Statistical Modeling & Statistical Causality

Programa C-Level Gestores na Educação

Hedibert Freitas Lopes

Professor de Estatística e Econometria

Coordenador do Núcleo de Ciências de Dados e Decisão

09/2025

V 1.0

https://hedibert.org/wp-content/uploads/2025/09/statistical-modeling-statistical-causality-byexamples.pdf

LISTA DE EXEMPLOS

https://hedibert.org/wp-content/uploads/2025/09/statistical-modeling-statistical-causality-byexamples.pdf

Statistical modeling

- Example 1: Height and weight by age group
- Example 1.5: Maps and heatmaps are also cool!
- Example 2: Birthday problem
- Example 3: Avaliando filmes (imdb.com)
- Example 4: Vacinação vs hospitalização (COVID-19)
- Example 5: Measuring vulnerability
- Example 6: Nowcasting

Statistical causality

- Example 7: Simpson's paradox
- Example 8: Exercise vs cholesterol
- Example 9: Drug vs recovery (by gender)
- Example 10: Education on wage (IV)
- Example 11: Return to education (IV)
- Example 12: Minimum wage on employment (DiD)
- Example 13: Birthdays and funerals (RDD)
- Example 14: Effects of semaglutide on weight
- Example 15: Impact of a full-time school program (DiD)
- Example 16: Effect of academic probation on education
- Example 17: Media manipulation

Example 1

In data we trust?
NO!

In reliable data we trust?
YES

The catch:
Define “reliable”.
Reliable in which sense?
Reliable for what purpose?

Accuracy of self-reported height, weight, and waist circumference in a general adult Chinese population
Popul Health Metr. 2016 Aug 11;14:30.
doi: 10.1186/s12963-016-0099-8

The difference between self-reported and measured (SR-DM) height, weight, BMI, and WC by sociodemographic factors in Jiangsu, China

		Height (cm)			Weight (kg)			BMI (kg/m^2)			WC (cm)		
		n	Mean	95 % CI	n	Mean	95 % CI	n	Mean	95 % CI	n	Mean	95 % CI
Total		5037	1.1	1.0,1.2	5867	−0.1	−0.2,0.0	4829	−0.4	−0.5,-0.3	3269	−1.5	−1.7,-1.3
Sex	Men	2377	1.1	0.9,1.2	2645	−0.1	−0.2,0.1	2278	−0.4	−0.4,-0.3	1285	−0.9	−1.2,-0.5
	Women	2660	1.2	1.1,1.3	3222	−0.1	−0.2,0.1	2551	−0.4	−0.5,−0.4	1984	−1.9	−2.1,−1.7
	P value	0.098			0.882			0.13			<0.001		
Age group (years)	18–29	569	0.5	0.4,0.7	594	−0.4	−0.8,0.1	534	−0.3	−0.5,−0.1	292	−1.2	−1.8,−0.6
	30–39	661	0.7	0.5,0.9	704	−0.6	−0.8,−0.3	637	−0.4	−0.5,−0.3	426	−1.9	−2.4,−1.5
	40–49	1221	0.9	0.8,1.0	1389	−0.1	−0.3,0.2	1169	−0.3	−0.4,−0.2	837	−1.9	−2.3,−1.5
	50–59	1271	1.2	0.9,1.4	1512	0.0	−0.2,0.1	1224	−0.4	−0.5,−0.3	851	−1.2	−1.5,−0.9
	60–69	893	1.6	1.4,1.8	1134	0.2	0.0,0.4	863	−0.5	−0.5,−0.4	598	−1.4	−1.9,−0.9
	70~	422	2.3	1.9,2.7	534	0.3	−0.2,0.7	402	−0.6	−0.9,−0.4	265	−1.0	−1.6,−0.3
	P value	< 0.001			0.002			0.025			0.015		
Locations	Urban	3982	1.2	1.1,1.3	4223	−0.2	−0.3,−0.1	3875	−0.4	−0.5,−0.4	2592	−1.3	−1.5,−1.1
	Rural	1055	1.0	0.8,1.2	1644	0.3	0.0,0.5	954	−0.2	−0.4,−0.1	677	−2.2	−2.7,−1.7
	P value	0.078			< 0.001			0.003			0.001		
Education	Primary	1495	1.4	1.2,1.7	2107	0.2	0.0,0.4	1392	−0.4	−0.5,−0.4	1027	−1.7	−2.1,−1.4
	Secondary	2994	1.1	1.0,1.2	3218	−0.1	−0.3,0.0	2902	−0.4	−0.5,−0.3	1927	−1.5	−1.7,−1.2
	University	548	0.6	0.4,0.8	542	−0.7	−1.0,−0.4	535	−0.4	−0.5,−0.3	315	−0.9	−1.4,−0.4
	P value	0.347			< 0.001			0.545			0.037		
Household income level	Low	1034	1.3	1.1,1.4	1392	0.3	0.0,0.6	962	−0.3	−0.5,−0.2	652	−1.6	−2.0,−1.1
	Moderate	2550	1.1	0.9,1.2	2911	−0.1	−0.3,0.0	2450	−0.4	−0.5,−0.3	1633	−1.6	−1.9,−1.4
	High	1453	1.1	1.0,1.3	1564	−0.3	−0.4,−0.1	1417	−0.4	−0.5,−0.4	984	−1.2	−1.5,−0.9
	P value	0.319			0.002			0.338			0.125		

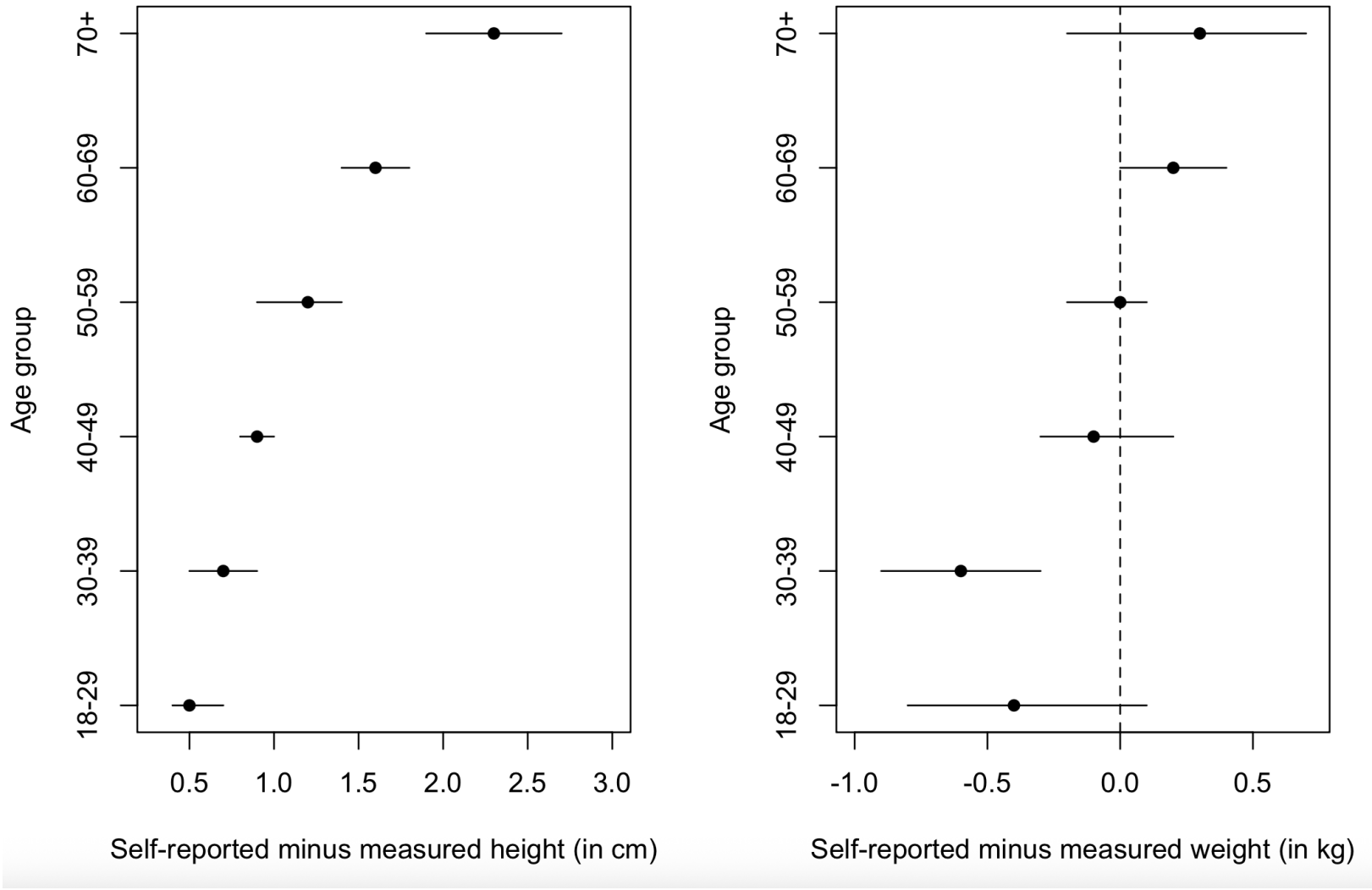
Abbreviations: BMI body mass index, *WC* waist circumference, *CI* confidence interval, *SR* self-reported, *DM* directly measured

Height and weight by age group

		Height (cm)			Weight (kg)			BMI (kg/m^2)			WC (cm)		
		n	Mean	95 % CI	n	Mean	95 % CI	n	Mean	95 % CI	n	Mean	95 % CI
Total		5037	1.1	1.0,1.2	5867	−0.1	−0.2,0.0	4829	−0.4	−0.5,-0.3	3269	−1.5	−1.7,-1.3
Sex	Men	2377	1.1	0.9,1.2	2645	−0.1	−0.2,0.1	2278	−0.4	−0.4,-0.3	1285	−0.9	−1.2,-0.5
	Women	2660	1.2	1.1,1.3	3222	−0.1	−0.2,0.1	2551	−0.4	−0.5,−0.4	1984	−1.9	−2.1,−1.7
	P value	0.098			0.882			0.13			<0.001		
Age group (years)	18–29	569	0.5	0.4,0.7	594	−0.4	−0.8,0.1	534	−0.3	−0.5,−0.1	292	−1.2	−1.8,−0.6
	30–39	661	0.7	0.5,0.9	704	−0.6	−0.8,−0.3	637	−0.4	−0.5,−0.3	426	−1.9	−2.4,−1.5
	40–49	1221	0.9	0.8,1.0	1389	−0.1	−0.3,0.2	1169	−0.3	−0.4,−0.2	837	−1.9	−2.3,−1.5
	50–59	1271	1.2	0.9,1.4	1512	0.0	−0.2,0.1	1224	−0.4	−0.5,−0.3	851	−1.2	−1.5,−0.9
	60–69	893	1.6	1.4,1.8	1134	0.2	0.0,0.4	863	−0.5	−0.5,−0.4	598	−1.4	−1.9,−0.9
	70~	422	2.3	1.9,2.7	534	0.3	−0.2,0.7	402	−0.6	−0.9,−0.4	265	−1.0	−1.6,−0.3
	P value	< 0.001			0.002			0.025			0.015		

Insper

Graphical summaries are impactful, but can be expensive (why?)

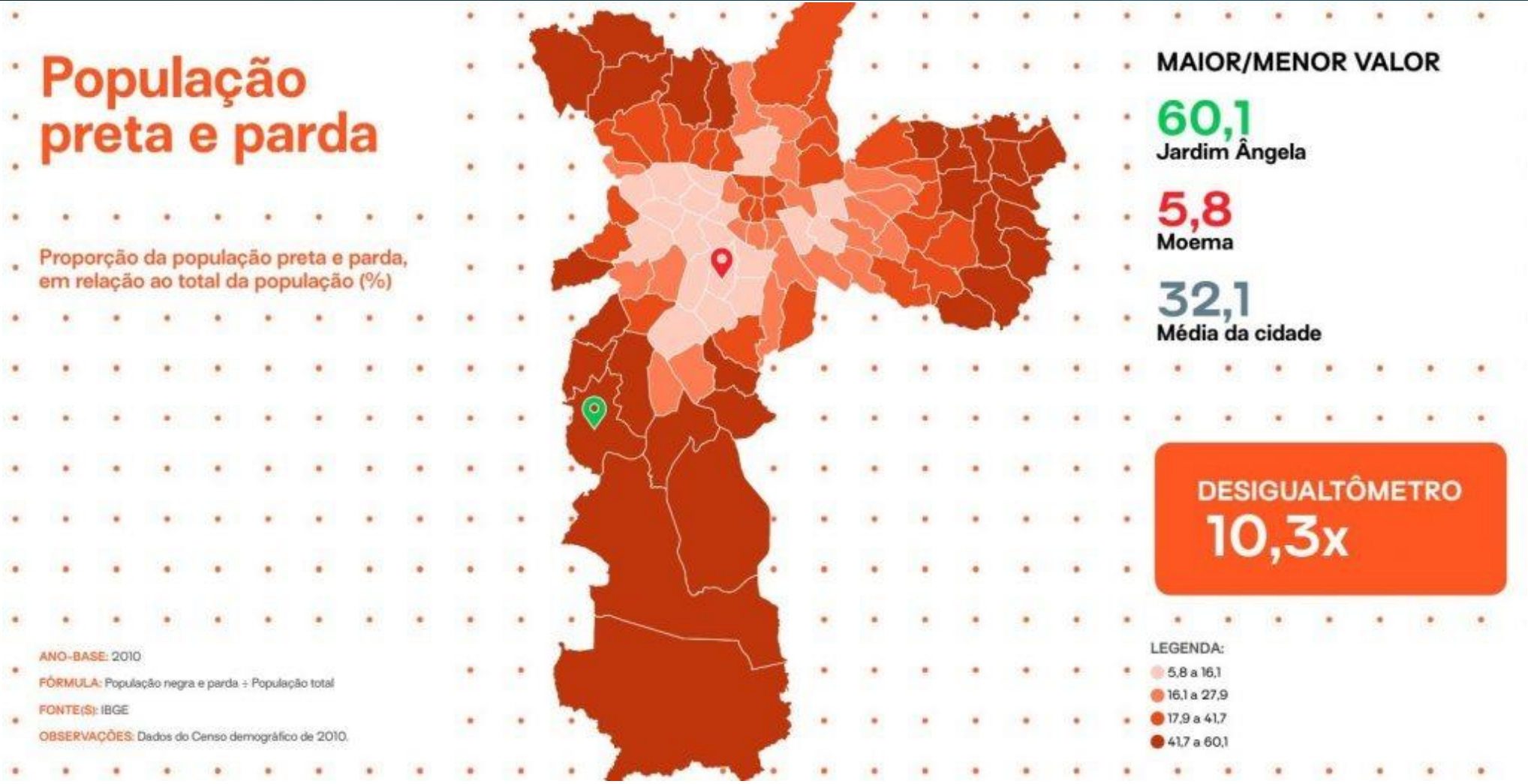


Insper

Takeaways

- ALWAYS verify the source and reliability of the data you intend to use
- Tables and graphs help in this initial exploratory data analysis
- Associate data, tables and graphs to the pursued scientific question
- Only then entertain possible statistical modeling and decision making

Example 1.5: Maps and heatmaps are also cool!



https://institutocidadessustentaveis.shinyapps.io/mapadesigualdadesaopaulo2024/

Insper

Idade média ao morrer

Média de idade com que as pessoas morreram

MELHOR/PIOR VALOR

80,6
Moema

57,3
Cidade Tiradentes

68,7
Média da cidade

DESIGUALTÔMETRO
1,4x

LEGENDA:
- 57 a 63
- 63 a 69
- 69 a 75
- 75 aa 81

ANO-BASE: 2018

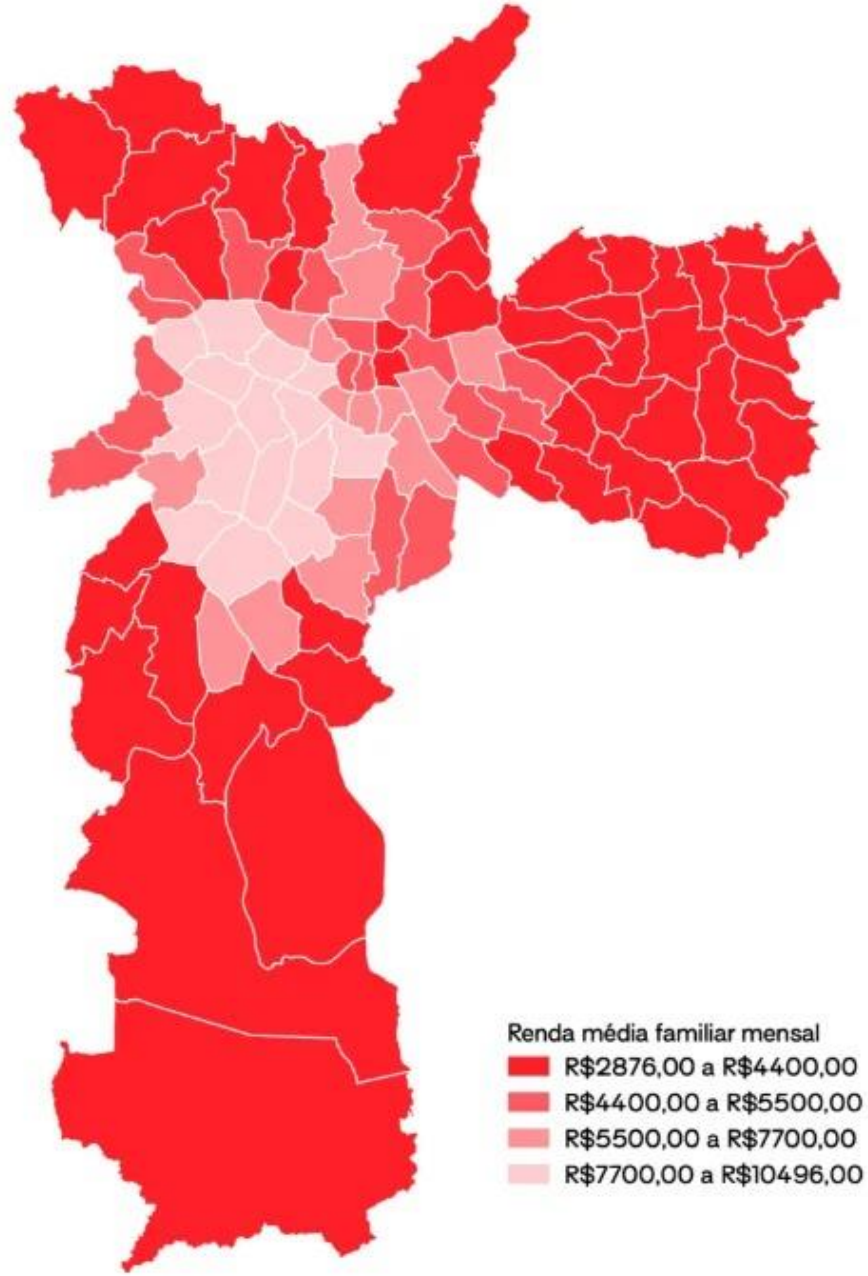
FÓRMULA: Soma das idades ao morrer ÷ Número total de óbitos

FONTE(S): SIM

OBSERVAÇÕES: Dados de 2018 tabulados em maio de 2019.

