Causal methods with observational data

Francisco Villamil

Research Design for Social Sciences
MA Computational Social Science, UC3M
Fall 2023

Roadmap

Intro and overview

Methods in brief

Paper discussion

Re-cap and essay guidelines

Re-cap

1. Problem/topic

2. Stories, arguments about mechanisms

3. Research question

4. Proper theory, concepts and operationalization

5. Measurement, unit of analyses, data sources, etc

6. Inference strategy

7. Results & interpretation

Methods and causal inference

- Most of the time is **impossible** to control for all relevant variables (i.e. not able to close all back-door paths)

- So what do we do? We try to find ways to control for unobserved confounding

- One option is to rely on additional controlling techniques
  - → Matching : also depends on observables, but parametric advantages, etc (anyway, not a solution for U)
  - → Fixed effects : can control for group-level unobservables, and correct issues related to ecological fallacy

- But there are other methods often related to causal inference, not because they uncover causal relationships, but because they allow you to exploit 'typical' exogenous sources of variation

- (In H-K's *The Effect*, they're called 'template causal diagrams')

Methods and causal inference

- Another option is to find a source of exogenous variation and exploit

- There are a few methods related to causal inference not because they uncover causal relationships but because they are designed to exploit *'typical' exogenous sources of variation*

- Situations where there is *something* that introduces variation in the treatment that is independent from confounders and you can exploit to analyse

  $\rightarrow$ Same as randomization (of treatment assignment) in experiments

- (In H-K's *The Effect*, they're called 'template causal diagrams')
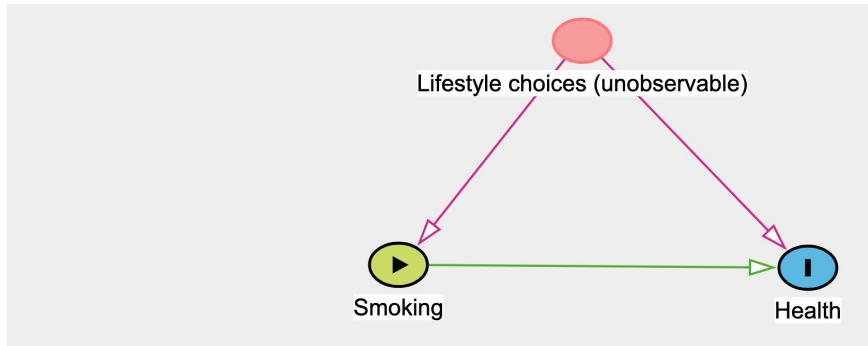
Methods and causal inference

Methods and causal inference
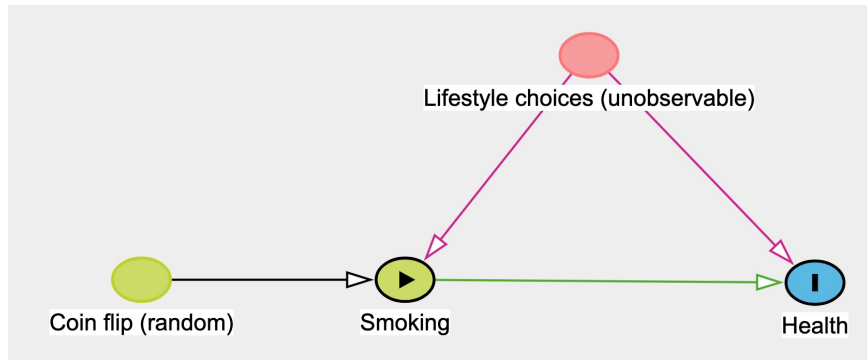
Methods and causal inference

- Five techniques commonly used in causal inference
    - $\rightarrow$ two of them use for controlling (closing back-doors), and the other three to exploit 'pre-made' causal models
    - $\rightarrow$ btw, what is controlling? when you control for Z, you remove the variation in X and Y that is *explained by Z* (see this)

