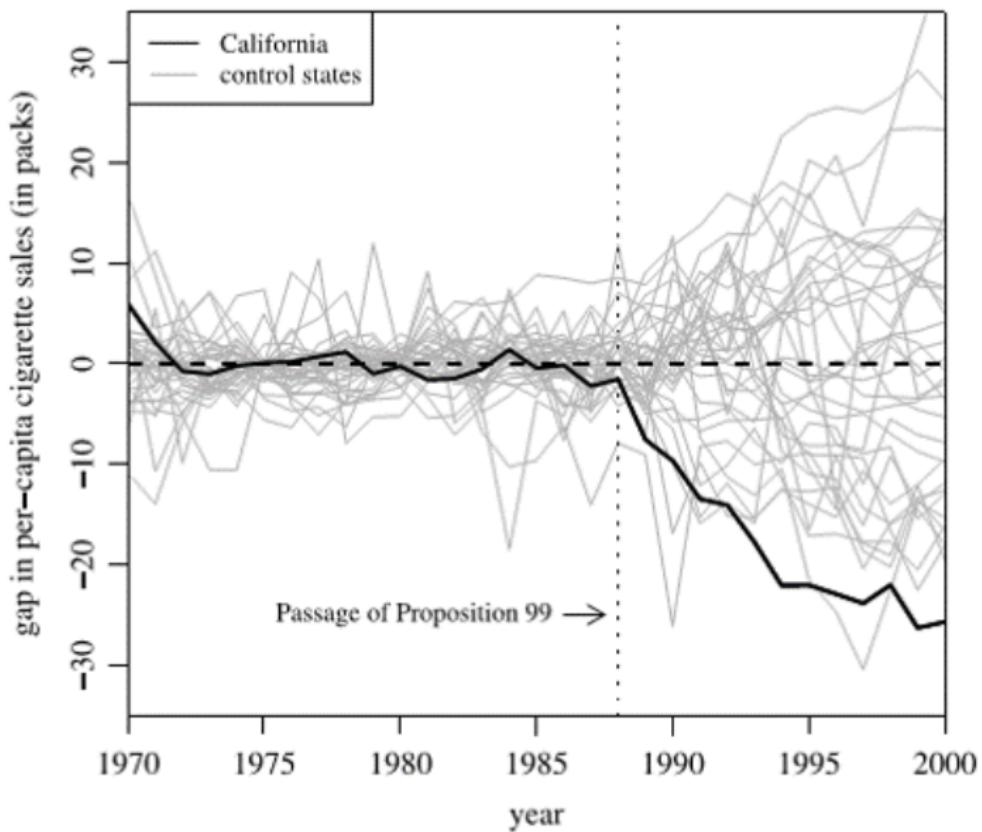
Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Controle Sintético: Aplicações



Controle Sintético: Aplicações



Resumo

Controle Sintético: Atentar para

- ▶ Número de períodos pré-tratamento (T_0) vs. Choques transitórios.
- ▶ Excluir controles com potencial contaminação do donor pool (efeitos indiretos destes subestimam o efeito do tratamento).
- ▶ Incluir controles com características parecidas/balanceadas com relação à unidade tratada (dentro do conjunto convexo).
- ▶ Excluir controles que sofreram grandes choques em Y no período.
- ▶ Testar diferentes inclusões de covariadas.