MAIS INFORMAÇÕES: Observatório Cidadão https://www.redesocialdecidades.org.br/br/SP/sao-paulo/regiao/+aricanduva/idade-media-ao-morrer

Insper



Mapa 2: Renda média familiar mensal, por distrito. Fonte: Pesquisa OD 2017. Elaboração: Rede Nossa São Paulo, 2021.

Capa Introdução Mapa Informações adicionais

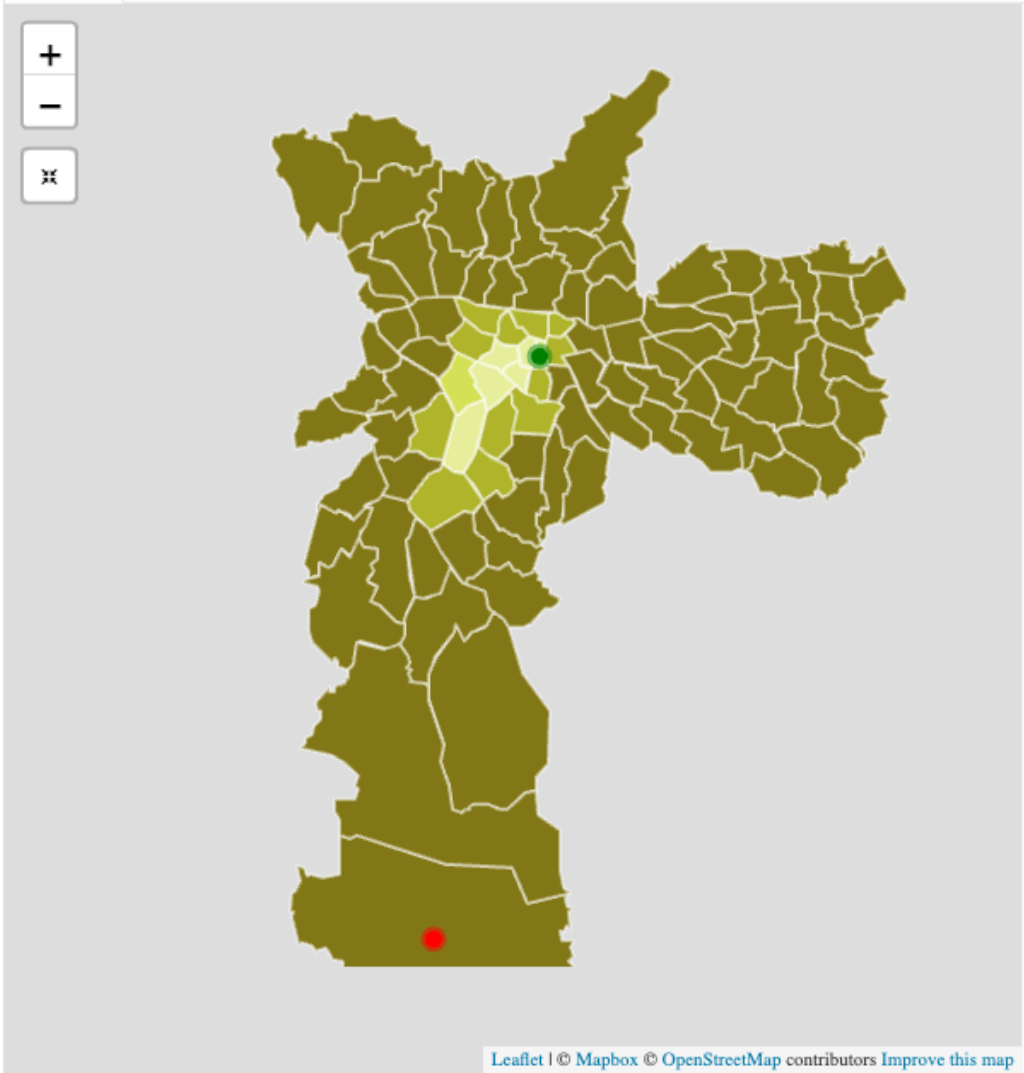
Capturar tela em Imagem

Indicador:

Acesso internet móvel (por área/km²)

Distribuição de antenas (Estações Radio-base - ERBs), por área (km²) dos distritos

ANO-BASE: 2024
FÓRMULA: Número total de antenas (Estações Radio-base - ERBs) ÷ área (km²) do distrito
FONTE(s): Anatel ; Infocidade/SMUL
ELABORAÇÃO: RNSP
Observações: (1) Foram consideradas apenas antenas dentro dos limites administrativos do Município de São Paulo, conforme os arquivos shapfiles disponíveis no portal Geosampa em 2021; (2) Os dados se referem apenas a oferta móvel de internet, através das antenas de Estação-rádio-base (ERB); (3) Há limitações em relação ao mapeamento detalhado da cobertura de internet disponível aos distritos administrativos, não estando disponíveis os microdados de acesso à outras tipoogias de infraestrutura de internet, como por exemplo: banda larga, fibra-ótica e dentre outras; (4) não foram considerados as estações com data de validade inferior a 2023-01-01.



MELHOR/PIOR VALOR

43,84
Sé

0,09
Marsilac

Média dos Distritos
8,18

DESIGUALTÔMETRO
480,4x

- 0,09 - 11,03
- 11,03 - 21,96
- 21,96 - 32,9
- 32,9 - 43,84

Insper

Capa Introdução Mapa Informações adicionais

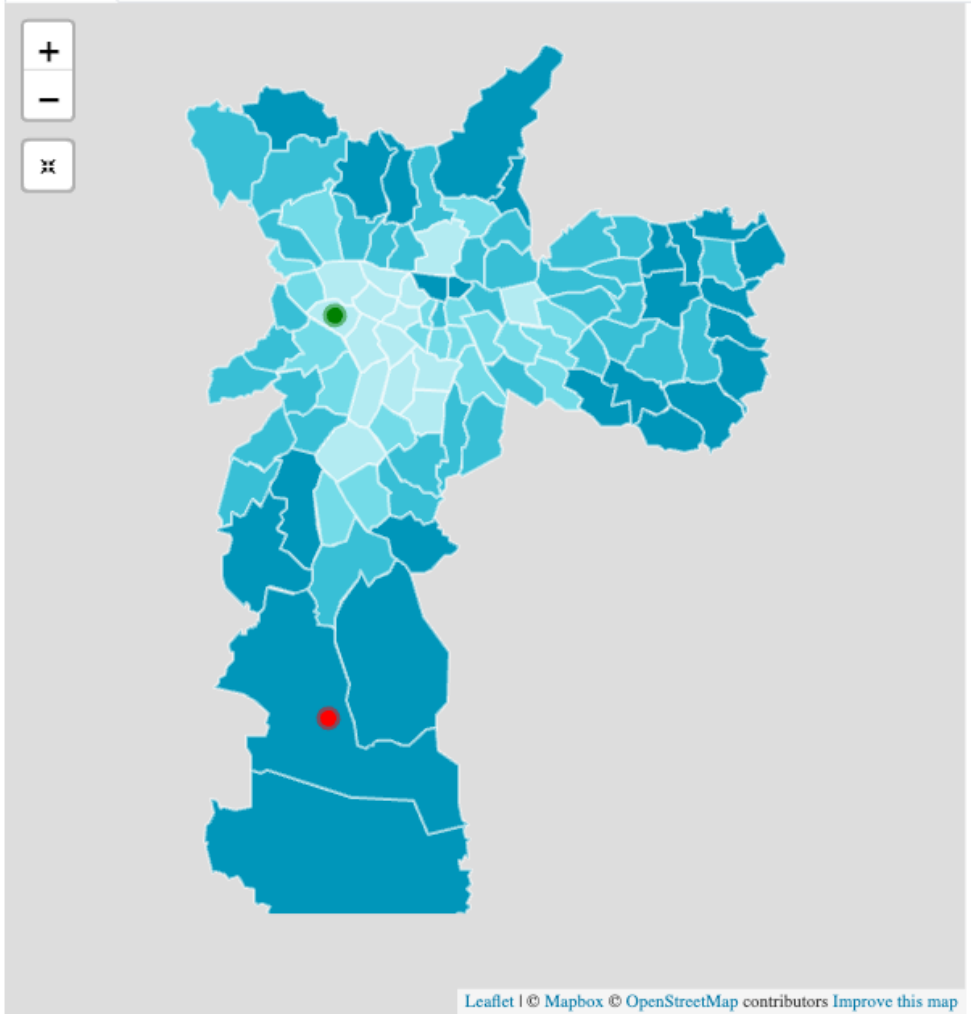
Capturar tela em Imagem

Indicador:

Gravidez na adolescência

Proporção (%) de nascidos vivos de parturientes com menos de 20 anos em relação ao total de nascidos vivos

ANO-BASE: 2023
FÓRMULA: Número de nascidos vivos de parturientes com menos de 20 anos ÷ Número total de nascidos vivos x 100
FONTE(s): SINASC/CEInfo/SMS-SP
ELABORAÇÃO: RNSP
Observações: Atualizado em 08/08/2024.



MELHOR/PIOR VALOR

0

Alto de Pinheiros

11,26

Parelheiros

Média dos Distritos

6,3

DESIGUALTÔMETRO

81x

- 0 - 2,81
- 2,81 - 5,63
- 5,63 - 8,44
- 8,44 - 11,26

Insper

Dia do mês de maio	Aniversariantes
18	4
19	3
20	3
21	5
22	3
23	3
24	1
25	7
26	5
27	1
28	2
29	3
30	4
31	5

Example 2: Probabilidade pode te enganar!

Tenho cerca de 1.500 “amigos” no FB.

Com 100% de chance, num universo de 1500 pessoas, pelo menos duas fazem aniversário no mesmo dia.

Não precisamos de probabilidade pra saber disso!

Pergunta: E se ao invés de 1.500 pessoas, estivéssemos numa sala de aula de um curso de graduação do INSPER com 40 alunos (Estatística I ou Introdução à Economia, por exemplo), qual a chance de pelo menos duas pessoas fazerem aniversário no mesmo dia e no mesmo mês?

(a) 0% (b) 10% (c) 50% (d) 90% (e) 100%

E se tivéssemos 70 alunos?

Evento A: “Pelo menos 2 pessoas aniversariando no mesmo dia e no mesmo mês do ano”.

Evento A^c: “Nenhuma das **n** pessoas aniversariam no mesmo dia e no mesmo mês do ano”.

Então o que estamos procurando é $\Pr(A) = 1 - \Pr(A^C)$.

$$\Pr(A^c) = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

$$= \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$

$$= \frac{365!}{365^n (365 - n)!} = \frac{n! \cdot \binom{365}{n}}{365^n}$$

n	Pr(A)
10	11.7%
20	41.1%
23	50.7%
30	70.6%
40	**89.1%**
50	97.0%
60	99.4%
70	99.9%
75	99.97%

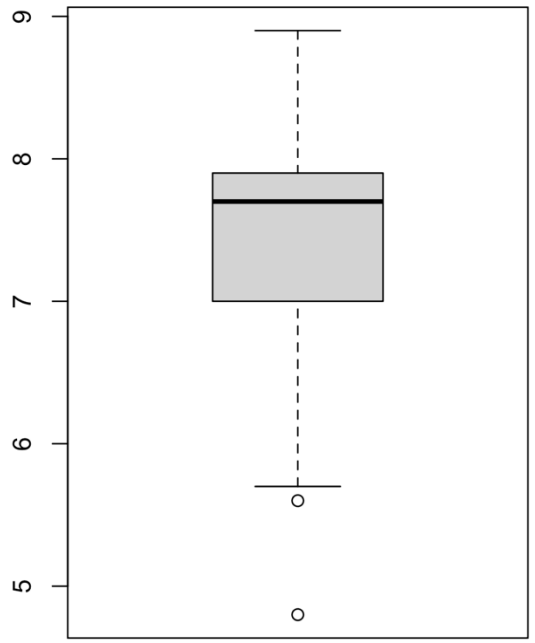
Takeaways

Sofremos muito ainda em quantificar o grau de confiança em eventos aleatórios.

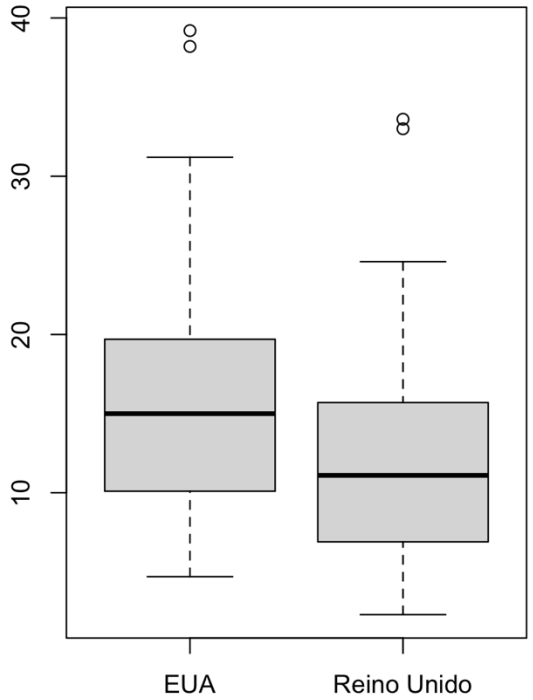
Somos inconsistentes em relação à probabilidade: subestimamos ou superestimamos.

A Teoria (e prática) de probabilidade permite o acompanhamento racional da combinação DADOS x MODELOS.

Exemplo 3: Avaliando filmes (www.imdb.com)



Médias IMDB



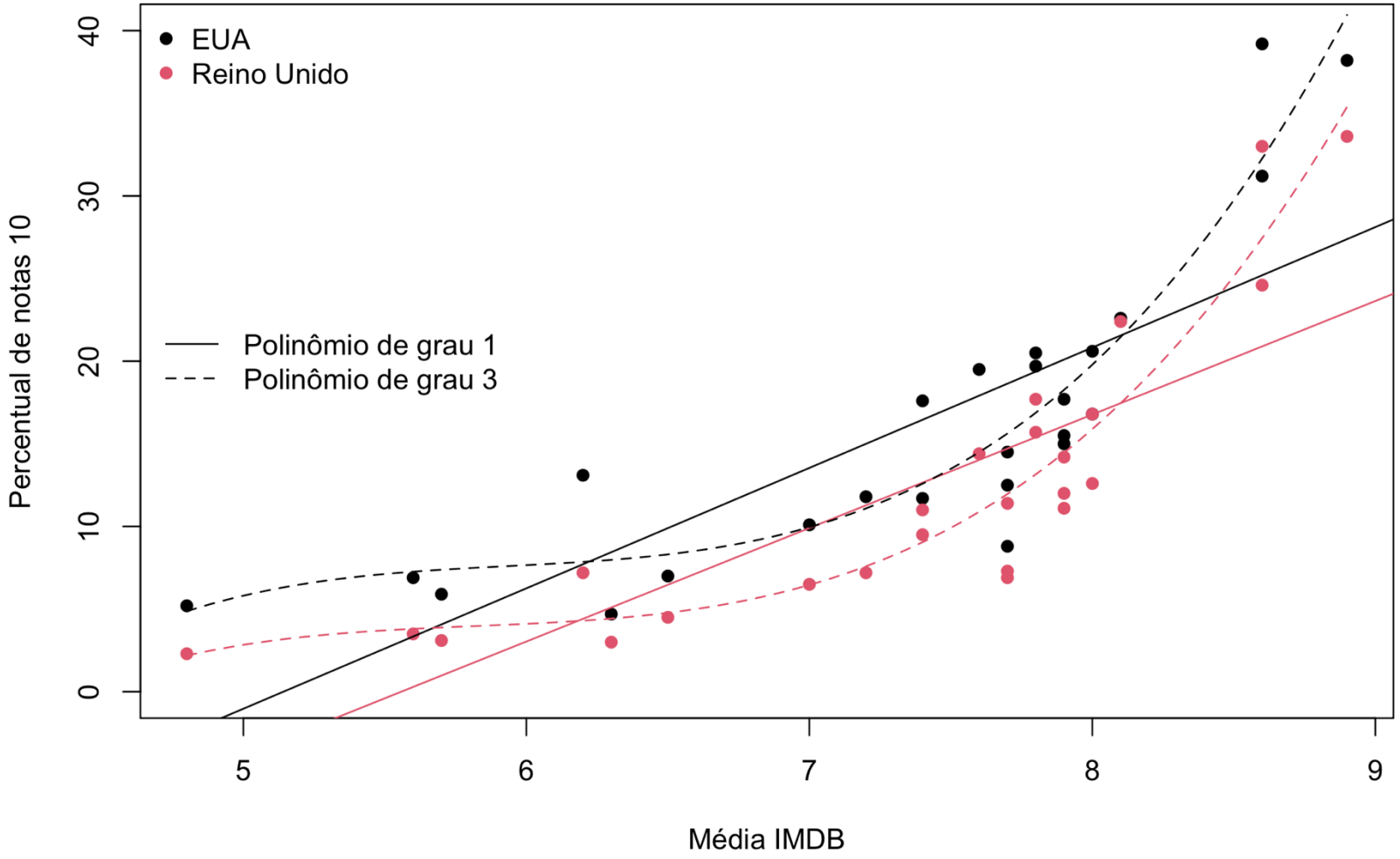
Percentual de notas 10 (máxima)

	Média IMDB	Nota 10 USA (%)	Nota 10 UK (%)
[1,]	4.8	5.2	2.3
[2,]	5.6	6.9	3.5
[3,]	5.7	5.9	3.1
[4,]	6.2	13.1	7.2
[5,]	6.3	4.7	3.0
[6,]	6.5	7.0	4.5
[7,]	7.0	10.1	6.5
[8,]	7.2	11.8	7.2
[9,]	7.4	11.7	9.5
[10,]	7.4	17.6	11.0
[11,]	7.6	19.5	14.4
[12,]	7.7	8.8	6.9
[13,]	7.7	12.5	7.3
[14,]	7.7	14.5	11.4
[15,]	7.8	20.5	17.7
[16,]	7.8	19.7	15.7
[17,]	7.9	15.5	12.0
[18,]	7.9	15.0	11.1
[19,]	7.9	17.7	14.2
[20,]	8.0	20.6	16.8
[21,]	8.0	16.8	12.6
[22,]	8.1	22.6	22.4
[23,]	8.6	39.2	33.0
[24,]	8.6	31.2	24.6
[25,]	8.9	38.2	33.6

Regressão linear

Regressão polynomial

Variáveis dummies



Insper

Takeaways

- Regressão explora associação entre variáveis
- Sem hipóteses adicionais, regressão não evidencia relação causal
- Dados observacionais são inundados de viéses e podem comprometer estudos causais.
- Randomized control trials (RCT) são poderosas ferramentas para elucidar efeitos de tratamentos.