Methods and causal inference

- Five techniques commonly used in causal inference
  - $\rightarrow$ two of them use for controlling (closing back-doors), and the other three to exploit 'pre-made' causal models
  - $\rightarrow$ btw, what is controlling? when you control for Z, you remove the variation in X and Y that is *explained by Z* (see this)

1. Fixed effects
2. Difference-in-differences
3. Regression discontinuity design
4. Instrumental variables
5. Matching

Roadmap

Fixed effects

- The problem with covariate adjustment is that we need to *observe* those variables, but we usually have unobserved confounders
- An strategy in that case is to try within-group comparisons, which blocks all group-wide confounding
- For example, imagine cases when our U variable is:
  $\rightarrow$ *city of origin*, in an individual-level analysis; *school effects* in an analyses of students grades; *individual background*, in a panel survey analysis...
- It corrects for ecological fallacy (esp. Simpson's paradox)
- Estimation: dummy variable for each group

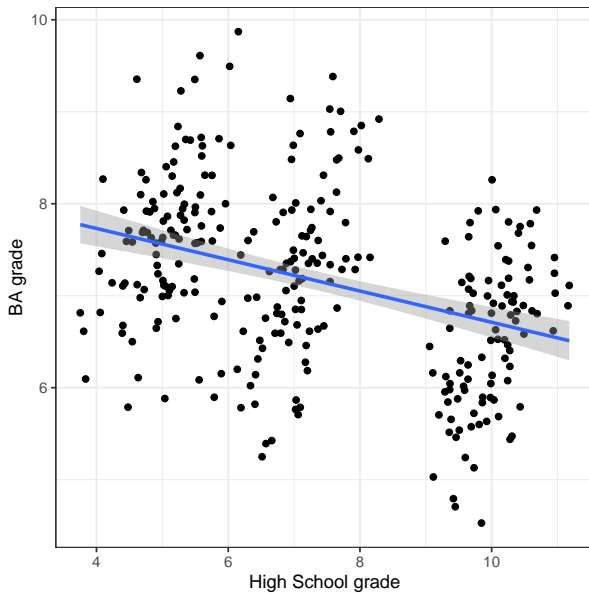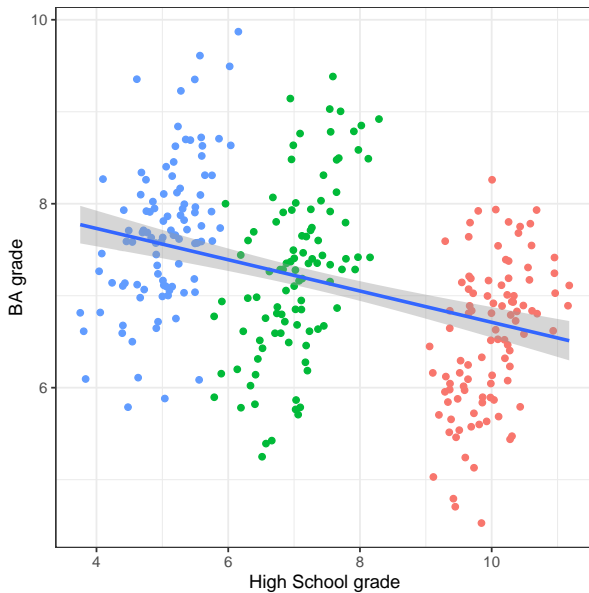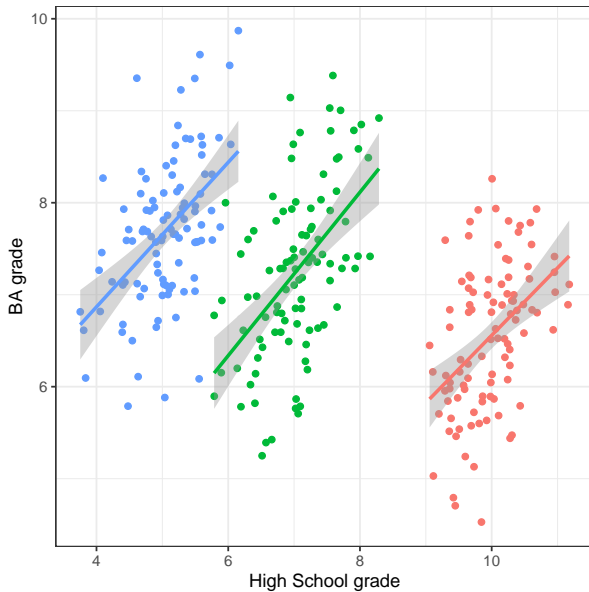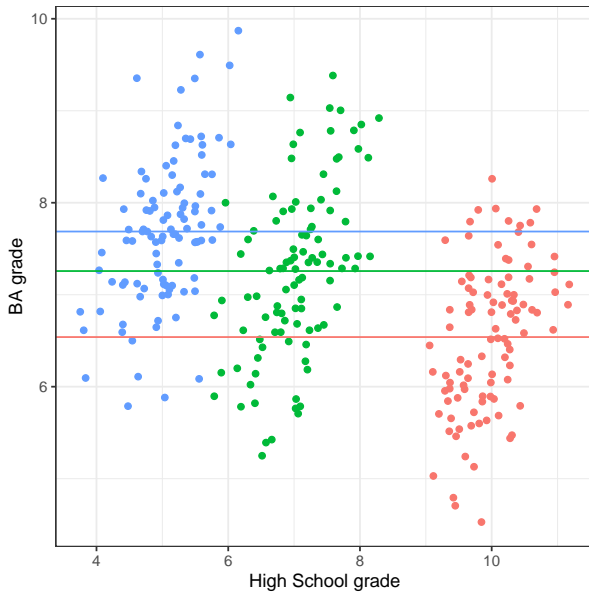# Fixed effects
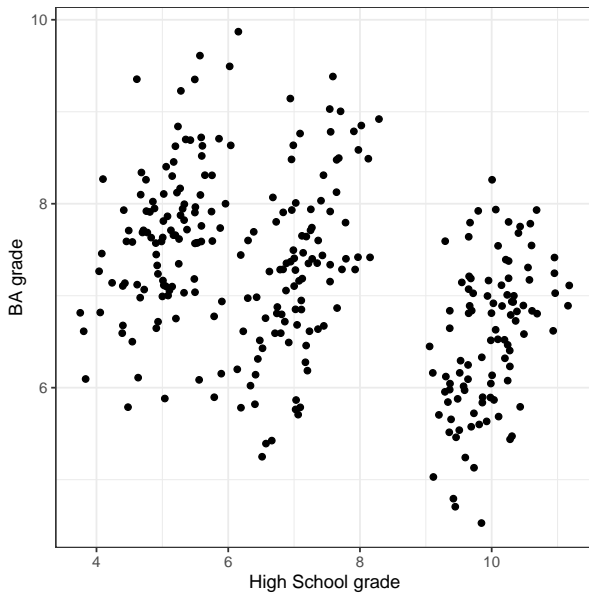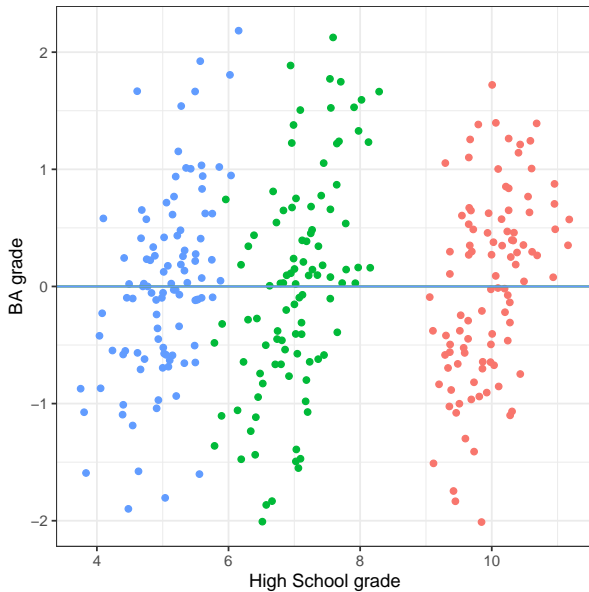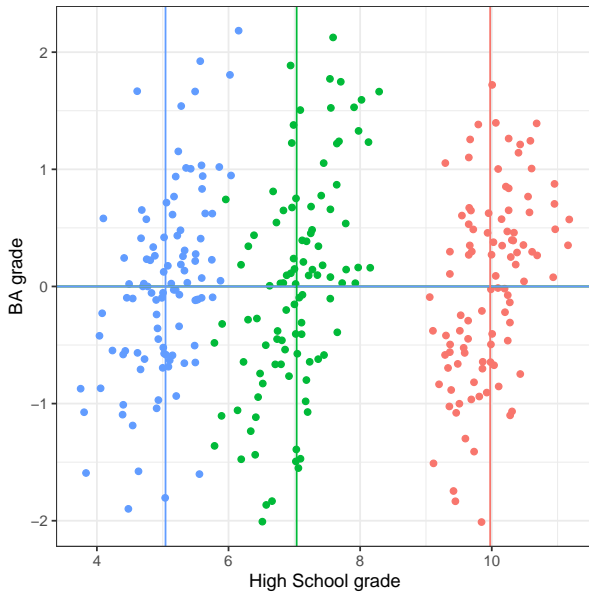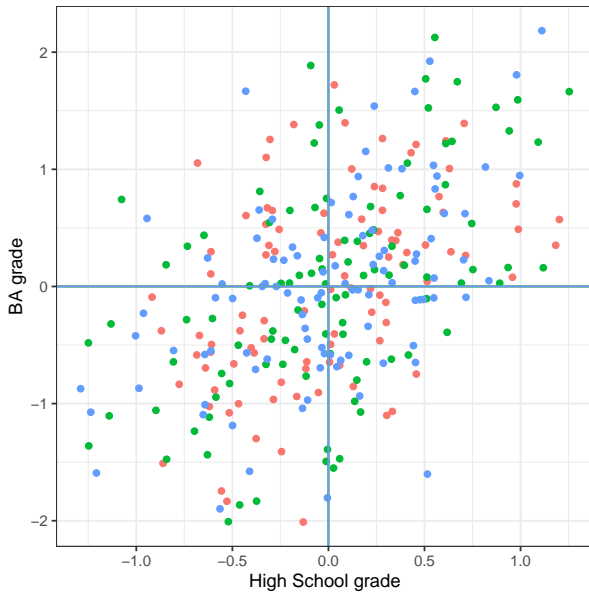
Fixed effects
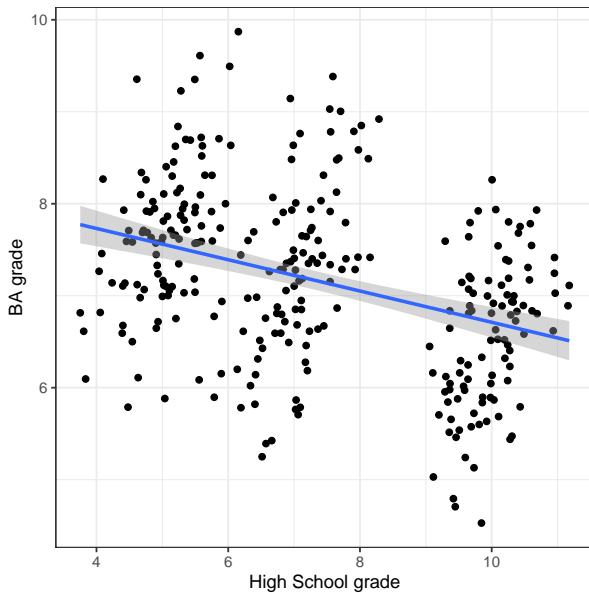
Fixed effects