Controle Sintético: Inclusões de covariadas

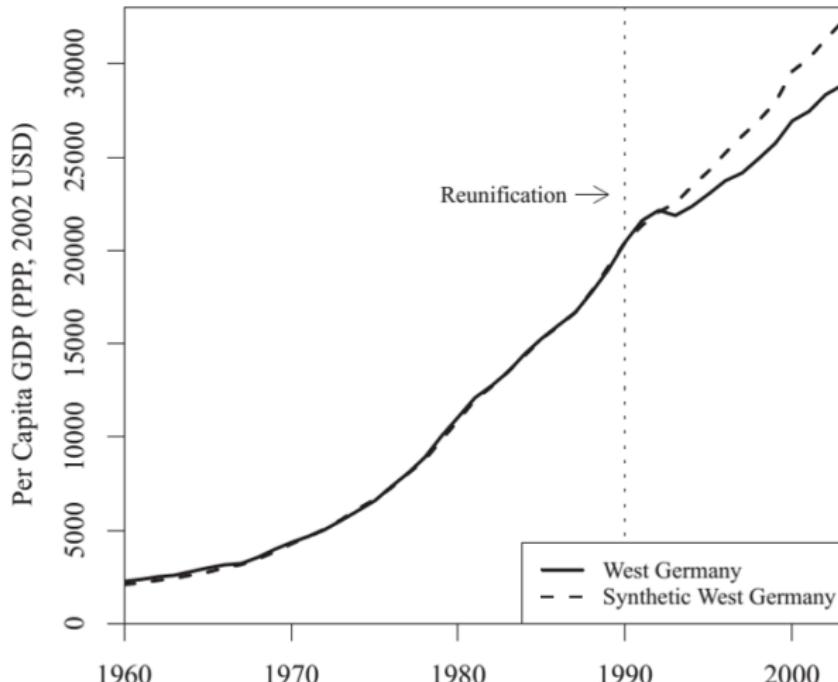
**TABLE 2 Economic Growth Predictor Means
before German Reunification**

	West Germany	Synthetic West Germany	OECD Sample
GDP per capita	15808.9	15802.2	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

Notes: GDP per capita, inflation rate, trade openness, and industry share are averaged for the 1981–90 period. Investment rate and schooling are averaged for the 1980–85 period. The last column reports a population-weighted average for the 16 OECD countries in the donor pool.

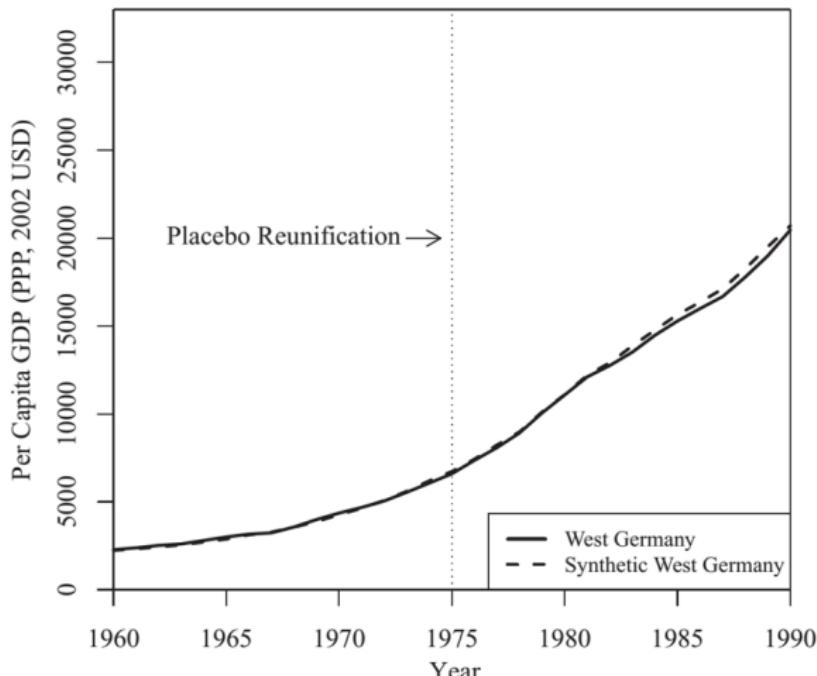
Controle Sintético: Gráfico do Efeito (exemplo)

