Data Science for Public Policy

Applied Micro Methods I

Dr. Sergio Galletta

ETHZ Zurich

27/02/2025

Empirical methods

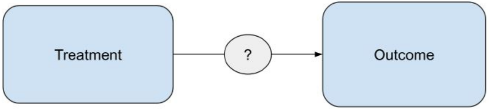
- While social scientists are often motivated by **why** questions, in research, we proceed to address **what if** questions

- This is because we are typically interested in estimating a **causal effect** (if any) of a "**treatment**" on an "**outcome**", **but is not easy!!!**

Empirical methods

- While social scientists are often motivated by **why** questions, in research, we proceed to address **what if** questions
- This is because we are typically interested in estimating a **causal effect** (if any) of a "**treatment**" on an "**outcome**", **but is not easy!!!**
- Examples:
  - How does taking this course affect the grade in your master thesis?
    - This is different from the predictive question: "What is the grade that students taking this course will obtain with their master thesis?"
  - If Zurich imposed a special tax on Uber drivers, how would that affect the supply of Uber rides?

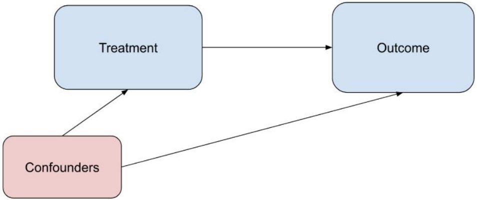
Empirical methods



Treatment

?

Outcome

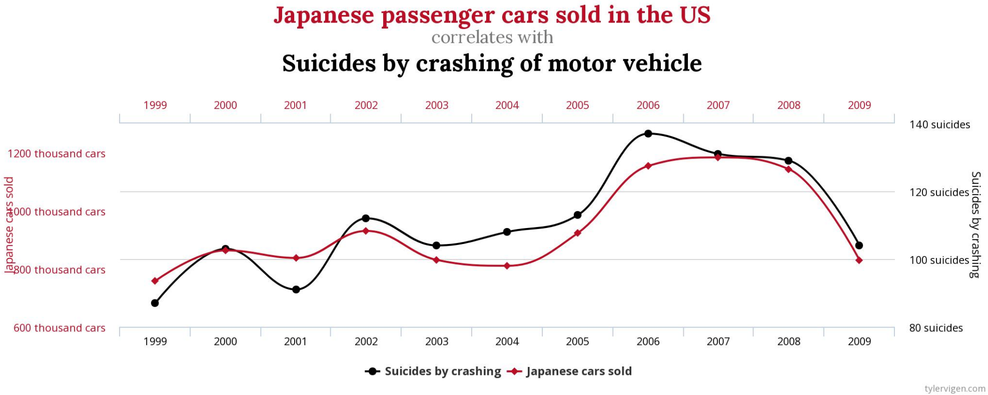


Treatment

Outcome

Confounders

Correlation does not imply causation



A formal framework (Neyman-Rubin Causal model)

- In the **potential outcomes** framework, a causal effect is defined as a comparison between **two states of the world**. For example:
  - In the first state of the world, a man takes aspirin for his headache and later reports the severity of his headache.
  - In the second state of the world, that **same** man takes nothing for his headache and later reports the severity of his headache.
- The **causal effect** of the aspirin – *a treatment*– is the **difference** in the severity of his headache – *an outcome* – **between two states of the world**

A formal framework (Neyman-Rubin Causal model)

The outcome of interest is denoted by $Y_i(D_i)$, where the notation indicates that it may depends on D_i

- $Y_i(1) = \text{if } D_i = 1$
- $Y_i(0) = \text{if } D_i = 0$

The outcome for each individual i can be written as:

- $Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$

For every individual i, the event $\{D_i = 1 \text{ instead of } D_i = 0\}$ causes the effect

- $\Delta_i = Y_i(1) - Y_i(0)$

The fundamental problem of causal inference

The "Fundamental Problem of Causal Inference"

- It is impossible to observe for the same individual i the values $D_i = 1$ and $D_i = 0$ as well as the values $Y_i(1)$ and $Y_i(0)$ and, therefore, it is impossible to observe the effect of D on Y for unit i (Holland, 1986)
- Another way to express this problem is to say that **we cannot infer the effect of a treatment because we do not have the counterfactual evidence**

The fundamental problem of causal inference

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
1	1	0	0	?	?
2	0	1	?	1	?
3	1	0	0	?	?
4	1	1	1	?	?
5	0	0	?	0	?
6	1	0	0	?	?
7	0	1	?	1	?