Fixed effects
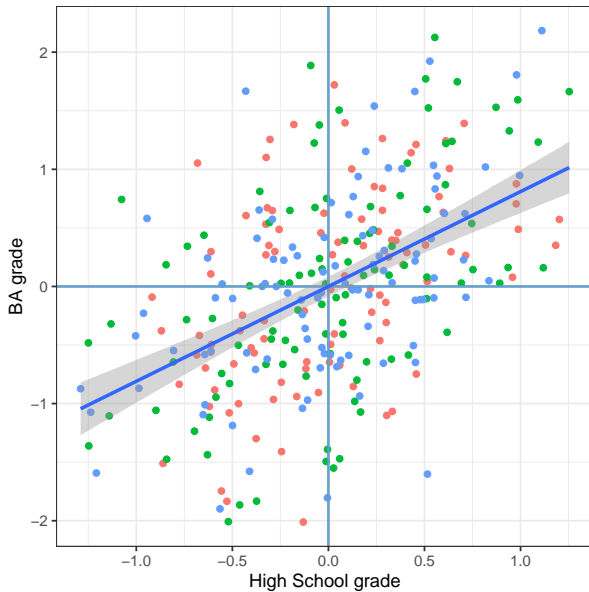
Fixed effects

Fixed effects

Fixed effects

Fixed effects

Fixed effects

Fixed effects

Difference-in-differences

- Treatments usually occur at a particular moment in time, e.g.:
  - $\rightarrow$ Minimum wage increase, terrorist attacks, influx of refugees, ...
- A naive idea would be to compare how it was *before* the treatment with how it is *after* the treatment, right?
  - $\rightarrow$ e.g. in municipalities that removed Francoist streets, how much did Vox grow between 2016 and 2019?
- or we could just compare treated vs control *after* treatment
  - $\rightarrow$ e.g. did Vox get more votes in 2019 in municipalities that removed F streets?
- DiD idea: use a control group that works as a counterfactual for how much would Y have changed without treatment

Difference-in-differences
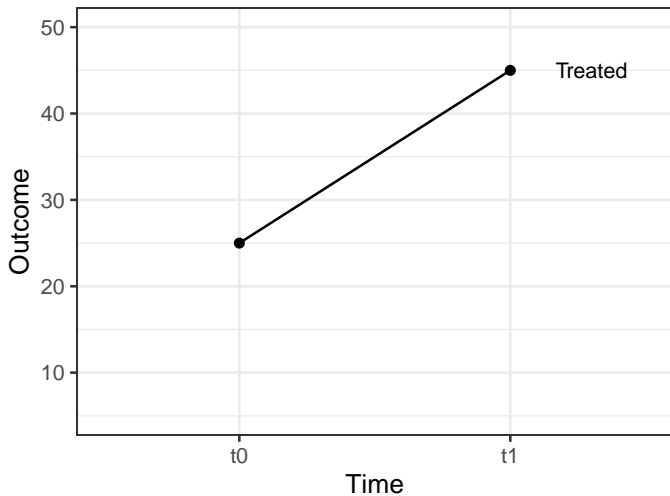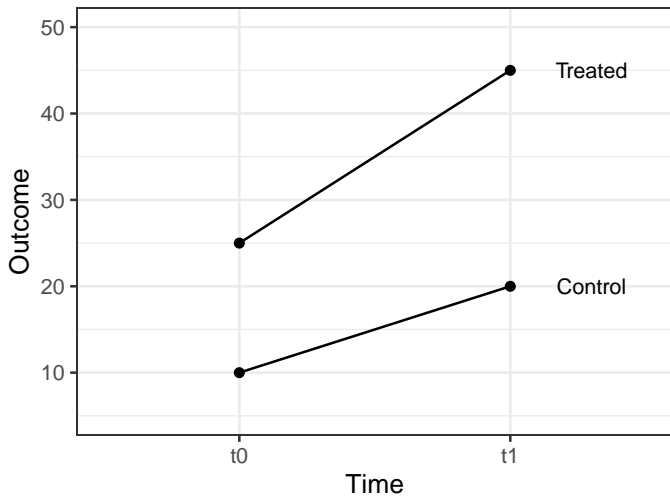
Difference-in-differences