FIGURE 2 Trends in per Capita GDP: West Germany versus Synthetic West Germany

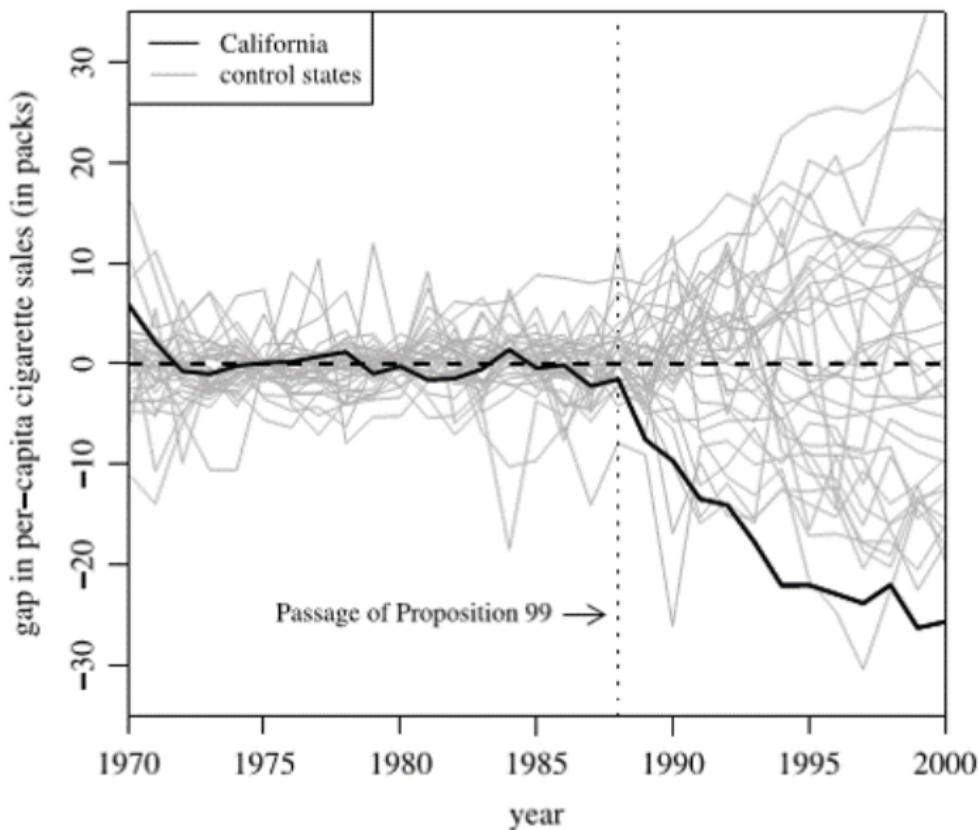


CS: Falsificação do Efeito - antecipação (exemplo)

FIGURE 4 Placebo Reunification 1975–Trends in per Capita GDP: West Germany versus Synthetic West Germany

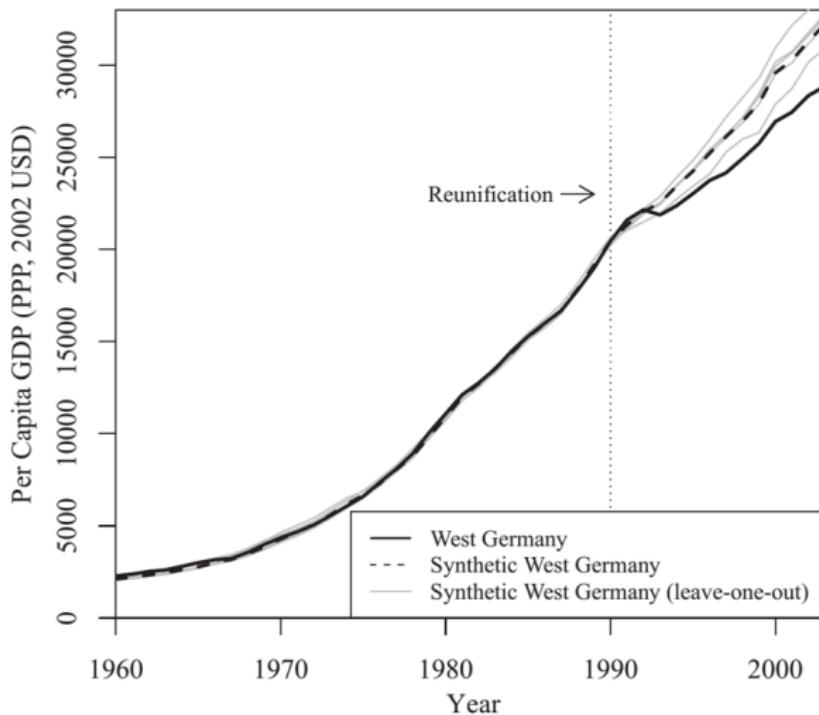


Controle Sintético: Teste de Permutação (exemplo)



Controle Sintético: Leave-one-out (exemplo)

FIGURE 6 Leave-One-Out Distribution of the Synthetic Control for West Germany



Obrigada!