Causal estimands

- We could approach the problem by focusing on the **Average Treatment Effect** (ATE) for the entire population

$$E\{\Delta_i\} = E\{Y_i(1) - Y_i(0)\} = E\{Y_i(1)\} - E\{Y_i(0)\}$$

- Alternatively, one could focus on the **Average Treatment Effect on the Treated** (ATT):

$$E\{\Delta_i|D_i = 1\} = E\{Y_i(1) - Y_i(0)|D_i = 1\} = E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\}$$

Statistical solution

- **ATE, and ATT are not identified** as we cannot compute the expectations on the right-hand side (**because of the missing data**)
- Comparing the outcome based on the observed treatment status would result in a biased estimate of the ATT.

Statistical solution

- **ATE, and ATT are not identified** as we cannot compute the expectations on the right-hand side (**because of the missing data**)
- Comparing the outcome based on the observed treatment status would result in a biased estimate of the ATT.

$$E\{Y_i|D_i = 1\} - E\{Y_i|D_i = 0\}$$

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

$$= \underbrace{E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 1\}}_{\text{Treatment Effect on Treated}} + \underbrace{E\{Y_i(0)|D_i = 1\} - E\{Y_i(0)|D_i = 0\}}_{\text{"Selection Bias"}}$$

Statistical solution

- The observed difference in treatment status adds to this causal effect a term called **selection bias**
- This selection bias term is the difference in average $Y_i(0)$ between those who were and those who were not treated
- Alternatively, selection bias is the phenomenon that the distribution of the observed group is not representative to the group we are interested in

Randomized experiments

- Randomization solves this problem!

Randomized experiments

- Randomization solves this problem!
- Random assignment makes D_i independent of potential outcomes: $(Y_i(1), Y_i(0)) \perp D_i$
- Very often we deal with (ignorability assumption) $(Y_i(1), Y_i(0)) \perp D_i | X_i$

Randomized experiments

- Randomization solves this problem!
- Random assignment makes D_i independent of potential outcomes: $(Y_i(1), Y_i(0)) \perp D_i$
- Very often we deal with (ignorability assumption) $(Y_i(1), Y_i(0)) \perp D_i | X_i$
- Consider two random samples C and T, from the population - by construction, these samples are statistically identical to the entire population therefore:

$$E\{Y_i(0)|i \in C\} = E\{Y_i(0)|i \in T\} = E\{Y_i(0)\}$$

and

$$E\{Y_i(1)|i \in C\} = E\{Y_i(1)|i \in T\} = E\{Y_i(1)\}$$

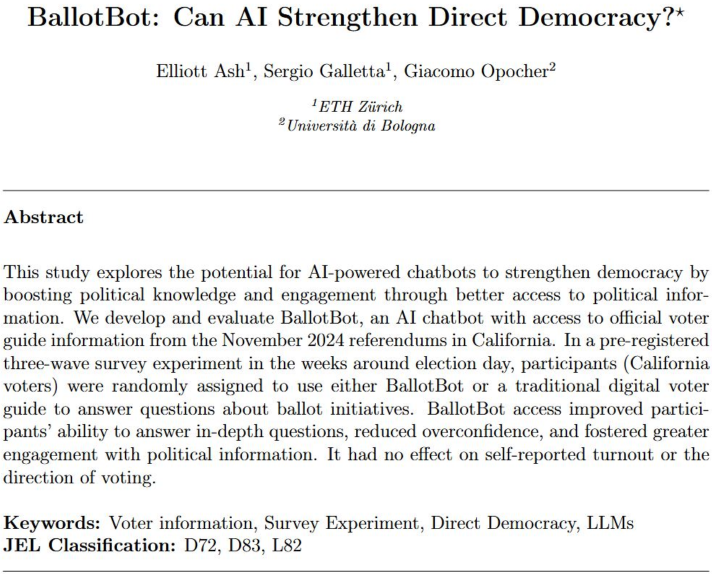
Randomized experiments

- **Randomization allows us to use the control units C as an image of what would happen to the treated unit T in the counterfactual situation of no treatment, and vice-versa**
- Going back to the first equation

$$E\{\Delta_i\} = E\{Y_i(1) - Y_i(0)\} = E\{Y_i(1)|i \in T\} - E\{Y_i(0)|i \in C\}$$