Exemplo 4:

Quantas hospitalizações a vacinação contra a COVID-19 já preveniu em São Paulo?

Izbicki, Bastos, Izbicki, Lopes and Santos, Clinics (Sao Paulo). 2021; 76: e3250, Publicado online 27/Agosto/2022. doi: 10.6061/clinics/2021/e3250

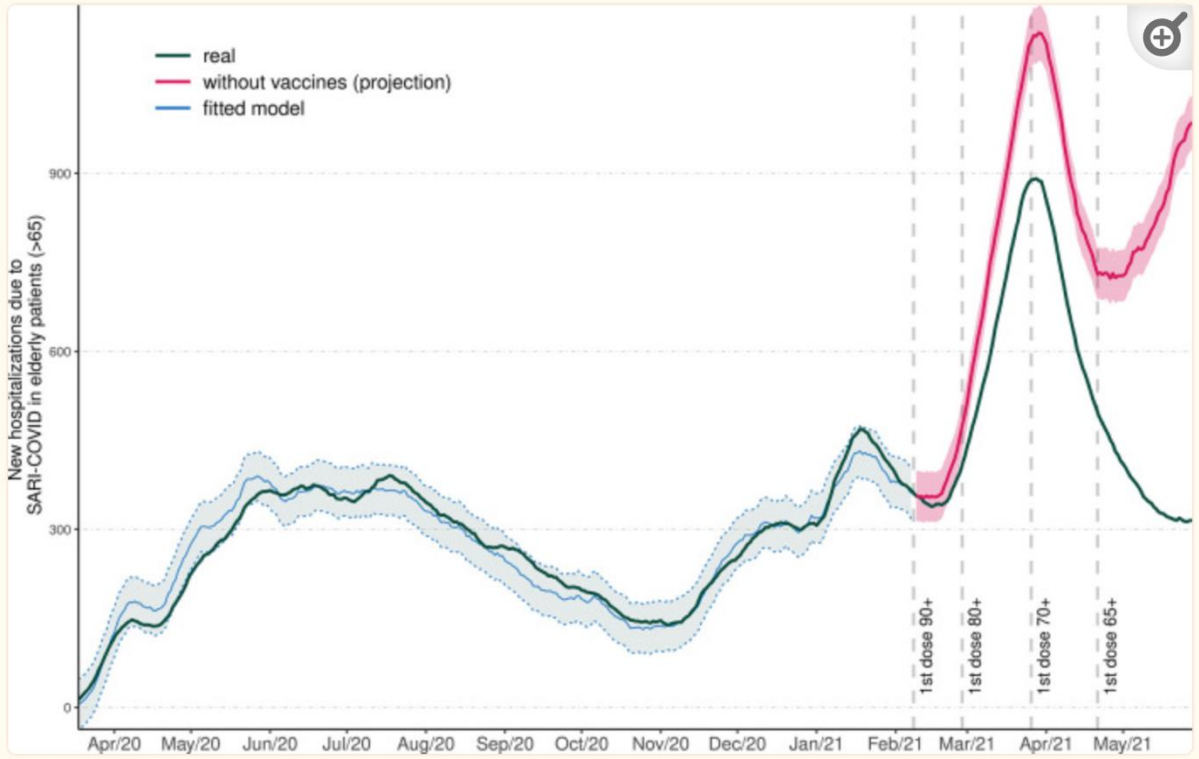


Figure 1

Number of hospitalizations due to SARI-COVID in patients aged >65 years (dark green), fitted pre-vaccination model (blue), and estimated counterfactual curve for the setting without vaccines (red).

Insper

Takeaways

Usando as áreas entre as curvas, observamos que aproximadamente 24.364 hospitalizações foram evitadas em São Paulo devido à vacinação antes de 28 de maio de 2021.

Considerando que a taxa de mortalidade estimada de pacientes hospitalizados com mais de 65 anos com SARI-COVID-19 é de aproximadamente 45%, cerca de 10.964 mortes podem ter sido prevenidas pela vacinação durante esse período.

Usando estimativas dos custos médios de hospitalização por COVID-19, aproximadamente 297 milhões de dólares podem ter sido economizados.

Esse valor é suficiente para comprar quase 30 milhões de doses adicionais de vacina (a 10 dólares cada).

Example 5: Measuring vulnerability

The Annals of Applied Statistics
2012, Vol. 6, No. 1, 284–303
DOI: 10.1214/11-AOAS497

MEASURING THE VULNERABILITY OF THE URUGUAYAN POPULATION TO VECTOR-BORNE DISEASES VIA SPATIALLY HIERARCHICAL FACTOR MODELS

BY HEDIBERT F. LOPES[1], ALEXANDRA M. SCHMIDT[2], ESTHER SALAZAR[3], MARIANA GÓMEZ AND MARCEL ACHKAR

University of Chicago, Universidade Federal do Rio de Janeiro, Duke University, Universidad de la República and Instituto Nacional de Salud Pública de Mexico, and Universidad de la República

Insper

FIG. 1. *Map of Uruguay with department boundaries and capitals together with the census tracts of Melo.*

TABLE 1

Description of the $p = 11$ variables, observed in the census tract level of the departmental capitals, to build the vulnerability index of the population of Uruguay to vector-borne diseases

Level of vulnerability	**Variables**
Personal characteristic	Illiteracy rate (ILL)
	Population with access to public health care (PHC)
	Male without formal jobs (UQW)
Household characteristic	Owed houses (OWH)
	Households headed by a woman (WHF)
	Households without sewage system (AHS)
	Average number of persons per household (APH)
	Households with more than two persons per room (OVC)
	Households without access to treated, drinkable water (ADW)
	Households with air conditioner (ACO)
	Households poorly built (HOQ)

Our proposed model

i: departamental capital (I=19)
j: census tract
k: region-specific measurement (p=11)

$$y_{ijk} = \mu_k + \beta_k f_{ij} + \sigma_k \varepsilon_{ijk}, \qquad k = 1, \ldots, p,$$

$$f_{ij} = \theta_i + \tilde{f}_{ij} + \sqrt{\omega_i} u_{ij},$$

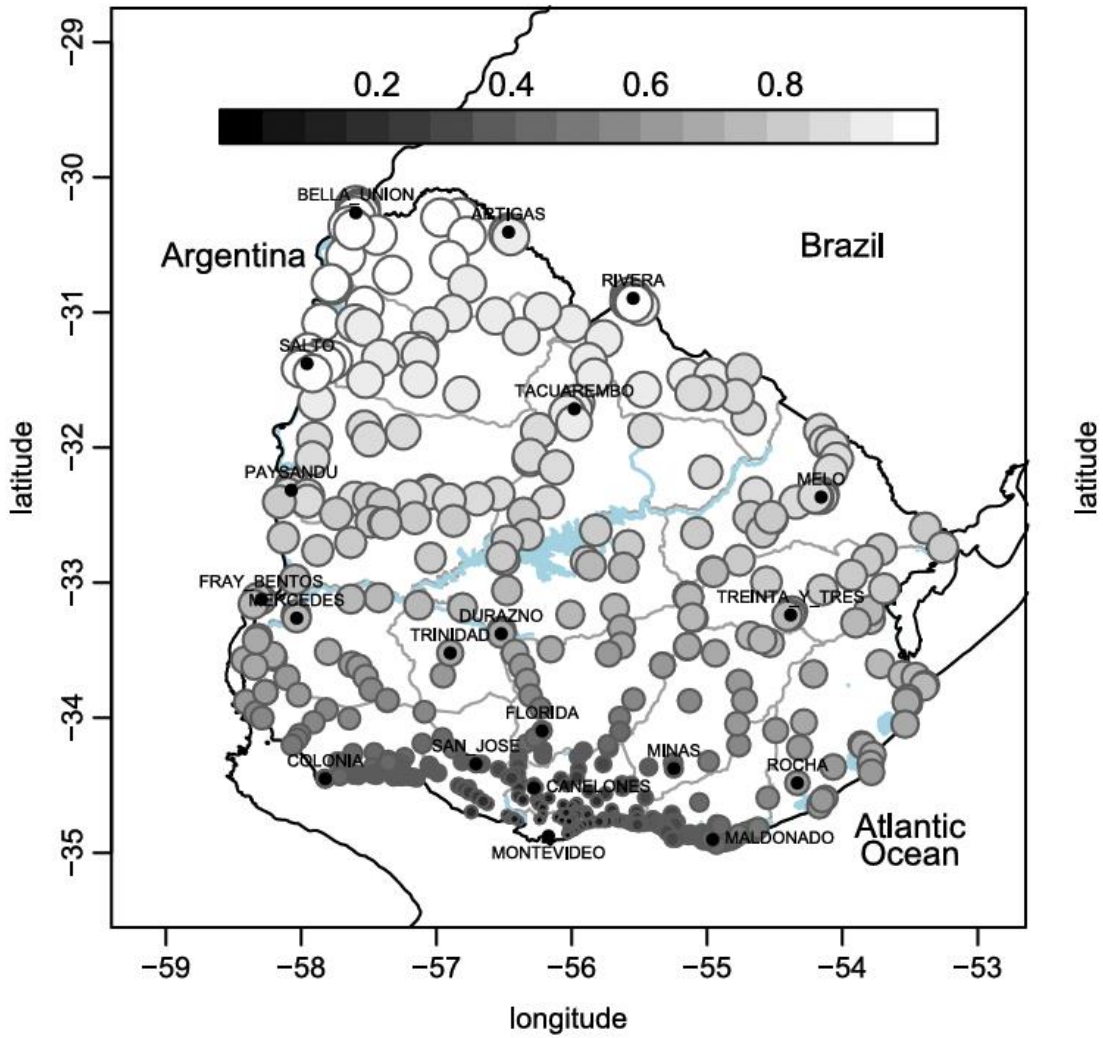
$$\tilde{f}_i \sim N(0, \tau_i^2 P_i),$$

$$\theta \sim N(1_I \theta_0, \delta^2 H(\lambda)),$$

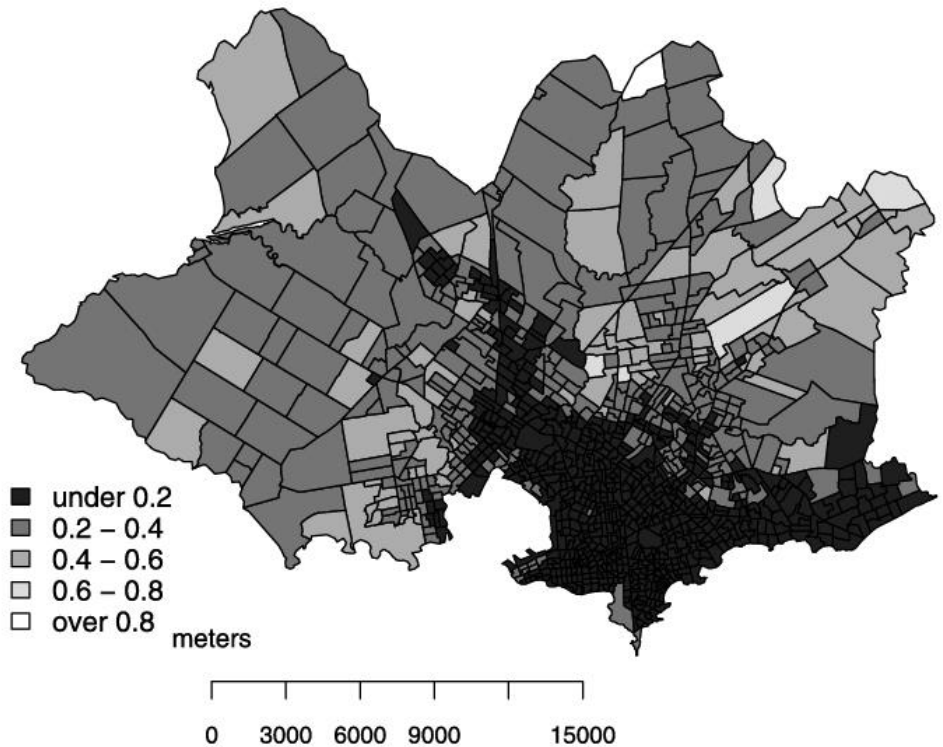
It pays to be Bayes!

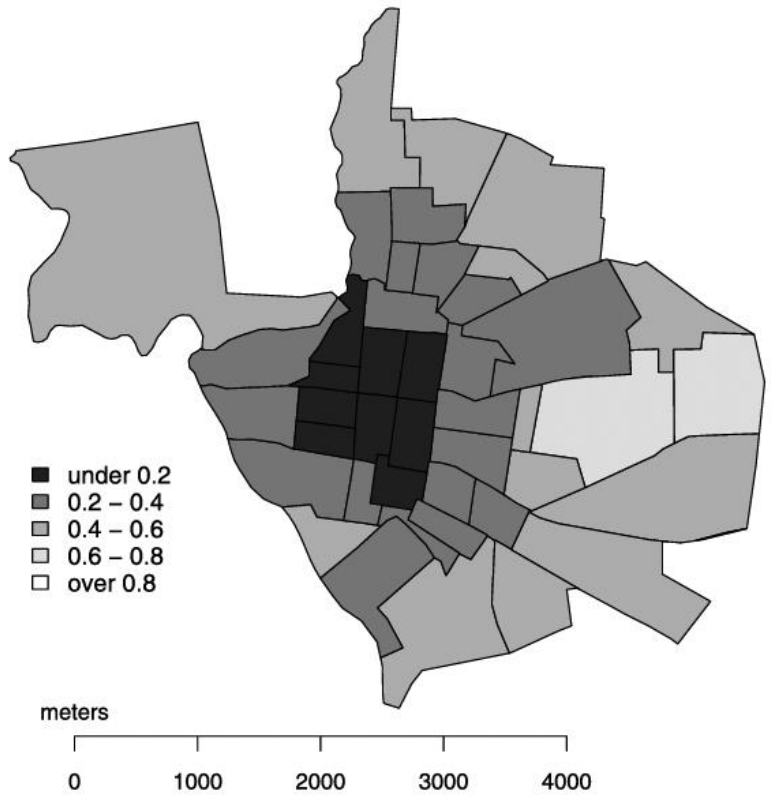
Insper

The darker the less vulnerable



Insper


under 0.2
0.2 – 0.4
0.4 – 0.6
0.6 – 0.8
over 0.8
meters
0
3000
6000
9000
15000


under 0.2
0.2 – 0.4
0.4 – 0.6
0.6 – 0.8
over 0.8
meters
0
1000
2000
3000
4000

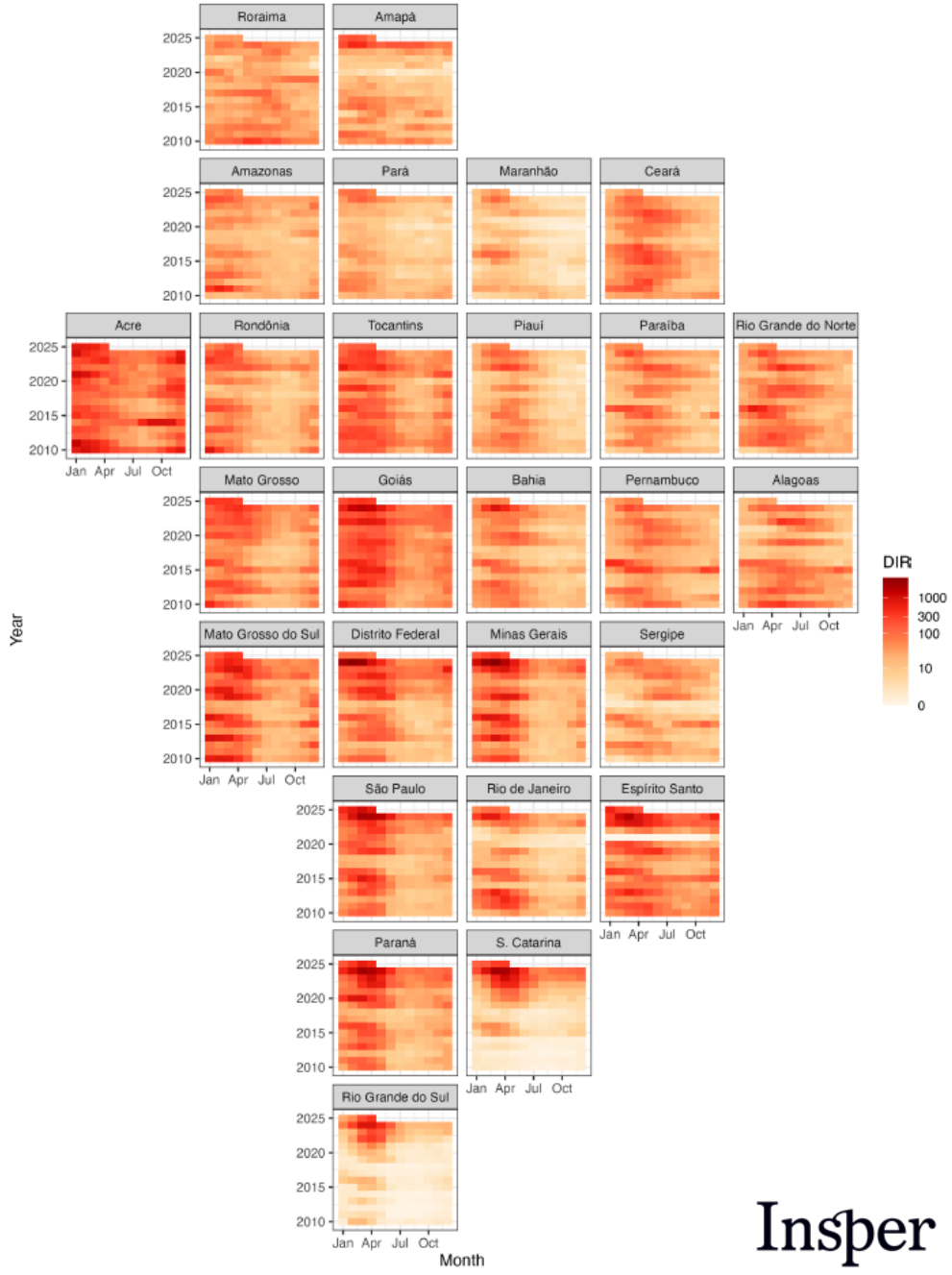
Rank
1
3
5
7
9
11
13
15
17
19
103
1
20
34
21
33
36
31
28
27
35
39
22
29
72
43
38
45
84
11
Montevideo
Canelones
San Jose
Colonia
Minas
Maldonado
Florida
Rocha
Trinidad
Durazno
Mercedes
Fray Bentos
Treinta y Tres
Paysandu
Melo
Tacuarembo
Rivera
Salto
Bella Union

Example 6: Nowcasting

Dengue nowcasting in Brazil by combining official surveillance data and Google Trends information.