- However, randomization is rarely a feasible solution for ethical concerns and technical implementation
- But: always useful benchmark. And increasingly feasible in some settings (also because of increasing awareness among policymakers and researchers)

Example: BallotBot



BallotBot: Can AI Strengthen Direct Democracy?*

Elliott Ash[1], Sergio Galletta[1], Giacomo Opocher[2]

[1] *ETH Zürich*
[2] *Università di Bologna*

Abstract

This study explores the potential for AI-powered chatbots to strengthen democracy by boosting political knowledge and engagement through better access to political information. We develop and evaluate BallotBot, an AI chatbot with access to official voter guide information from the November 2024 referendums in California. In a pre-registered three-wave survey experiment in the weeks around election day, participants (California voters) were randomly assigned to use either BallotBot or a traditional digital voter guide to answer questions about ballot initiatives. BallotBot access improved participants' ability to answer in-depth questions, reduced overconfidence, and fostered greater engagement with political information. It had no effect on self-reported turnout or the direction of voting.

Keywords: Voter information, Survey Experiment, Direct Democracy, LLMs
JEL Classification: D72, D83, L82

Example: BallotBot

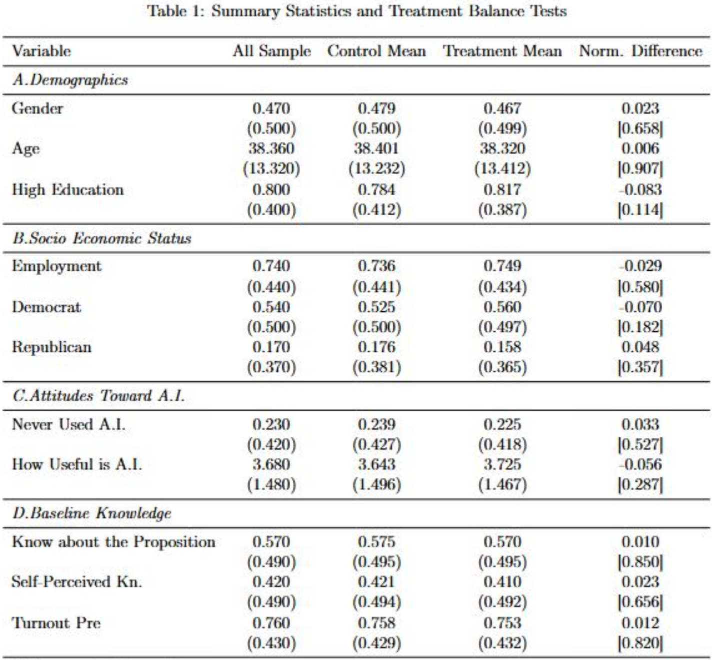


Table 1: Summary Statistics and Treatment Balance Tests

Variable	All Sample	Control Mean	Treatment Mean	Norm. Difference
A.Demographics				
Gender	0.470	0.479	0.467	0.023
	(0.500)	(0.500)	(0.499)	[0.658]
Age	38.360	38.401	38.320	0.006
	(13.320)	(13.232)	(13.412)	[0.907]
High Education	0.800	0.784	0.817	-0.083
	(0.400)	(0.412)	(0.387)	[0.114]
B.Socio Economic Status				
Employment	0.740	0.736	0.749	-0.029
	(0.440)	(0.441)	(0.434)	[0.580]
Democrat	0.540	0.525	0.560	-0.070
	(0.500)	(0.500)	(0.497)	[0.182]
Republican	0.170	0.176	0.158	0.048
	(0.370)	(0.381)	(0.365)	[0.357]
C.Attitudes Toward A.I.				
Never Used A.I.	0.230	0.239	0.225	0.033
	(0.420)	(0.427)	(0.418)	[0.527]
How Useful is A.I.	3.680	3.643	3.725	-0.056
	(1.480)	(1.496)	(1.467)	[0.287]
D.Baseline Knowledge				
Know about the Proposition	0.570	0.575	0.570	0.010
	(0.490)	(0.495)	(0.495)	[0.850]
Self-Perceived Kn.	0.420	0.421	0.410	0.023
	(0.490)	(0.494)	(0.492)	[0.656]
Turnout Pre	0.760	0.758	0.753	0.012
	(0.430)	(0.429)	(0.432)	[0.820]