Xiao Y, Soares G, Bastos L, Izbicki R, Moraga P (2025)
PLoS Neglected Tropical Diseases 19(8): e0012501.
https://doi.org/10.1371/ journal.pntd.0012501

- Nowcasting methods are needed to estimate underreported cases, enabling more timely decision-making.
- This study evaluates the value of using Google Trends indices of dengue-related keywords to complement official dengue data for nowcasting dengue in Brazil.
- The study demonstrates the value of digital data sources in enhancing dengue nowcasting, and emphasizes the value of integrating alternative data streams into traditional surveillance systems for better-informed decision-making.



Dengue incidence rate (cases per 100k people) on a log10 scale in Brazilian states from January 2010 to April 2025, showing a seasonal pattern, with outbreaks typically occurring between January and May.

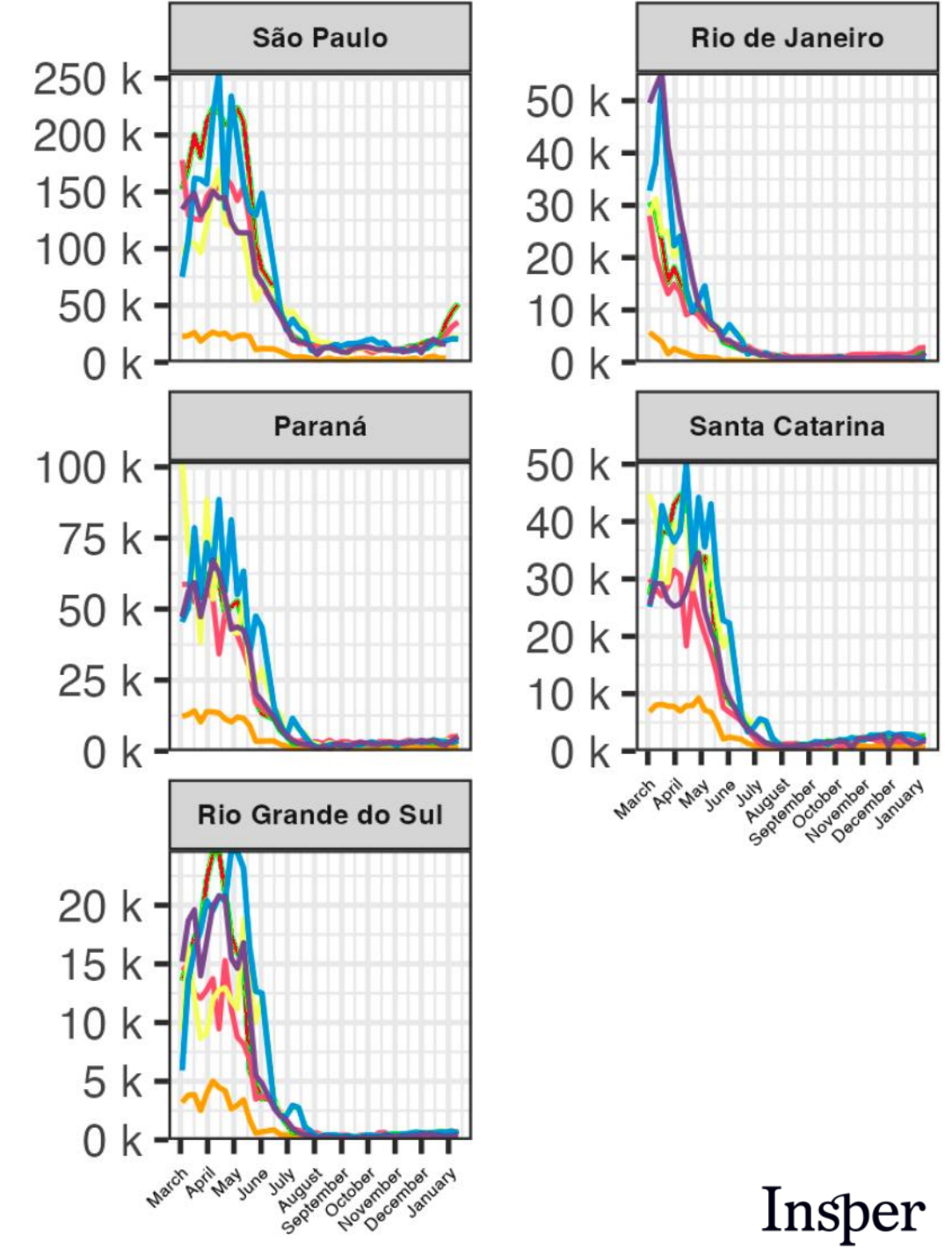
Insper

Predictions

The baseline represents the true number of cases (reported after a 15-week delay) used as the benchmark.

The orange line shows the suspected cases that are reported each week, reflecting reporting delays.

Overall, Goggle Trends (GT) and InfoDengue provide the most accurate forecasts.



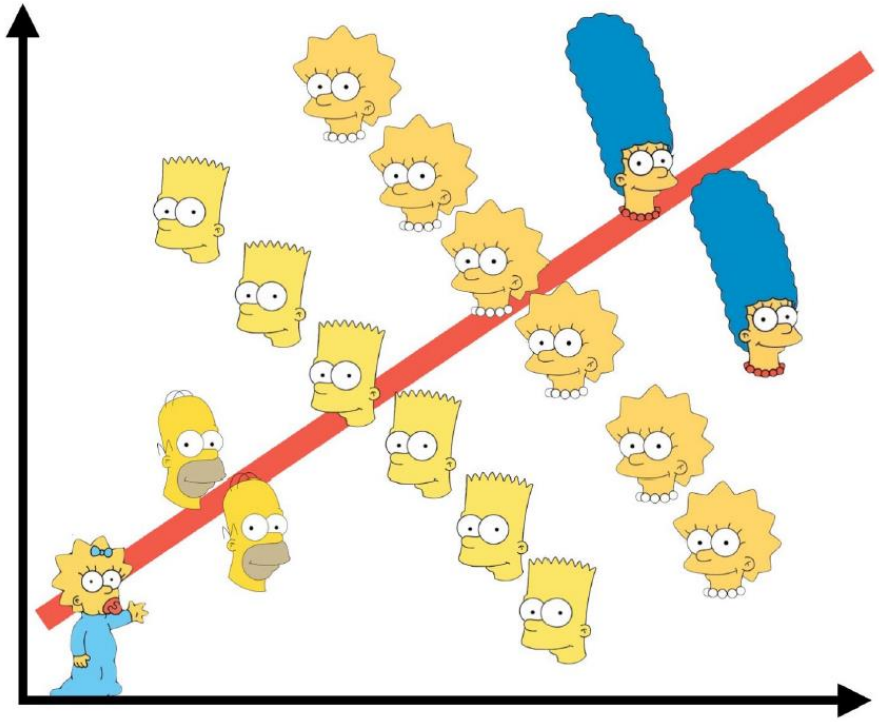
Insper

Slides are downloadable

https://hedibert.org/wp-content/uploads/2025/09/statistical-m statistical-causality-byexamples.pdf

Insper

Let us talk about STATISTICAL CAUSALITY



THE SIMPSONS PARADOX

The following slides are based on Pearl, Glymour and Jewell's Causal Inference in Statistics: Chapters 1 and 2.

Example 7: Simpson's paradox

Named after Edward Simpson (born 1922), the statistician who first popularized it, the paradox refers to the existence of data in which a statistical association that holds for an entire population is reserved in every subpopulation."

"We record the recovery rates of 700 patients (343 women and 357 men) who were given access to the drug. A Total of 350 patients chose to take the drug and 350 patients did not."

	Drug	No drug
Patients	273 out of 350 - 78%	289 out of 350 - 83%

Total and percentage of recovered.

Question: Based on this data, should a doctor recommend the drug or not?

What if you have additional information about blood pressure?

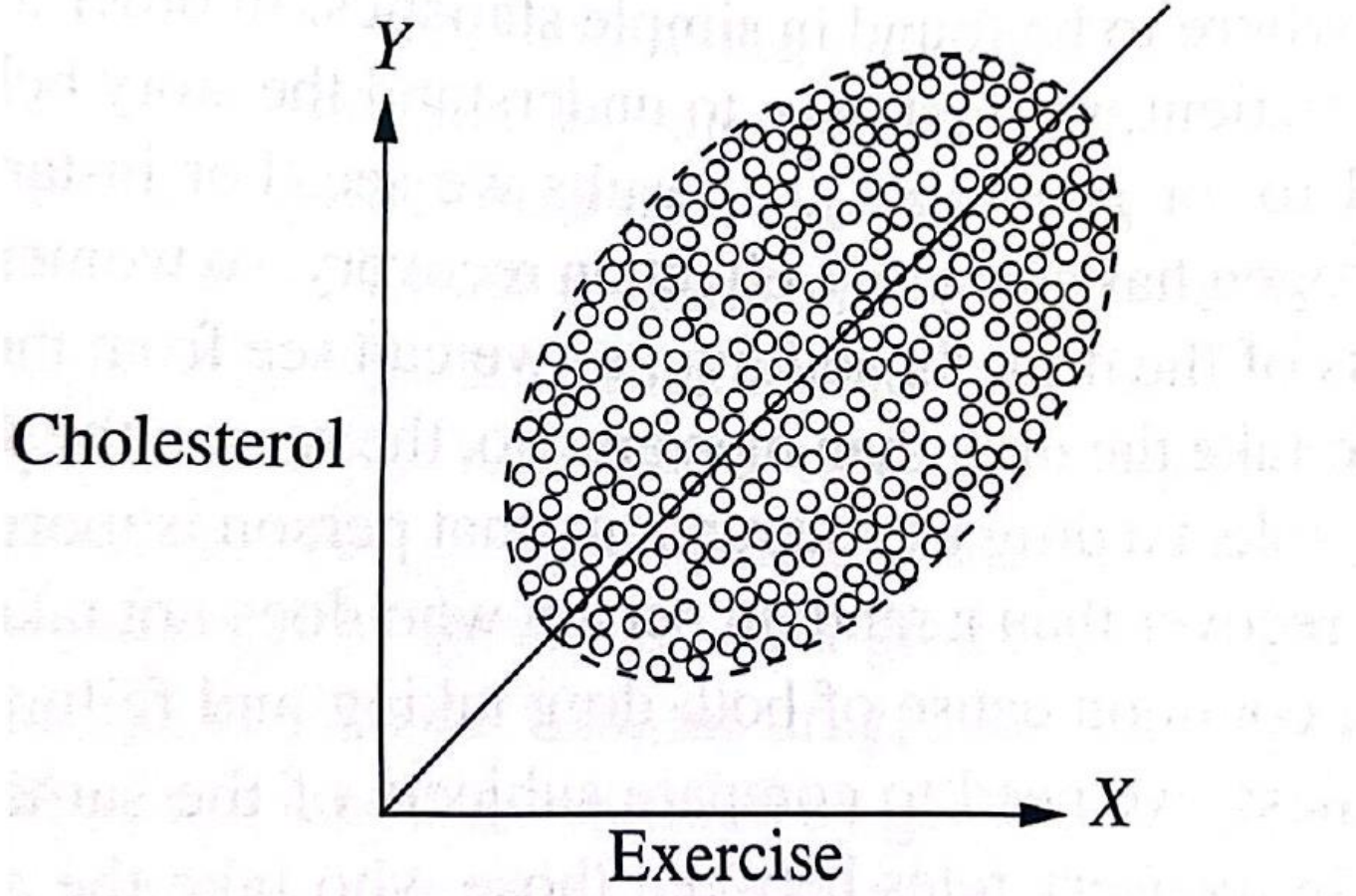
	Drug	No drug
Low BP	81 out of 87 - 93%	234 out of 270 - 87%
High BP	192 out of 263 - 73%	55 out of 80 - 69%

Total and percentage of recovered.

Question: Based on this data, should a doctor recommend the drug or not?

Example 8:

The more a person exercises, the higher their cholesterol is!



Why study causation?

Because we need

- To make sense of data,
- To guide actions and policies, and
- To learn from our successes and failures.

We need to estimate the effect of

- Smoking on lung cancer;
- Education on salaries;
- Carbon emissions on the climate;
- Turning 21 years old and increase of car accidents.

We need to understand HOW and WHY causes influence effects”

Numerous applications: health, labor, education, transportation, marketing, public economics, etc.

Example 9

Women are significantly more likely to take the drug than men are.

	Drug	No drug
Men	81 out of 87 - 93%	234 out of 270 - 87%
Women	192 out of 263 - 73%	55 out of 80 - 69%
Combined	273 out of 350 - 78%	289 out of 350 - 83%

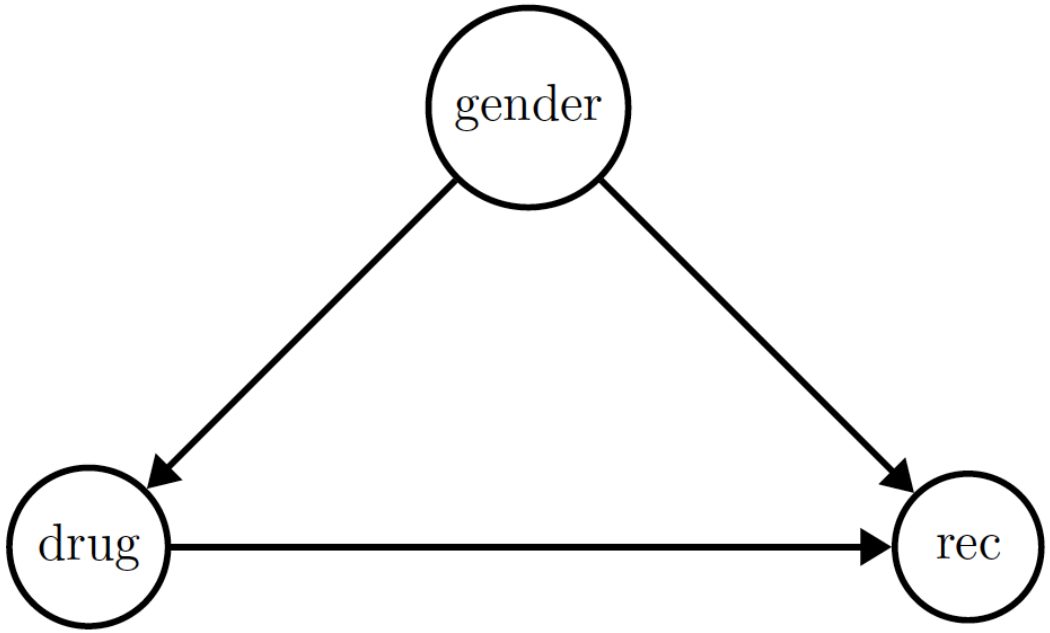
Total and percentage of recovered.

The data seem to say that

if we know the patient's gender we can prescribe the drug, but

if the gender is unknown we should not.

Obviously, that conclusion is ridiculous!

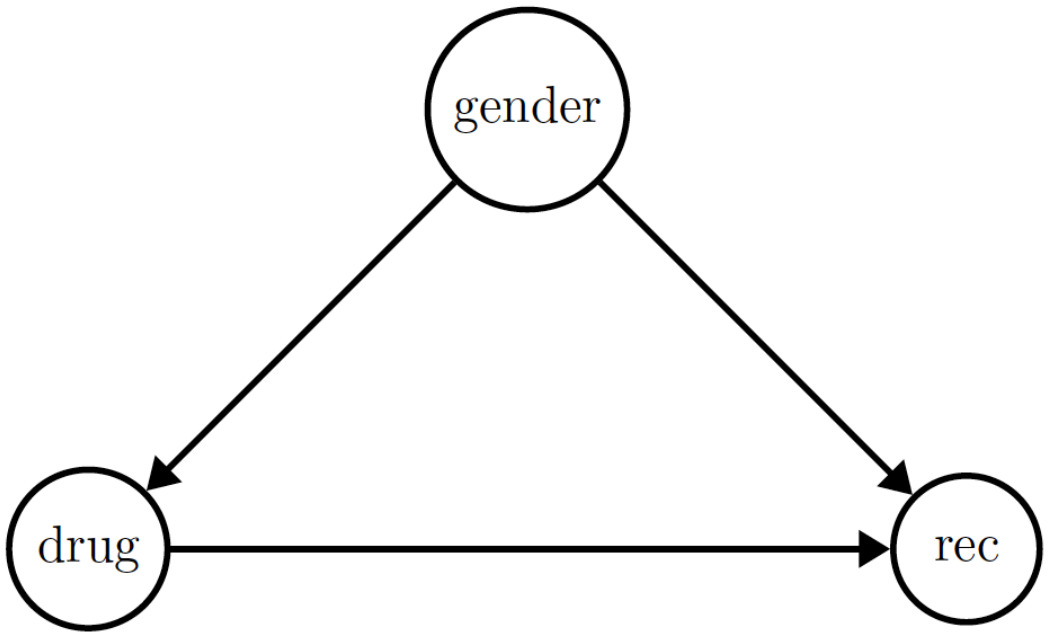


The effect of the drug on the patient's recovery is **CONFOUNDED** by the patient's gender

Suppose we knew an additional fact:
Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug.

The answer is nowhere to the found in simple statistics.

In order to decide whether the drug will harm or help a patient, **we first have to understand the story behind the data** -- the causal mechanism that let to, or generated, the results we see.

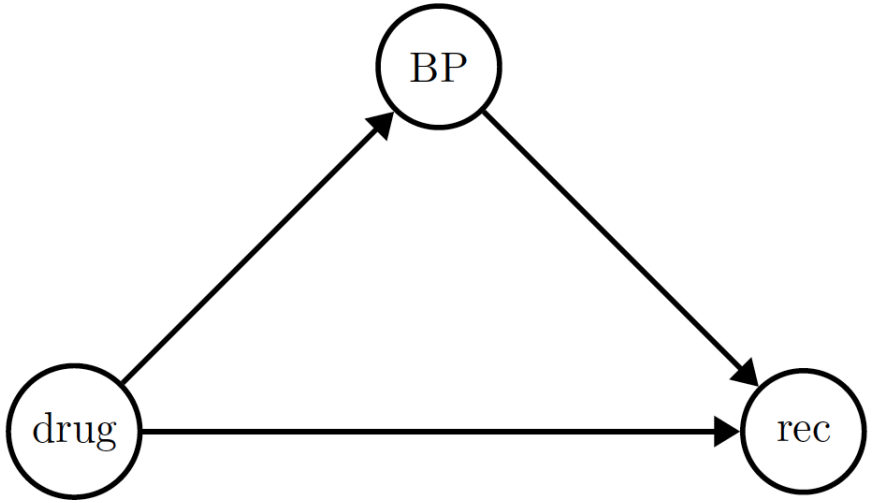


Insper

Examples 7 and 9

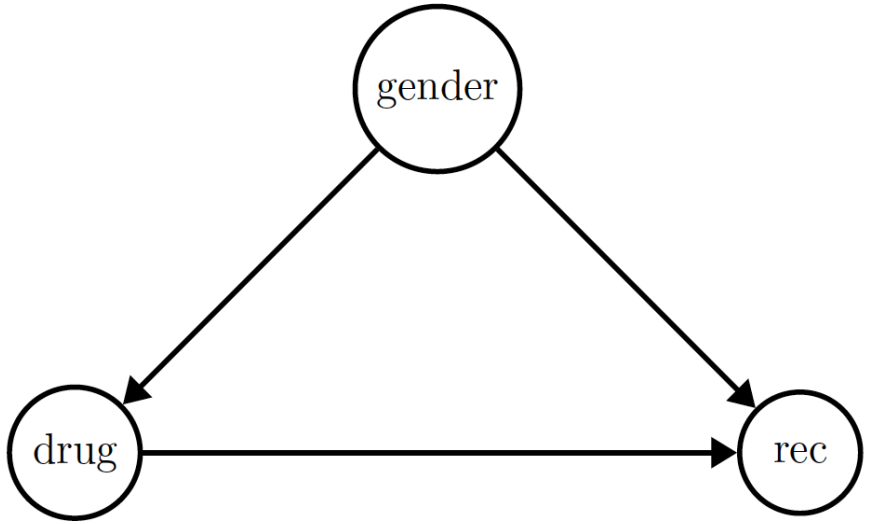
	Drug	No drug
Low BP	81 out of 87 - 93%	234 out of 270 - 87%
High BP	192 out of 263 - 73%	55 out of 80 - 69%
Combined	273 out of 350 - 78%	289 out of 350 - 83%

Total and percentage of recovered.