Example: BallotBot

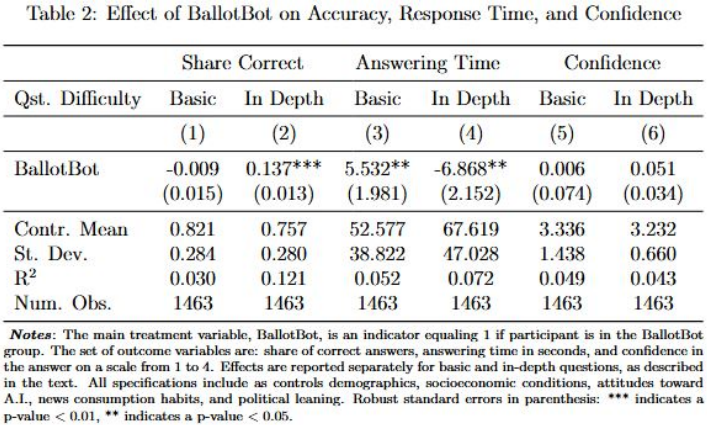


Table 2: Effect of BallotBot on Accuracy, Response Time, and Confidence

	Share Correct		Answering Time		Confidence	
Qst. Difficulty	Basic	In Depth	Basic	In Depth	Basic	In Depth
	(1)	(2)	(3)	(4)	(5)	(6)
BallotBot	-0.009	0.137***	5.532**	-6.868**	0.006	0.051
	(0.015)	(0.013)	(1.981)	(2.152)	(0.074)	(0.034)
Contr. Mean	0.821	0.757	52.577	67.619	3.336	3.232
St. Dev.	0.284	0.280	38.822	47.028	1.438	0.660
R^2	0.030	0.121	0.052	0.072	0.049	0.043
Num. Obs.	1463	1463	1463	1463	1463	1463

Notes: The main treatment variable, BallotBot, is an indicator equaling 1 if participant is in the BallotBot group. The set of outcome variables are: share of correct answers, answering time in seconds, and confidence in the answer on a scale from 1 to 4. Effects are reported separately for basic and in-depth questions, as described in the text. All specifications include as controls demographics, socioeconomic conditions, attitudes toward A.I., news consumption habits, and political leaning. Robust standard errors in parenthesis: *** indicates a p-value < 0.01, ** indicates a p-value < 0.05.

Causality without experiments

- The **research design**, **identification strategy**, or **empirical strategy** is the approach used with observational data (i.e., data not generated by a randomized trial) to approximate a randomized experiment.
- Standard methods
  - Linear Regression
  - Difference-in-differences
  - Event Studies, Synthetic Control + Synthetic DinD
  - Instrumental Variables
  - Regression Discontinuity

Introduction to Regression

- How does schooling affect income?
- Assume a linear model

$$Y_i = \alpha + s_i\beta + \epsilon_i$$

- Y_i is wage as a function of s_i, years of education
- β is the slope parameter summarizing how wages vary with schooling.
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
- ϵ_i includes all other factors affecting income besides schooling, including randomness

Ordinary Least Squares (OLS) Estimation

$$Y_i = \alpha + \beta\, s_i + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n$$

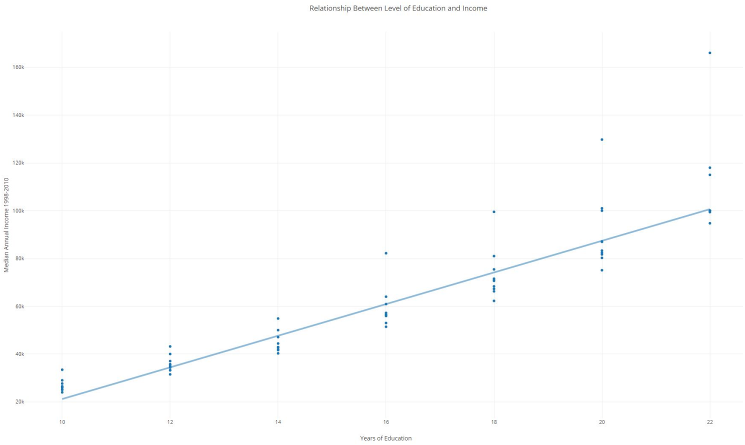
- The Ordinary Least Squares (OLS) estimator is a fundamental tool in applied microeconometrics .
- OLS chooses α and β to minimize the *sum of squared residuals (SSR)*:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \big(Y_i - \alpha - \beta s_i\big)^2.$$

- For the simple linear regression model,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(s_i - \bar{s})}{\sum_{i=1}^{n}(s_i - \bar{s})^2} = \frac{\mathrm{Cov}(Y, s)}{\mathrm{Var}(s)} \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\, \bar{s}.$$

OLS Estimator



- $\hat{\beta}$ is the slope of the regression – how Y responds to a change of 1 in s
- Is this relationship causal?
- No, we need specific assumptions

Unbiased Estimates

- The **OLS exogeneity assumption** is Cov(s, ϵ)=0
- This should mean that the treatment is uncorrelated with the error term, i.e., no confounders
- When **conditional independence** is not satisfied, we say that "s is endogenous":
  - That is, an explanatory variable s_i is said to be endogenous if it is correlated with unobserved factors (confounders) that are also correlated with the outcome variable.
- Since the error term ϵ_i includes all unobserved factors affecting the outcome, we can define endogeneity when there is correlation between an explanatory variable and the error term Cov$(s, \epsilon)\neq 0$

Omitted variable bias

- Assume individuals who choose to get more education likely differ from those who don't: maybe they have a higher innate ability, enjoy schooling, and are good at it
- The true model could be

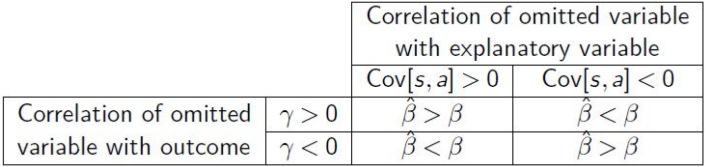
$$Y_i = \alpha + s_i\beta + \gamma a_i + \eta_i$$

- $\epsilon_i = a_i + \eta_i$ and we cannot measure a_i, while η_i is random error
- This would make $\hat{\beta}$ biased estimate of β

$$\hat{\beta} = \beta + \underbrace{\gamma\frac{\text{Cov}(s_i, a_i)}{\text{Var}(s_i)}}_{\text{Omitted Variable Bias}} + \underbrace{\frac{\text{Cov}(s_i, \eta_i)}{\text{Var}(s_i)}}_{=0 \text{ by assumption}}$$

Omitted variable bias

$$\underbrace{\gamma \frac{\text{Cov}(s_i, a_i)}{\text{Var}(s_i)}}_{\text{Omitted Variable Bias}}$$



		Correlation of omitted variable with explanatory variable	
		$\text{Cov}[s, a] > 0$	$\text{Cov}[s, a] < 0$
Correlation of omitted variable with outcome	$\gamma > 0$	$\hat{\beta} > \beta$	$\hat{\beta} < \beta$
	$\gamma < 0$	$\hat{\beta} < \beta$	$\hat{\beta} > \beta$

Understanding Statistical Significance

- The coefficient β estimates the impact of an explanatory variable on the outcome variable
- In causal analysis, it's crucial to assess whether the observed effect (as quantified by β) is statistically significant, meaning it's unlikely to have occurred by chance.
- Statistical significance is typically evaluated by calculating the *standard error* for β, which then aids in deriving *confidence intervals* and a *p-value*. These metrics test the null hypothesis that $\beta = 0$, helping to determine the reliability of the estimated effect.

Residuals and Standard Errors

- The residuals or errors from an OLS regression ($\neq$ st. errors) are defined as

$$\begin{aligned}\tilde{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} s_i\end{aligned}$$

- The standard error (SE) for the OLS estimate $\hat{\beta}$ is

$$\hat{\sigma}_\beta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_i^2},$$

  the square root of the average of the squared residuals.