	Drug	No drug
Men	81 out of 87 - 93%	234 out of 270 - 87%
Women	192 out of 263 - 73%	55 out of 80 - 69%
Combined	273 out of 350 - 78%	289 out of 350 - 83%

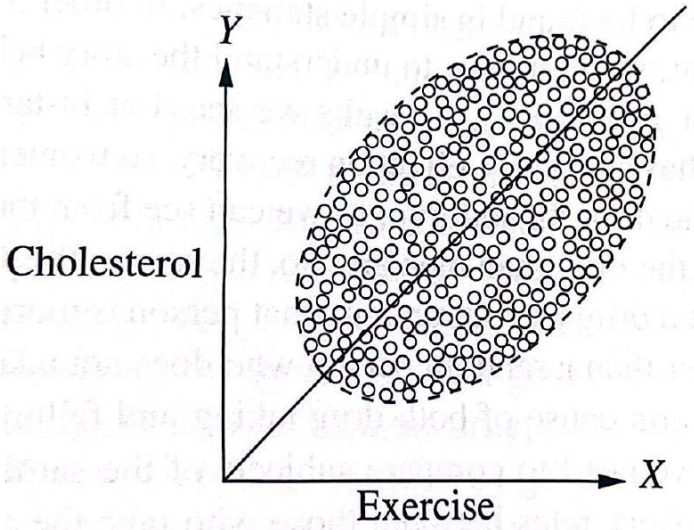
Total and percentage of recovered.

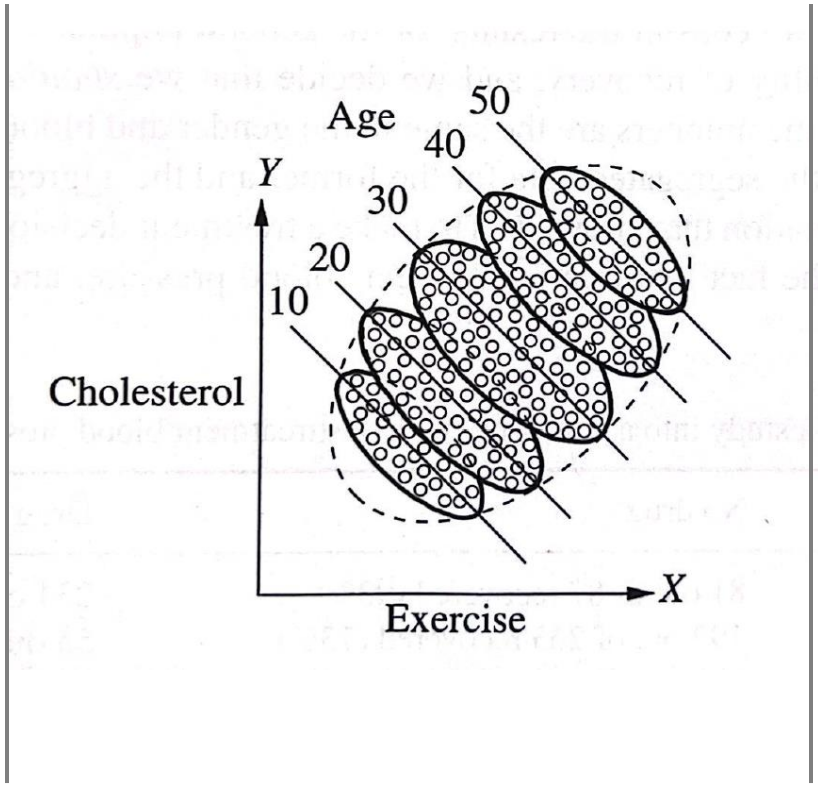


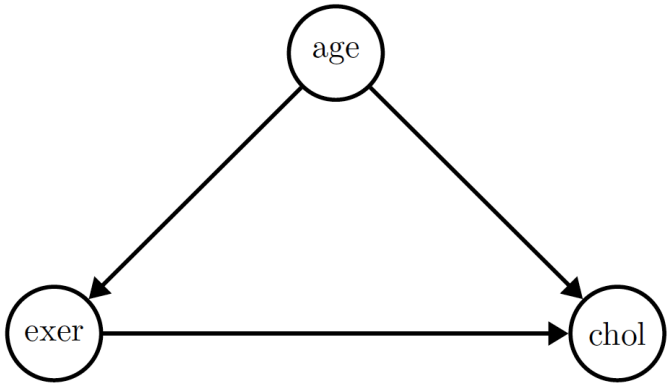
Remarkably, though the numbers are the same in the blood pressure and gender examples, the correct result lies in the aggregated data for the former (BP) and the segregated data for the latter (gender).

Insper

Back to Example 8



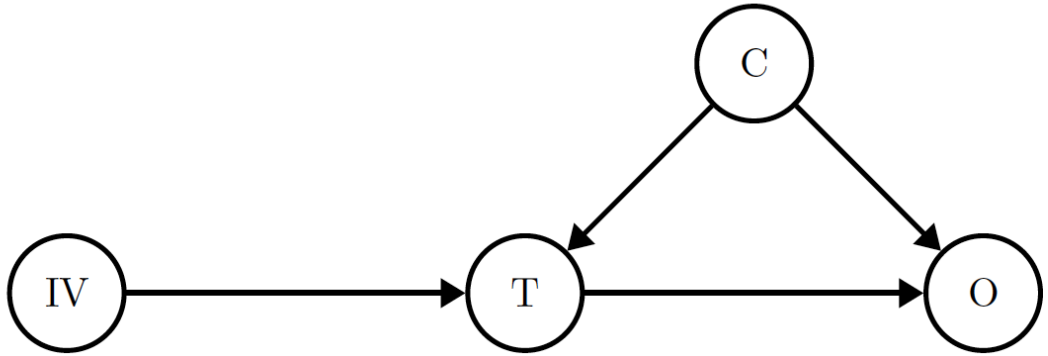




Insper

Instrumental variables

The following DAG illustrate the well-known instrumental variable solution to the endogeneity in the treatment/outcome scenario.



IV: Instrumental variable/Instrument
T: Treatment/Program/Policy/Risk Factors
C: Unmeasured Confounders/Measured Confounders/Confounding Factors
O: Outcome

Ordinary least squares (OLS)
O = a + **b***T + error ➔ **b** not causal!

Two-stage least squarse (2SLS)
O = c + **d***Tfit + error ➔ **d** causal
T = e + f*IV + error

	Confounders (C)	Treatment (T)	Outcome (O)	IV
i	Maternal characteristics	Smoking during pregnancy	Low birth weight	Cigarettes taxes
ii	Smoking, caffeine, alcohol	body mass index	Parkinson disease	FTO gene variant
iii	Prognostic factors	Catheter use	Mortality	Patients with catheter at facility
iv	Ability	Education	Earnings	Mother's education
v	Proximity	Tutoring program	GPA	Library hours

Insper

Example 10: Education on wage

College Proximity as IV: 3010 observations and 31 variables

Card (1995)[1] used wage and education data for a sample of men in 1976 to estimate the return to education.

Outcome: log(wage)

Treatment: Education

IV: Dummy variable for whether someone grew up near a four-year college.

Controls: i) experience, ii) a black dummy variable, iii) dummy variables for living in an Standard Metropolitan Statistical Area (SMSA), and iv) living in the South, and a few others.

$\widehat{\beta}_{ols} = 0.075$ and $\widehat{\beta}_{iv} = 0.132$ (twice as big!) - 95% CI: $(0.069, 0.081)$.

$se(\widehat{\beta}_{ols}) = 0.003$ and $se(\widehat{\beta}_{iv}) = 0.055$ (twenty times as big!) = 95% CI: $(0.022, 0.242)$.

Larger CIs: Price paid for consistent estimator of the return to (endogenous) education.

Card (1995) Using Geographic Variation in College Proximity to Estimate the Return to Schooling.
In Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp,
Ed. Christophides, Grant and Swidinsky, 201-222. Toronto: University of Toronto Press.

Insper

Example 11: Estimating the return to education for married women

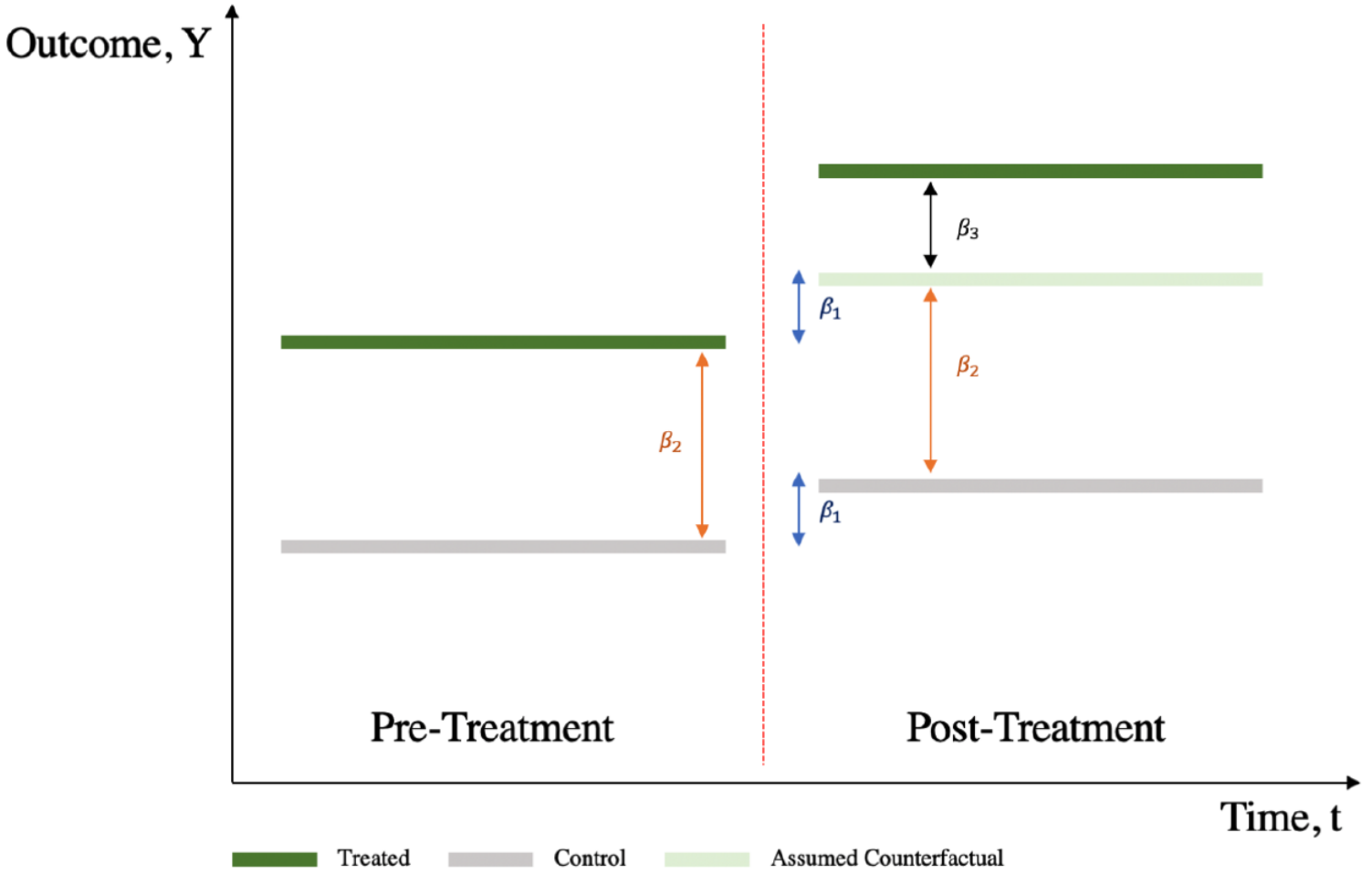
Outcome: log(wage)
Treatment: Education in years
IV: Father's education in years

OLS: 11% return for another year of education
2SLS: 5.9% return for another year of education

OLS suffers from omitted ability bias

Source:
http://hedibert.org/wp-content/uploads/2016/05/return-to-education-women.pdf
http://hedibert.org/wp-content/uploads/2016/05/return-to-education-women-R.txt

Difference-in-differences (DiD)



Insper

Difference-in-differences (DiD)

The following linear regression summarizes the above discussion:

$$y_{ti} = \beta_0 + \beta_1 P_t + \beta_2 T_i + \beta_3 P_t \times T_i + x'_{ti}\gamma + \varepsilon_{ti},$$

where $P_t = 1$ for the post-treatment period and $P_t = 0$ for the pre-treatment period, while $T_i = 1$ if individual i is in the treatment group and $T_i = 0$ if individual i is in the control group. The components of x_{ti} are additional control variables and will be cancelled out.

	Treatment Group ($T_i = 1$) (1)	Control Group ($T_i = 0$) (2)	Difference (1) – (2)
Post-Treatment Period ($P_t = 1$) (a)	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Pre-Treatment Period ($P_t = 0$) (b)	$\beta_0 + \beta_2$	β_0	β_2
Difference (a) – (b)	$\beta_1 + \beta_3$	β_1	$\boldsymbol{\beta_3}$

Insper

Example 12:
Increase in the state minimum wage on the employment

Treatment(=1): On April 1, 1992, New Jersey raised the state minimum wage from $4.25 to $5.05

Control(=0): Pennsylvania's minimum wage stays the same.

Data: Employment in fast food restaurants collected in **02/92 (Time=0)** and in **11/92 (Time=2)** - (384 restaurants).

Source: Card and Krueger (1994) Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, The American Economic Review, 84(4), 772-793.
https://davidcard.berkeley.edu/papers/njmin-aer.pdf

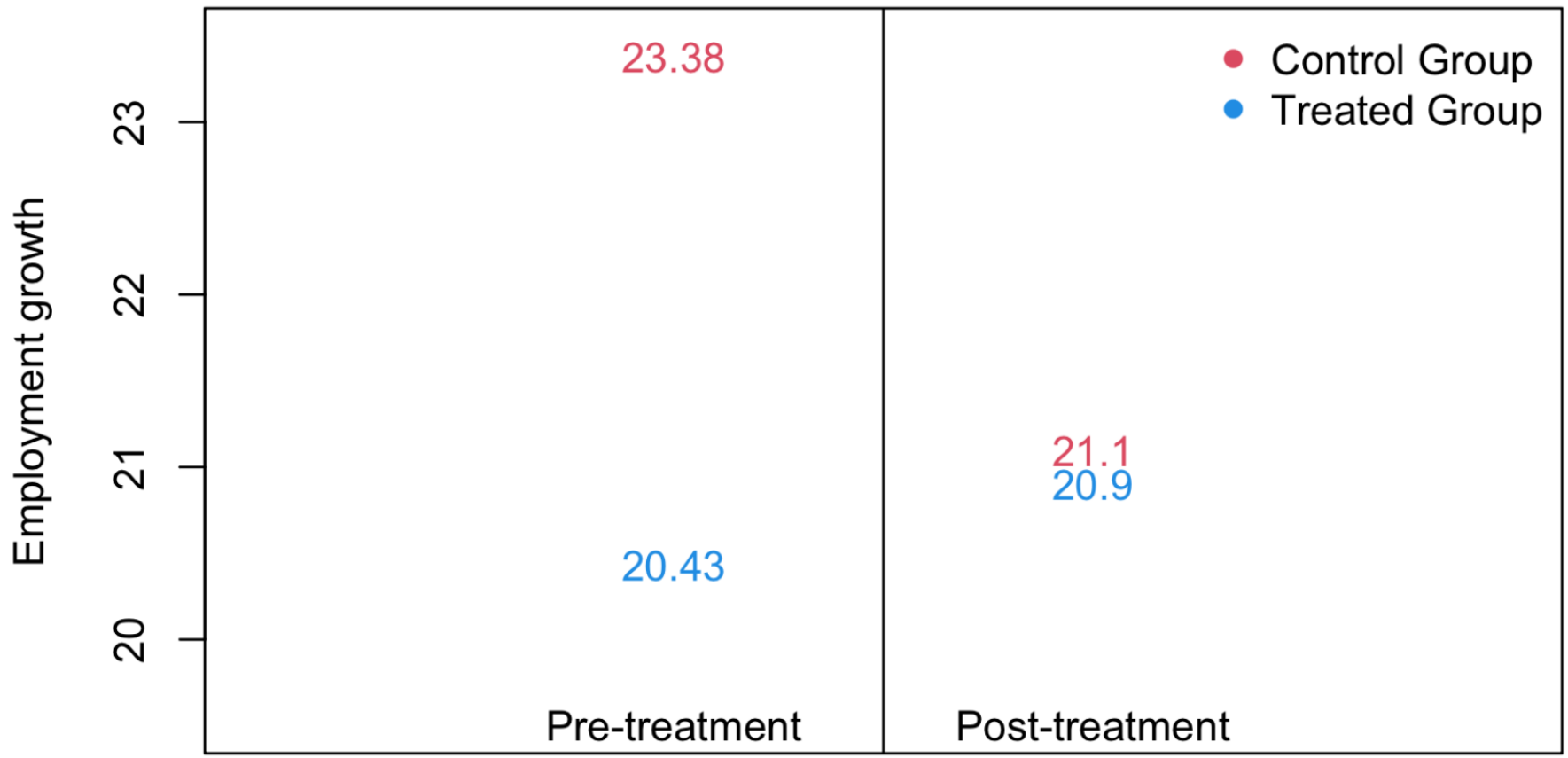
```
data = read.table("https://hedibert.org/wp-content/uploads/2024/10/card-krueger.txt",header=TRUE)
attach(data)

pretreatment.untreated  = mean(outcome[time==0 & treatment==0])
pretreatment.treated    = mean(outcome[time==0 & treatment==1])
posttreatment.untreated = mean(outcome[time==1 & treatment==0])
posttreatment.treated   = mean(outcome[time==1 & treatment==1])
A = posttreatment.treated   - pretreatment.treated
B = posttreatment.untreated - pretreatment.untreated
effect = A-B
c(A,B,effect)
#[1] 0.467 -2.283 2.750
```