- SE provides information about the estimate's precision: a lower standard error is a more precise estimate.

t-statistics, p-values and confidence intervals

- A rule of thumb for statistical significance is to compute the ***t*-statistic**:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$$

- $t > |2| \rightarrow$ statistically significant effect
- A high t (in absolute value) is associated with a small ***p*-value** (e.g., $t = \pm 1.96 \rightarrow p = .05$)
- 95% **confidence intervals** indicate (roughly) that the coefficient is 95% likely to reside within that interval

$$CI^{\beta}_{0.95} = [\hat{\beta} - 1.96 \times \hat{\sigma}_\beta, \hat{\beta} + 1.96 \times \hat{\sigma}_\beta]$$

Causal inference toolkit

- Difference-in-differences
- Event Studies, Synthetic Control + Synthetic DinD
- Instrumental Variables
- Regression Discontinuity

Difference-in-differences

- Assume we have n units (i) and T time periods (t)
- Consider a binary policy D_{it}, and we are interested in estimating its effect on outcomes Y_{it}
- The inherent problem is that D_{it} is *not* necessarily randomly assigned
- Diff-in-Diff works under the assumption that **in the absence of the treatment, the Y_{it} across units evolve in parallel – their γ_t are identical.**

$$Y_{it}(D_{it}) = \alpha_i + \gamma_t + \tau_i D_{it}$$
$$\text{s.t. } Y_{it}(1) - Y_{it}(0) = \tau_i$$

- Absent the policy, units may have different *levels* (α_i) but their changes would evolve in parallel

Difference-in-differences (using linear regression)

- A simple linear regression will identify $E(\tau_i | D_{i1} = 1)$ with two time periods:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta + \epsilon_{it} \tag{1}$$

- This setup is sometimes referred to as the Two-way Fixed Effects estimator (TWFE)

- Note: we could have also estimated τ directly:

$$\hat{\tau} = \underbrace{E(Y_{i1} - Y_{i0} | D_i = 1)}_{\Delta \overline{Y}_1} - \underbrace{E(Y_{i1} - Y_{i0} | D_i = 0)}_{\Delta \overline{Y}_0}$$

  - Intuitively, we generate a counterfactual for the treatment using the changes in the untreated units: $E(Y_{i1} - Y_{i0} | D_i = 0)$

- Necessary: two time periods! What if we have more?

Multiple time periods in basic setup

- Let's consider a policy that occurs all at t_0 (e.g. single timing rolled out to treated units)

- More time periods helps in several ways:
  1. If we have multiple periods *before* the policy implementation, we can partially test the underlying assumptions
  2. If we have multiple periods *after* the policy implementation, we can examine the timing of the effect

- How do we implement this?

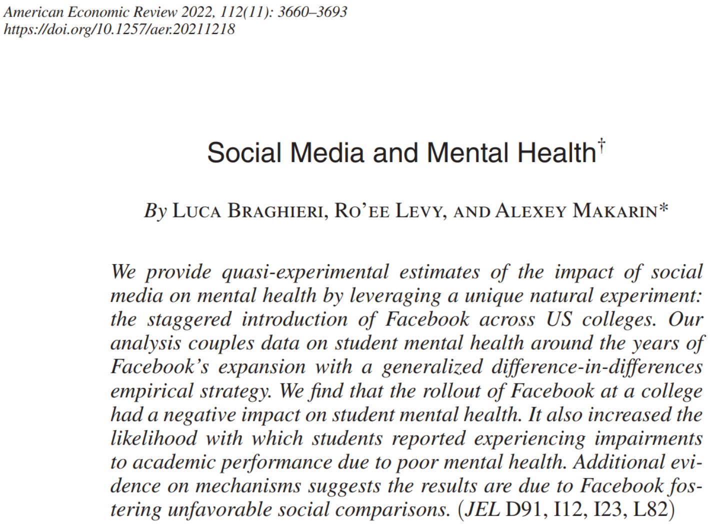
$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^{T} \delta_t D_{it} + \epsilon_{it},$$

  - One of the coefficients is fundamentally unidentified because of α_i
  - All coefficients measure the effect *relative* to period t_0.

Recent warning

- This literature has had a certain amount of upheaval over the past 5-6 years
- Tension: provide context for how people currently and historically have studied diff-in-diff
- More attention on multiple periods analysis, in checking parallel trends and inference
- Check Roth et al 2023 for a synthesis of the recent econometrics literature

Example: Facebook and Mental Health



American Economic Review 2022, 112(11): 3660–3693
https://doi.org/10.1257/aer.20211218

Social Media and Mental Health†

By Luca Braghieri, Ro'ee Levy, and Alexey Makarin*

We provide quasi-experimental estimates of the impact of social media on mental health by leveraging a unique natural experiment: the staggered introduction of Facebook across US colleges. Our analysis couples data on student mental health around the years of Facebook's expansion with a generalized difference-in-differences empirical strategy. We find that the rollout of Facebook at a college had a negative impact on student mental health. It also increased the likelihood with which students reported experiencing impairments to academic performance due to poor mental health. Additional evidence on mechanisms suggests the results are due to Facebook fostering unfavorable social comparisons. (*JEL* D91, I12, I23, L82)

Example: Facebook and Mental Health

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times \mathit{Facebook}_{gt} + \mathsf{X}_i \times \gamma + \mathsf{X}_c \times \lambda + \epsilon_{icgt}$$

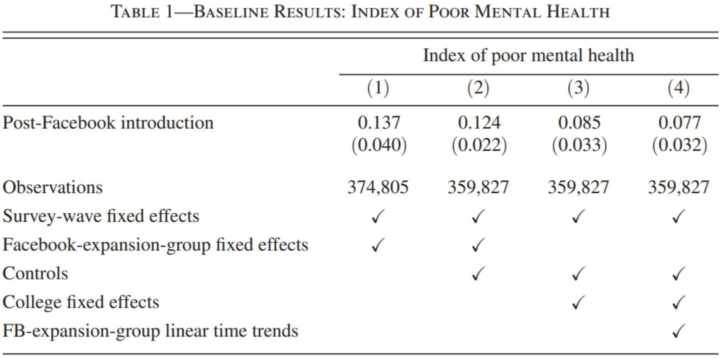


TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137 (0.040)	0.124 (0.022)	0.085 (0.033)	0.077 (0.032)
Observations	374,805	359,827	359,827	359,827
Survey-wave fixed effects	✓	✓	✓	✓
Facebook-expansion-group fixed effects	✓	✓		
Controls		✓	✓	✓
College fixed effects			✓	✓
FB-expansion-group linear time trends				✓

Example: Facebook and Mental Health

$$Y_{igt} = \alpha_g + \delta_t + \beta_k \times \sum_{k=-8}^{5} D_{k(gt)} + \epsilon_{icgt}$$

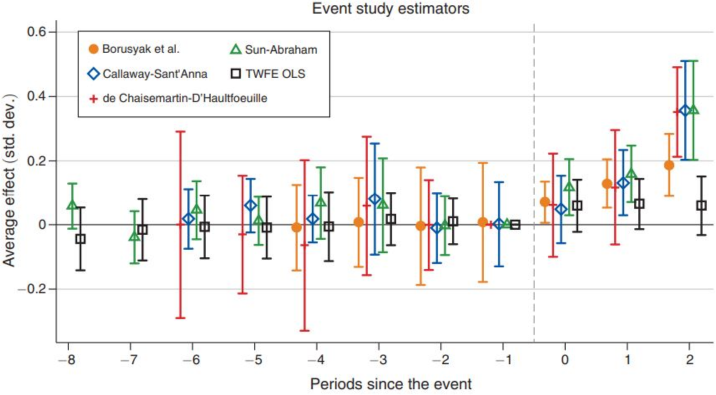
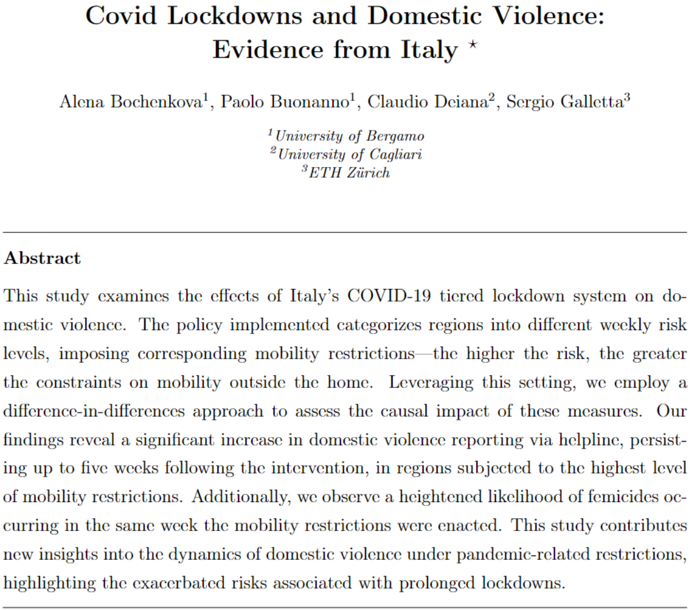