`DiDreg = lm(outcome ~ time + treatment + time*treatment)`

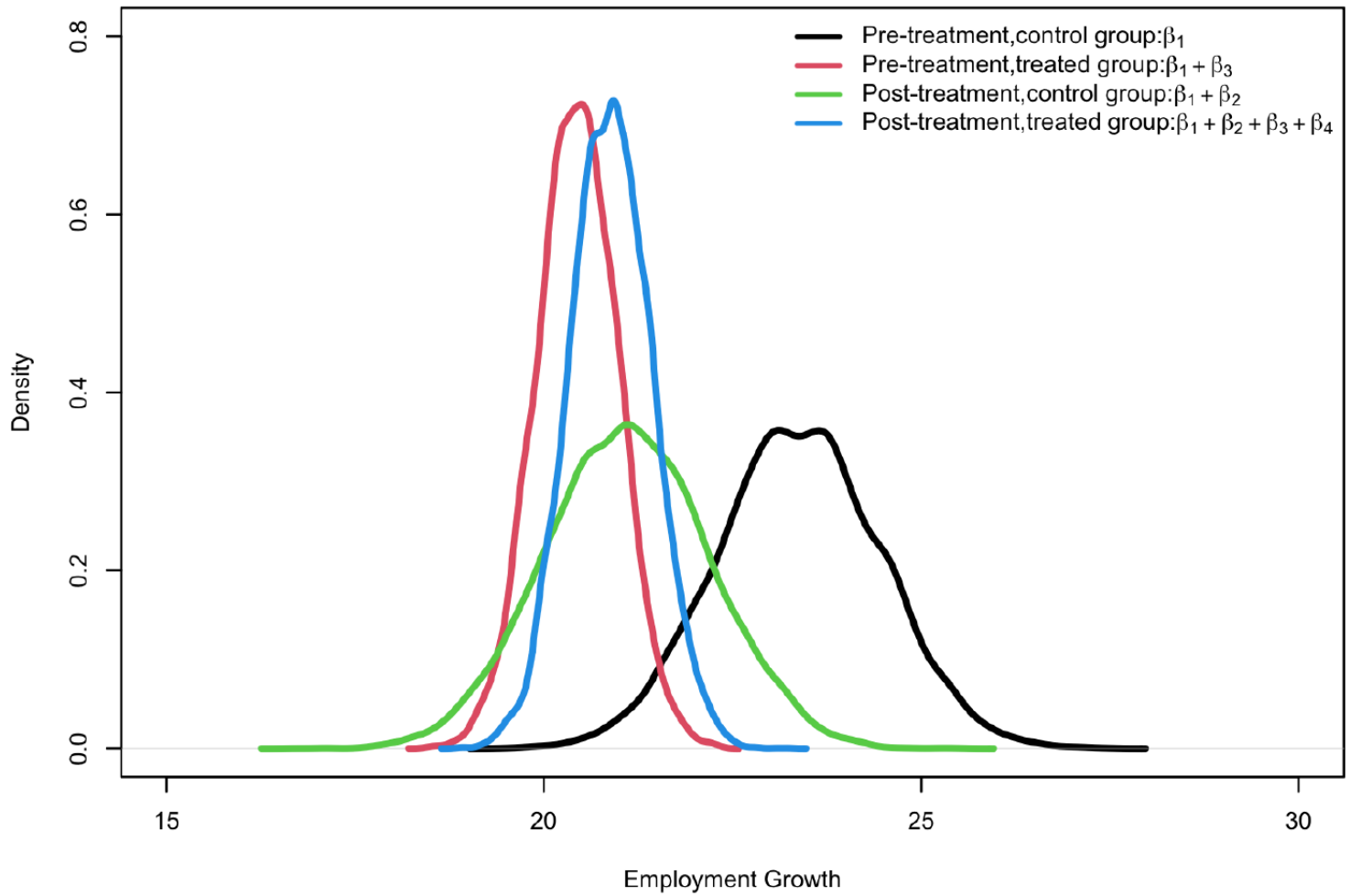
```
summary(DiDreg)
#lm(formula = outcome ~ time + treatment + time*treatment)
#
#Residuals:
# Min 1Q Median 3Q Max
#-21.097 -6.472 -0.931 4.603 64.569
#
#Coefficients:
#             Estimate Std. Error  t value  Pr(>|t|)
#(Intercept)   23.380      1.098   21.288   <2e-16 ***
#time          -2.283      1.553   -1.470   0.1419
#treatment     -2.949      1.224   -2.409   0.0162 *
#interaction    2.750      1.731    1.588   0.1126
#---
#Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
#Residual standard error: 9.511 on 764 degrees of freedom
#Multiple R-squared: 0.007587,Adjusted R-squared: 0.00369
#F-statistic: 1.947 on 3 and 764 DF, p-value: 0.1206
```

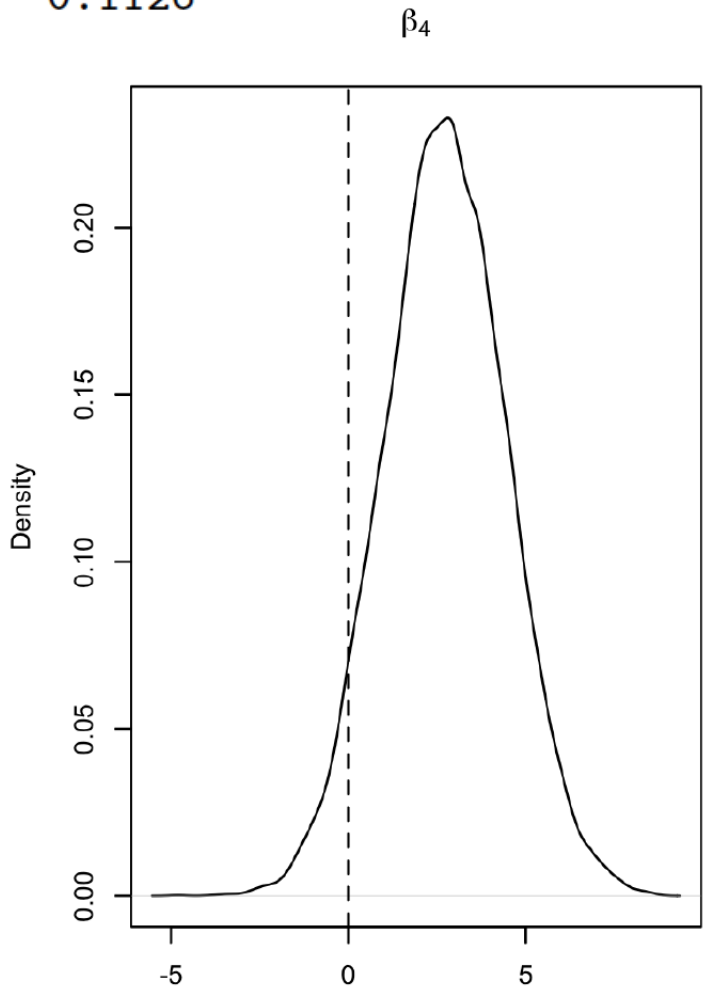
Insper

```
#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)   23.380      1.098  21.288   <2e-16 ***
#time          -2.283      1.553  -1.470   0.1419
#treatment     -2.949      1.224  -2.409   0.0162 *
#interaction    2.750      1.731   1.588   0.1126
```



```
#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)   23.380      1.098  21.288   <2e-16 ***
#time          -2.283      1.553  -1.470   0.1419
#treatment     -2.949      1.224  -2.409   0.0162 *
#interaction    2.750      1.731   1.588   0.1126
```





Insper

Regression Discontinuity Design

Source: Chapter 4, pages 147-164, of Angrist and Pischke (2015) *Mastering 'Metrics: The Path from Cause to Effect*.

Human behavior is constrained by rules.

- The State of California limits elementary school class size to 32 students; 33 is one too many.
- The Social Security Adm. won't pay you a penny in retirement benefits until you are 62.
- Armed forces recruits with test scores in the lower deciles are ineligible for military service.

For rules that constrain the role of chance in human affairs often generate interesting experiments.

Masters of 'metrics exploit these experiments with regression discontinuity (RD) design.

RD doesn't work for all causal questions, but it works for many.

And when it does, the results have almost the same causal force as those from a randomized trial.

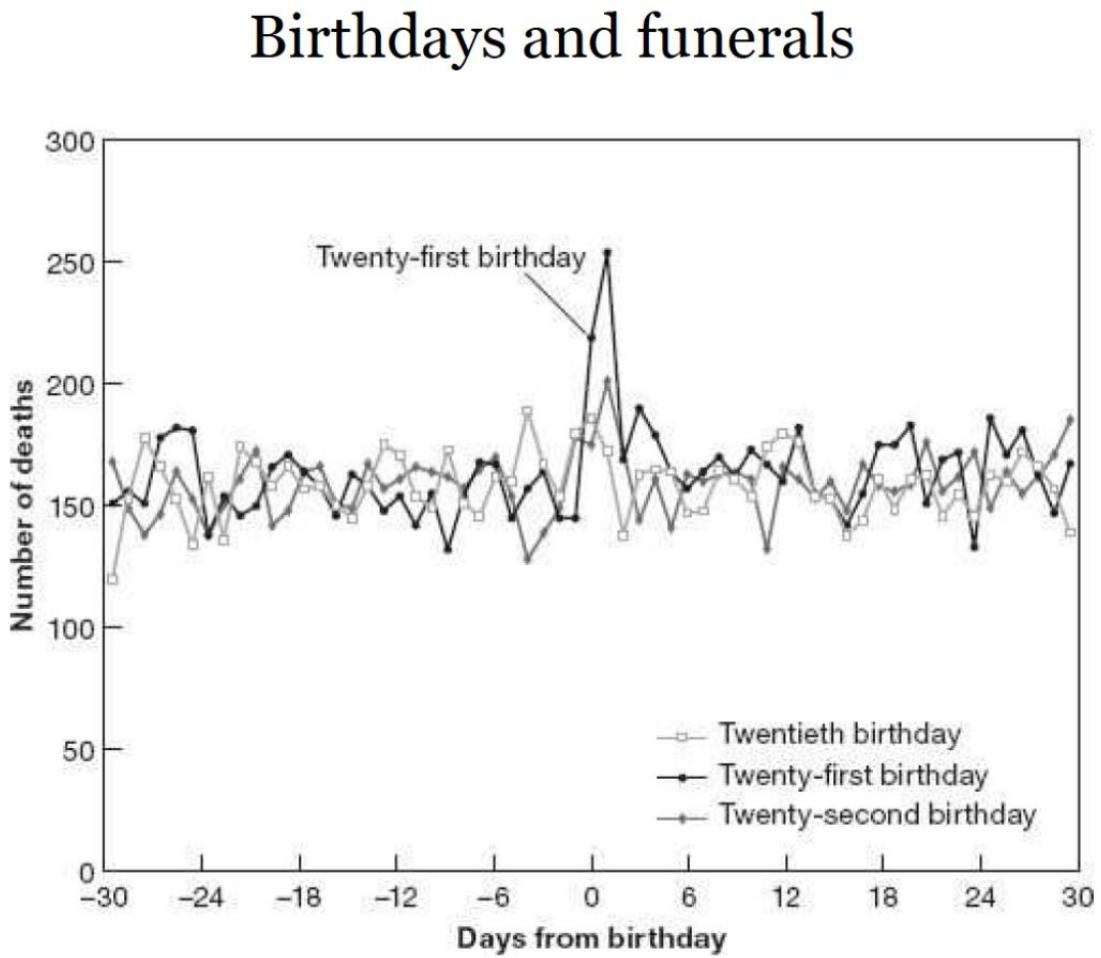
Example 13: Birthdays and Funerals

In the US the **minimum legal drinking age** (MLDA) is 21 years of age.

MLDA generates a **natural experiment** that can be used for a sober assessment of alcohol policy.

MLDA emerges from the fact that a small change in age (measured in months or even days) generates a big change in legal access.

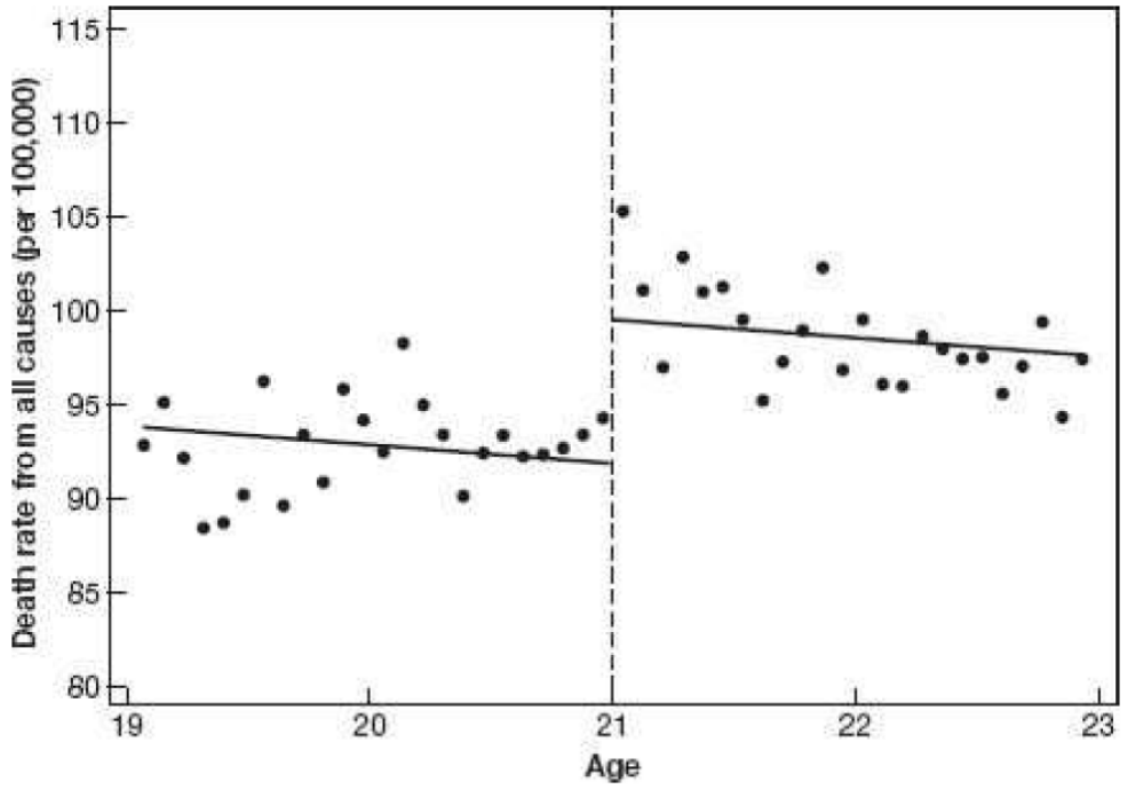
The difference a day makes can be seen in the figure below, which plots the relationship between birthdays and funerals, i.e. number of deaths among Americans aged 20-22 between 1997 and 2003.



Mortality risk shoots up on and immediately following a twenty-first birthday.
This spike adds about 100 deaths to a baseline level of about 150 per day.
There's something special about the twenty-first birthday.

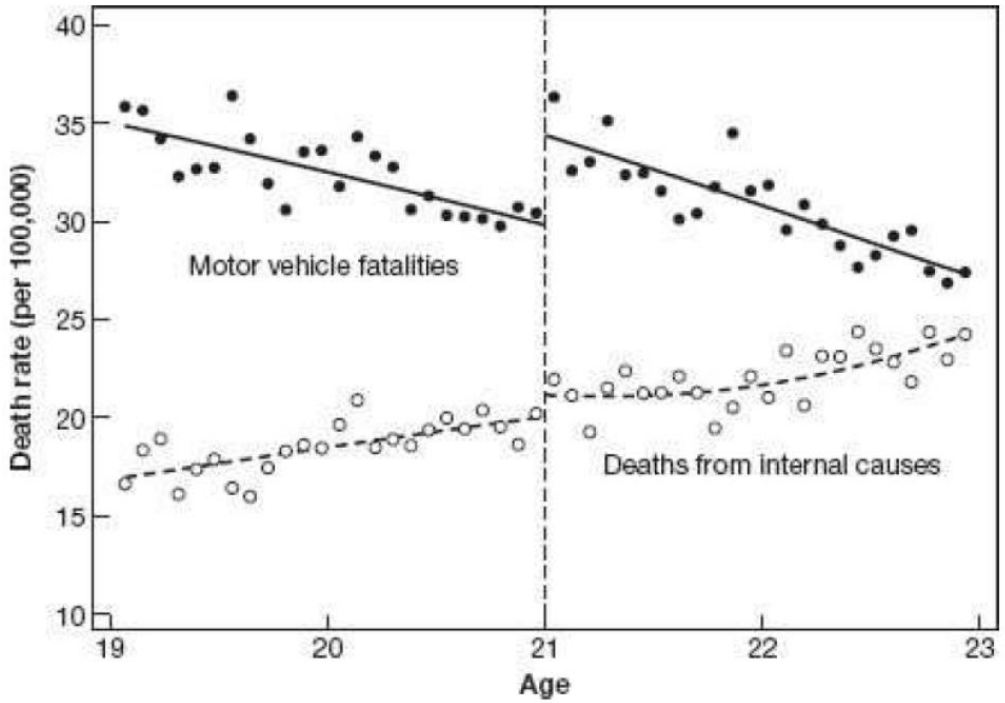
Insper

A sharp RD estimate of MLDA mortality effects



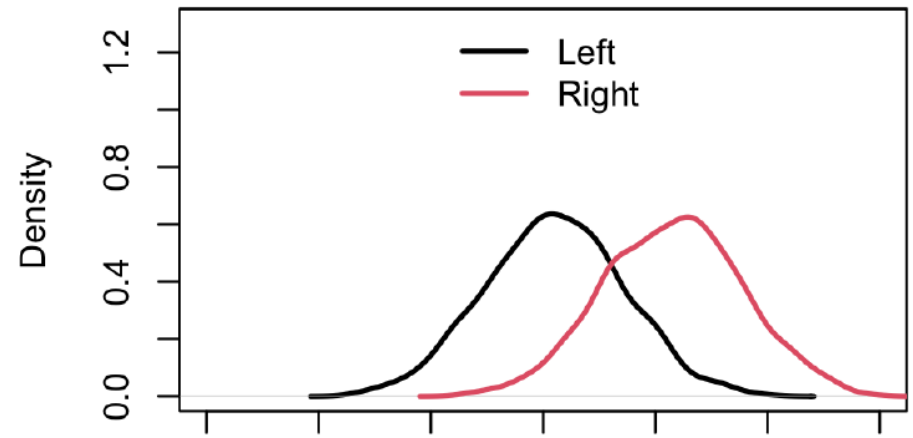
Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

RD estimates of MLDA effects on mortality by cause of death



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

Quadratic model (DIC)



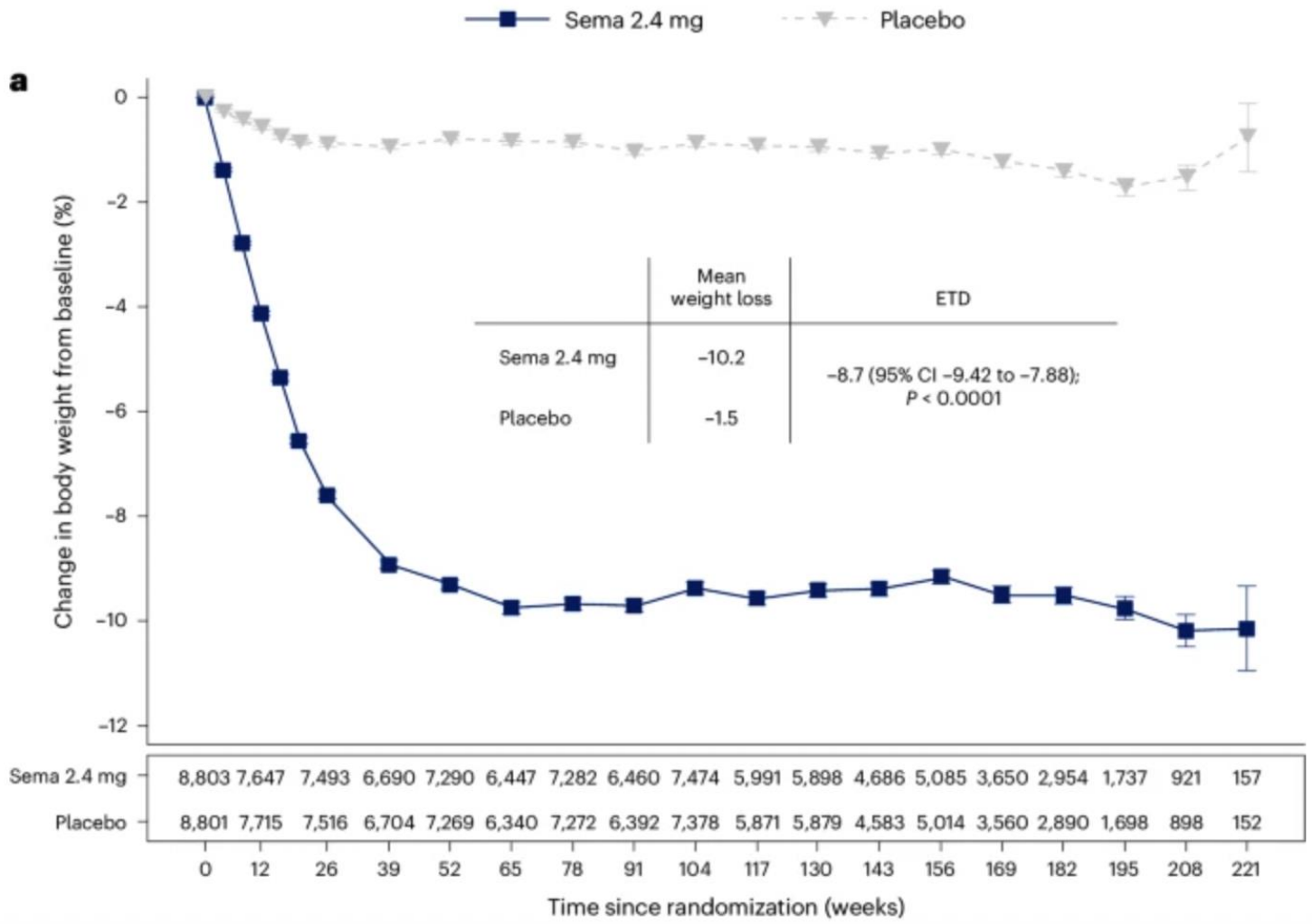
Quadratic model (MVA)

Insper

Example 14: effects of semaglutide on weight

Ryan, D.H., Lingvay, I., Deanfield, J. *et al*. **Long-term weight loss effects of semaglutide in obesity without diabetes in the SELECT trial.** *Nature Medicine* **30**, 2049–2057 (2024) - https://doi.org/10.1038/s41591-024-02996-7

Fig. 1: Percentage change in mean body weight from baseline through week 208 for all patients in-trial[21] and first on-treatment.



Insper

Example 15: Impacts of a full-time school program

Source:
Fukushima, Quintão and Pazello (2022) Impacts of a full-time school program on learning,
school's composition and infrastructure: The case of public schools in the state of São Paulo - Brazil,
Estudos Econômicos, 52(4):809-850 - DOI: 10.1590/1980-53575244ige

Resumo:

O artigo avalia o impacto do Programa de Educação Integral (PEI) implementado no Estado de São Paulo (Brasil) sobre o desempenho educacional (SAEB) e características das escolas participantes.

Usando diferenças em diferenças e lead and lags, encontramos efeitos positivos e significativos sobre o desempenho em matemática (0.469 desvio-padrão) e português (0.462 desvio-padrão) para os estudantes do 9º ano do ensino fundamental.

O impacto é maior se a escola recebe o programa há mais tempo.

O programa também reduziu a desigualdade denotas dentro das escolas.

Também identificamos que as escolas participantes apresentarammudanças em sua infraestrutura e perfil socioeconômico dos alunos.

Insper

Table 4 – Differences-in-Differences Regression Results

	Dependent variable:			
	mathematics scores		Portuguese scores	
	(1)	(2)	(3)	(4)
PEI	9.136***	7.069***	9.760***	6.979***
	(0.774)	(0.838)	(0.850)	(0.903)
Regular full-time		–0.031		–0.450
		(0.743)		(0.801)
Covariates		X		X
Observations	14.469	14.469	14.469	14.469
R^2	0.013	0.178	0.012	0.207

Note: *p<0.1;**p<0.05;***p<0.01
Note: Standard deviations in parentheses. School-level variables. ***significance at the 1% level, **significance at the 5% level, *significance at the 10% level. All regressions include school fixed effects and time fixed effects.

Impact of the PEI on mathematics (columns 1 and 2) and Portuguese proficiencies (columns 3 and 4), without and with the use of covariates.

The covariates used are related to school infrastructure, students' socioeconomic level, other students' characteristics, and dummies indicating the presence of regular full-time.

Table 5 – Leads and lags results

	Dependent variable:			
	mathematics scores		Portuguese scores	
	(1)	(2)	(3)	(4)
Lead 4	- 1.413	0.154	-2.606	- 0.705
	(1.704)	(1.561)	(1.870)	(1.682)
Lead 3	0.579	1.553	- 0.464	0.821
	(1.314)	(1.202)	(1.442)	(1.296)
Lead 2	- 2.069	- 1.106	-3.878***	- 2.707
	(1.701)	(1.558)	(1.867)	(1.680)
Lead 1	- 0.237	1.066	- 2.011	- 0.424
	(1.316)	(1.205)	(1.444)	(1.299)
Lag 0	5.090***	4.326***	4.734**	3.685**
	(1.722)	(1.654)	(1.890)	(1.783)
Lag 1	9.743***	8.651***	9.354***	7.702***
	(1.311)	(1.280)	(1.439)	(1.380)
Lag 2	14.600***	10.786***	13.184***	8.483***
	(2.458)	(2.295)	(2.697)	(2.474)
Regular full-time		0.171		-0.271
		(0.747)		(0.805)
Covariates		X		X
Observations	14.469	14.469	14.469	14.469
R^2	0.015	0.179	0.014	0.208

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

Note: Standard deviations in parentheses. School-level variables. ***significance at the 1% level, **significance at the 5% level, *significance at the 10% level. All regressions include school fixed effects and time fixed effects.

Example 16: effect of academic probation on education

Source: Rafael Alcantara, P. Richard Hahn, and Hedibert F. Lopes (2025) Learning Conditional Average Treatment Effects in Regression Discontinuity Designs using Bayesian Additive Regression Trees. http://hedibert.org/wp-content/uploads/2025/08/Alcantara-Hahn-Lopes-2025.pdf

Effect of **academic probation** in **educational outcomes** in a large Canadian university.

Students who, by the end of each term, present grade point average (GPA) lower than a certain threshold (which differs between each campus) are placed on **academic probation** and must improve their GPA in the next term.

Punishment if they fail to achieve this goal can range from one-year to permanent suspension from the university.

We focus on GPA in the term after a student is placed on probation.

Moderators

*1) **gender** ('male')*

*2) **age** upon entering university ('age_at_entry')*

*3) dummy for being **born** in North America ('bpl_north_america')*

*4) the number of **credits** taken in the first year ('totcredits_year1')*

*5) an indicator designating each of three **campuses** ('loc_campus' 1, 2 and 3)*

*6) **high school GPA** as a quantile w.r.t the university's incoming class ('hsgrade_pct')*

Insper

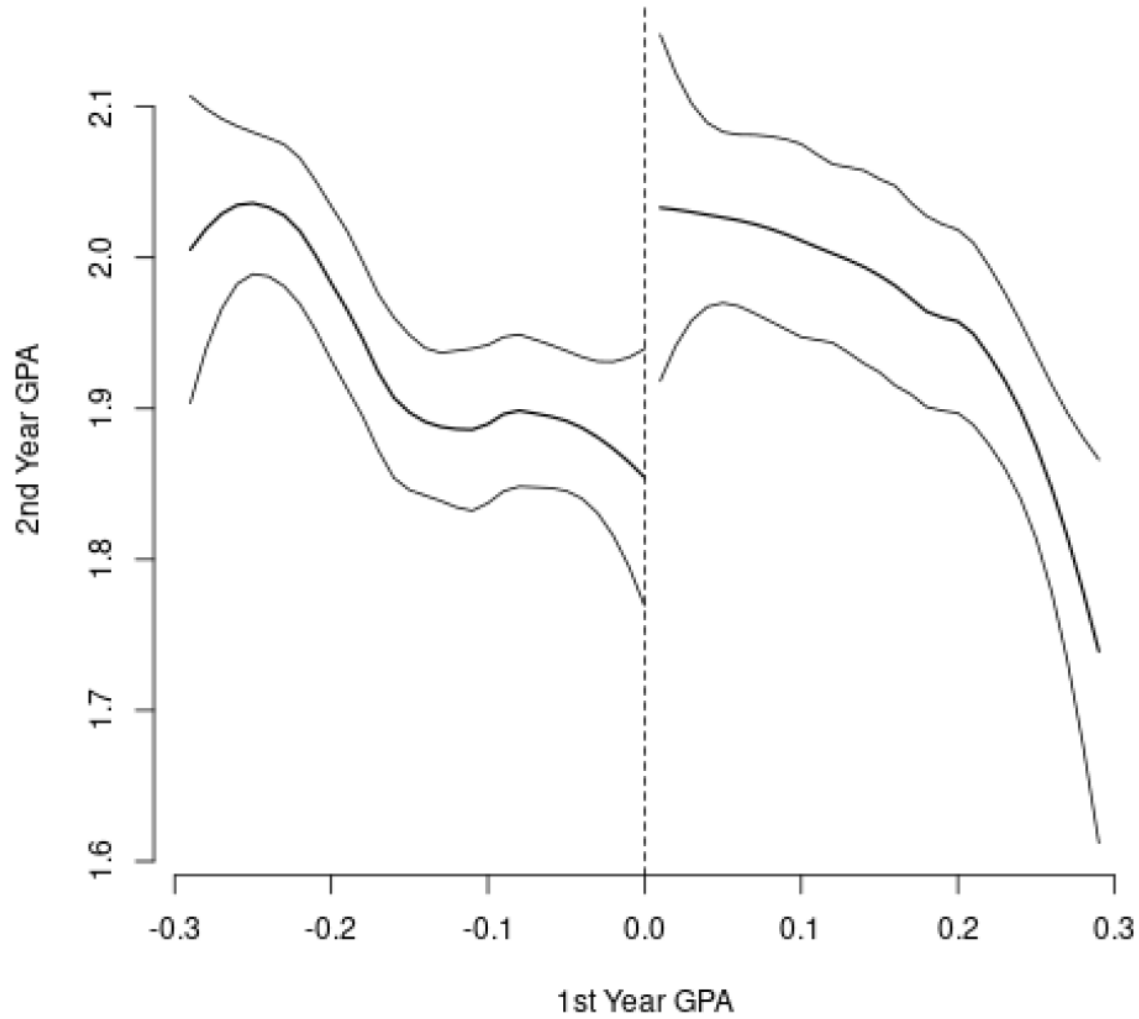


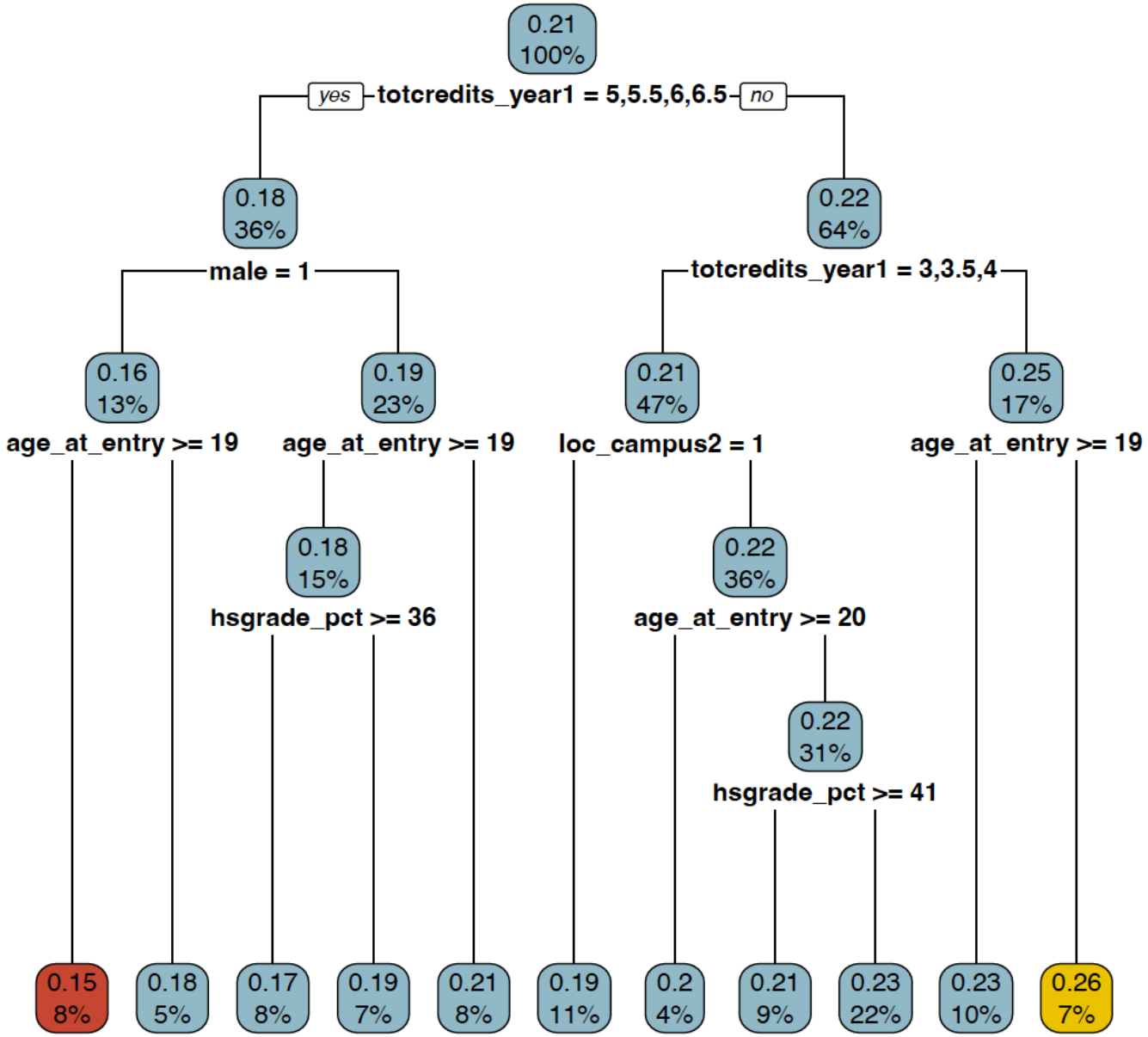
Figure 1: Effect of 1st year GPA cutoff on 2nd year GPA.

A summarizing tree

GROUP A

A male student that entered college older than 19 and attempted at least 5 credits in the first year.

128 individuals



GROUP B

A student of any gender who entered college younger than 19 and attempted more than 4, but less than 5 credits in the first year.

108 individuals

GROUP A

GROUP B

Insper

Example 17: Media manipulation

Original Article

Comparative Political Studies
2023, Vol. 0(0) 1–33

DOI: 10.1177/00104140231169027
journals.sagepub.com/home/cps
Sage

Media Manipulation in Young Democracies: Evidence From the 1989 Brazilian Presidential Election

Alexsandros Cavgias[1], Raphael Corbi[2], Luis Meloni[2], and Lucas M. Novaes[3]

Insper

Abstract

We investigate how dominant media networks can manipulate voters in young democracies. During the first presidential election after the democratic transition in Brazil, TV Globo, the largest and most-watched network in the country, unexpectedly manipulated the news coverage of the last debate 2 days before the decisive second round. In a video segment, Globo unfavorably depicted the left-wing candidate, Luiz Inácio Lula da Silva. Using the geographical distribution of broadcaster-specific TV signals and the timing of election events, we identify the effect of the manipulation net of the effect of the debate itself, showing that Globo's misleading reporting caused Lula to lose millions of votes. Our results showcase how the media can reshape an election in a single stroke, especially where the media is concentrated and politically inexperienced voters have few other sources of information.