Figure 2. Effects of Facebook on the Index of Poor Mental Health Based on Distance to/from Facebook Introduction

Example: Lockdown and Domestic Violence



Covid Lockdowns and Domestic Violence: Evidence from Italy *

Alena Bochenkova[1], Paolo Buonanno[1], Claudio Deiana[2], Sergio Galletta[3]

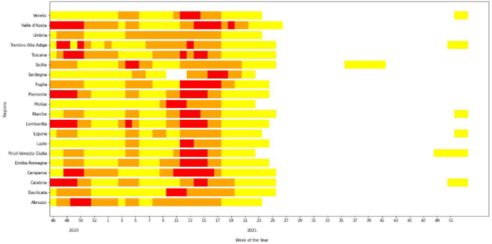
[1] *University of Bergamo*
[2] *University of Cagliari*
[3] *ETH Zürich*

Abstract

This study examines the effects of Italy's COVID-19 tiered lockdown system on domestic violence. The policy implemented categorizes regions into different weekly risk levels, imposing corresponding mobility restrictions—the higher the risk, the greater the constraints on mobility outside the home. Leveraging this setting, we employ a difference-in-differences approach to assess the causal impact of these measures. Our findings reveal a significant increase in domestic violence reporting via helpline, persisting up to five weeks following the intervention, in regions subjected to the highest level of mobility restrictions. Additionally, we observe a heightened likelihood of femicides occurring in the same week the mobility restrictions were enacted. This study contributes new insights into the dynamics of domestic violence under pandemic-related restrictions, highlighting the exacerbated risks associated with prolonged lockdowns.

Example: Lockdown and Domestic Violence

Figure 1: The four-tiered system of red, orange, yellow, and white zones



Notes: The figure is based on the color designation from the Ministry of Health's decree in the *Gazzetta Ufficiale* from November 2020 to December 2021.

Example: Lockdown and Domestic Violence

Figure 5: The effect of the orange-red zone on mobility

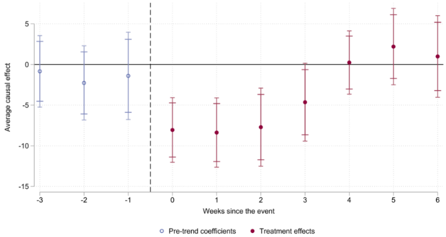


Figure 3: The effect of the orange-red zone on 1522 helpline calls

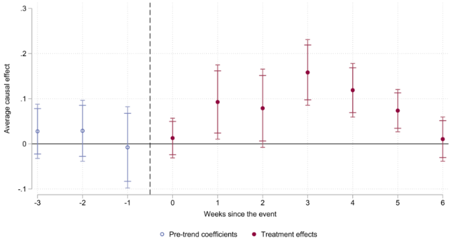
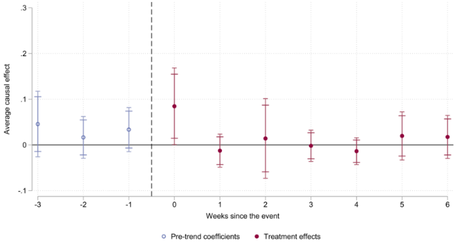


Figure 4: The effect of the orange-red zone on femicide


Average causal effect
Weeks since the event
○ Pre-trend coefficients
● Treatment effects