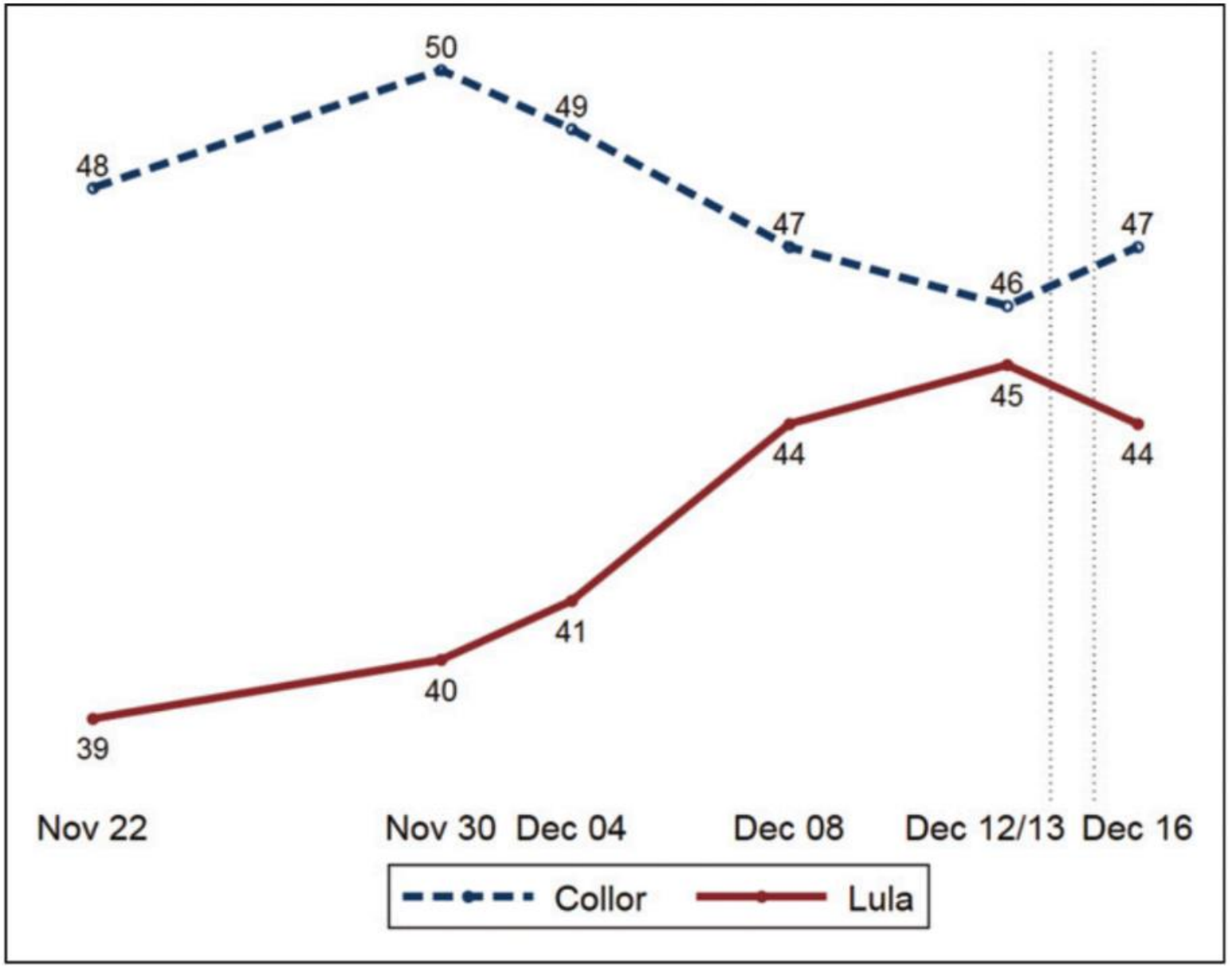


Figure 1. Voting intention trends during the second round.

Opinion polls picture how support for the two candidates evolved during the second round campaign period, suggesting that the trajectory was reversed in the last few days.

The dotted lines in Figure 1 represent the timing of the debate and Globo's coverage, registering a positive variation for Lula and a negative variation for Collor.

While the last variations are within margins of error, they also characterize a break in the previous trend.

Whether this break is the result of the debate itself or Globo's distorted portrayal of candidates' performance in the last debate.

Figure B2: Geographical Distribution of TV Signal per Broadcaster in 1989.

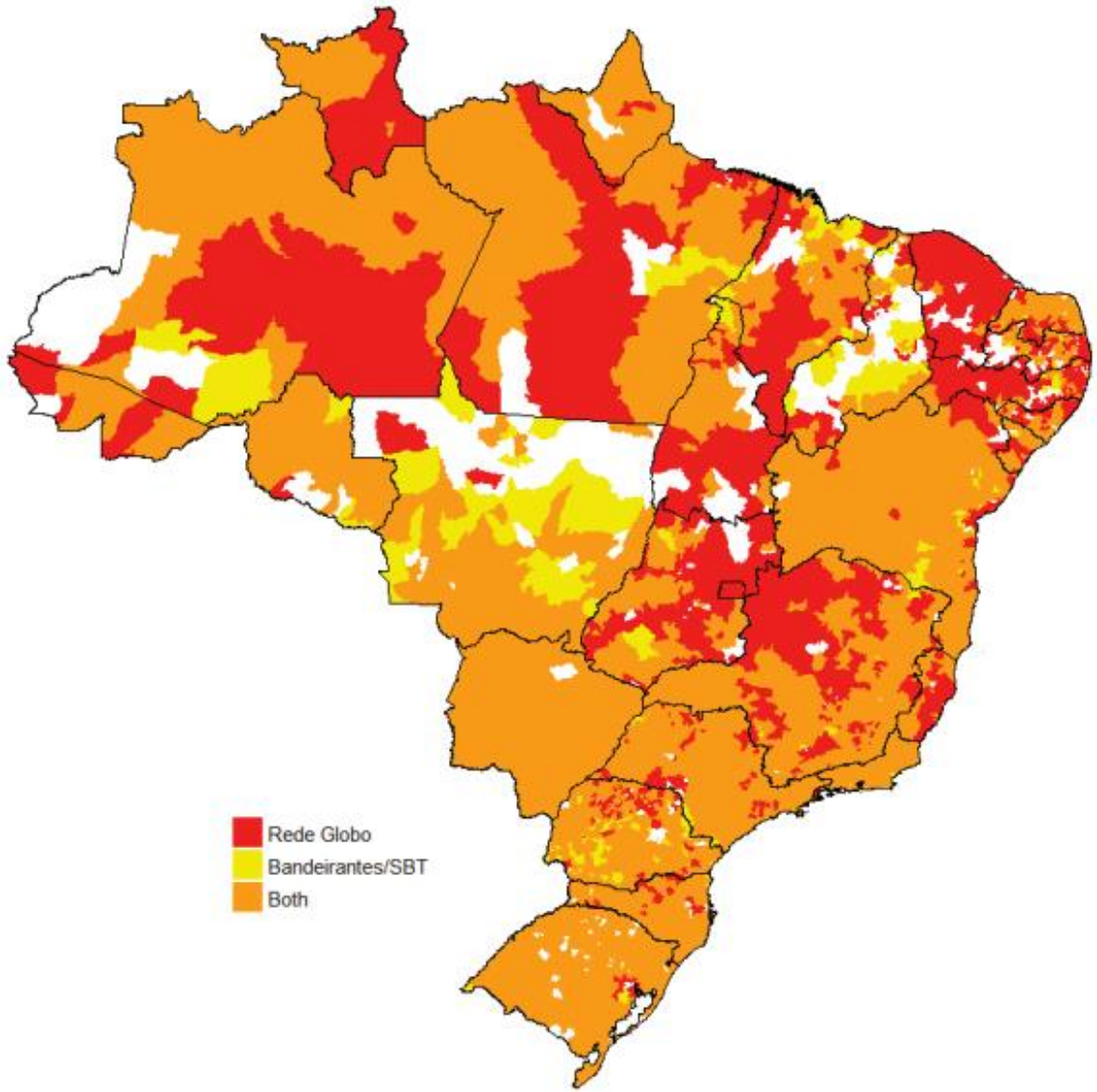


Figure B1: Describing the frequency of observations in the control and treatment groups

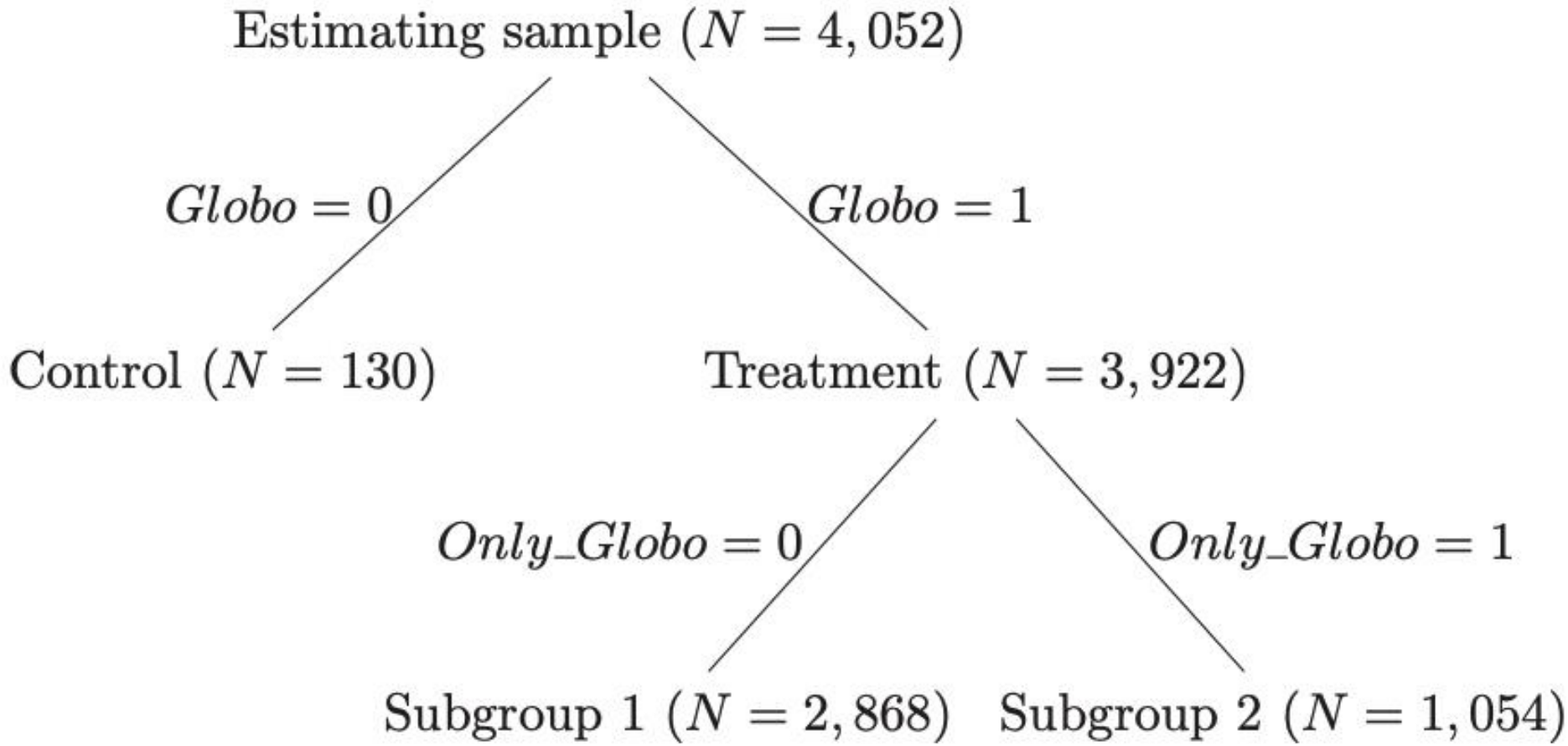


Figure 2 plots the average effect of Globo's coverage on changes in vote share of left- and right-wing candidates according to the three approaches discussed above, namely, (i) standard unconditional DiD, (ii) group FE, and (iii) p-score reweighted.

Across all three approaches, Globo's coverage is associated with a reduction of 1.2-1.8% points in the vote share of the left.

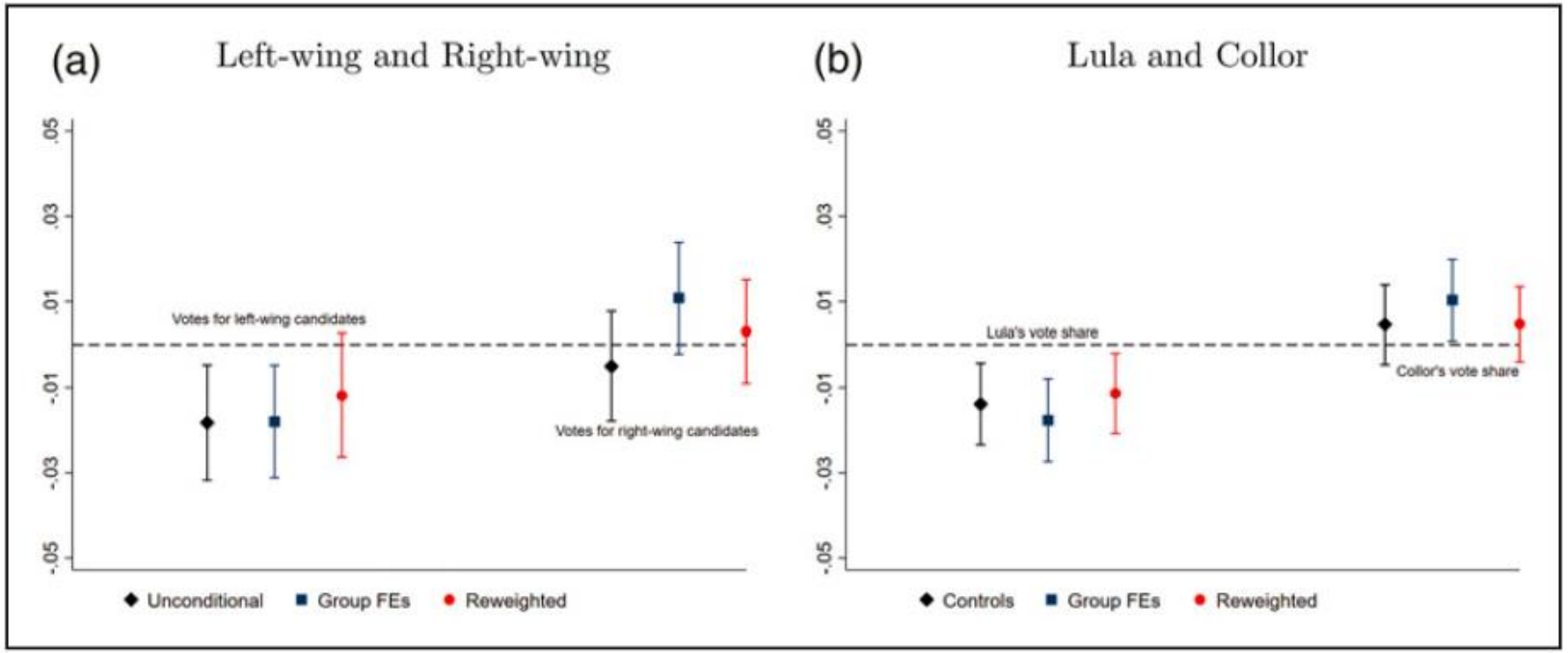


Figure 2. Globo's edited coverage and vote shares.

Insper

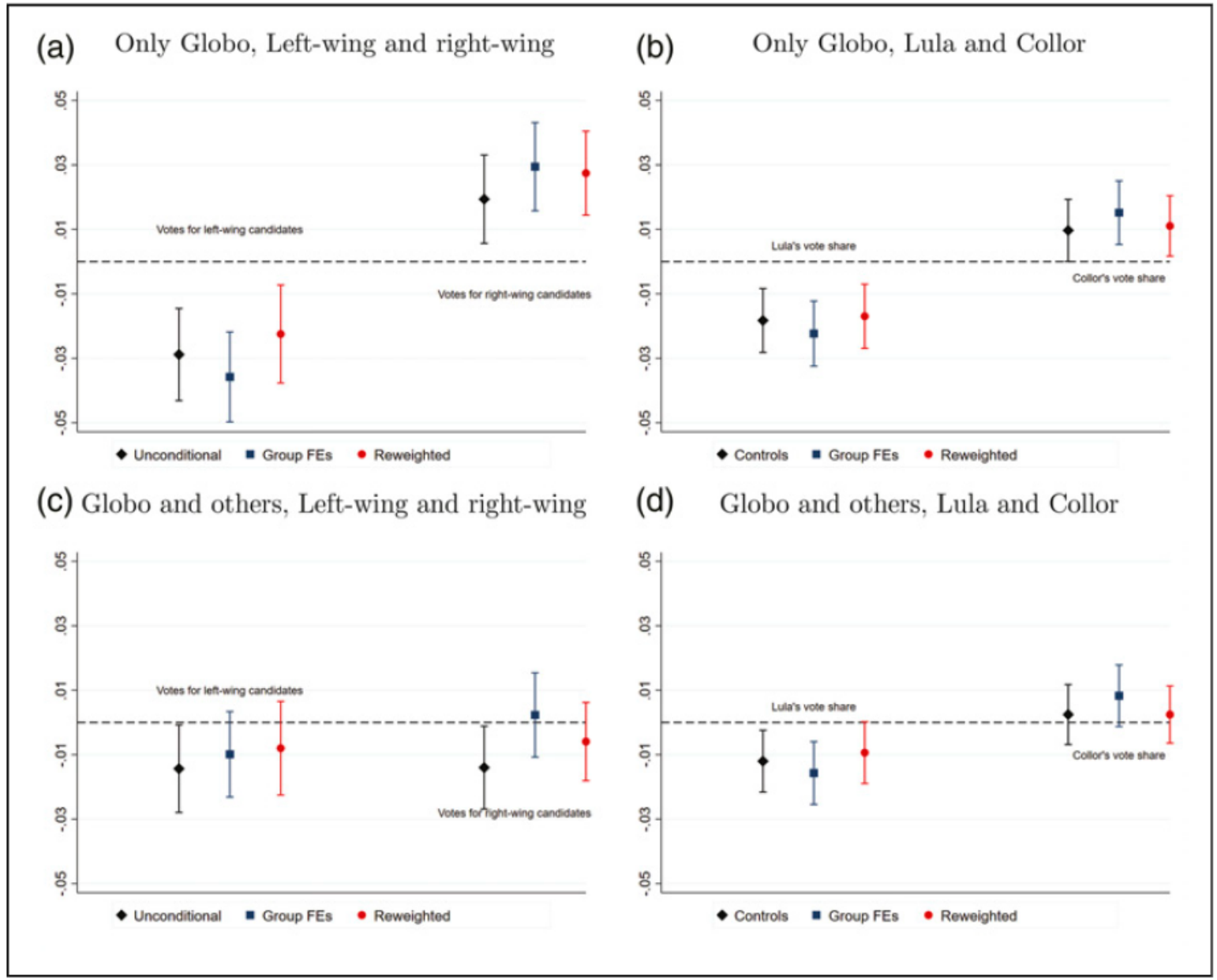


Figure 3. Globo's edited coverage and local media competition.

Conclusion

This paper shows that a powerful media group influenced an election by presenting a distorted coverage of a political debate. Although several studies identify media effects on political attitudes, ours brings new insights into the literature on media in two dimensions. First, the episode we study is not the result of exposure to partisan media in a consolidated democracy but comes from a non-openly partisan outlet during the first free presidential election after a regime change. Second, the network that edited the debate highlights was powerful enough to transmit the biased reporting to millions of voters, which under plausible assumptions, may have effectively changed the winner of a national election with a single event of manipulation.

Slides are downloadable

https://hedibert.org/wp-content/uploads/2025/09/statistical-m statistical-causality-byexamples.pdf

Insper