Difference-in-differences
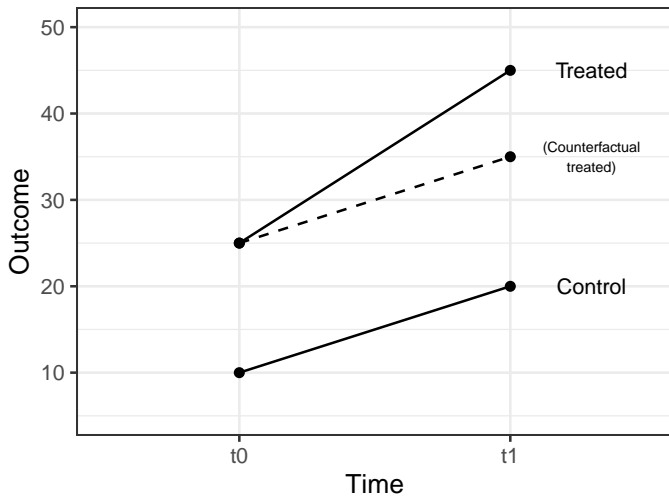
Difference-in-differences
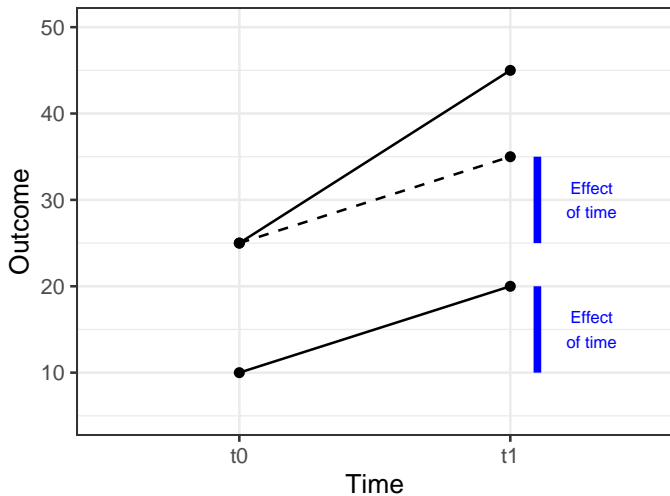
Difference-in-differences
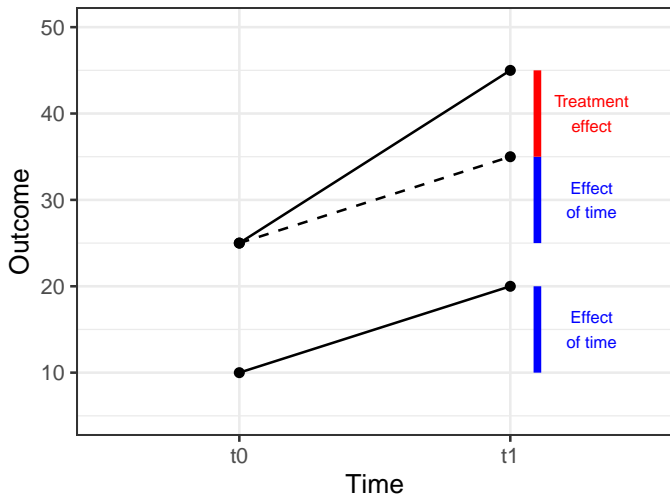
Difference-in-differences estimation

- You normally estimate it using a regression model (OLS or logit, depending on the outcome) on panel data
  - $\rightarrow$ each unit has one observation per period (at least, before/after treatment)

- Variables: time, group (treated or not), and their interaction

$$Y_{it} = \beta_0 + \beta_1 \, Treated_i + \beta_2 After_t + $$
$$\beta_3 (Treated_i \times After_t) + \beta^\top x_i + \epsilon_{it} \qquad (1)$$

- Key assumption : control group is a good counterfactual
  - $\rightarrow$ checking parallel trends assumption

Regression discontinuity design

- RDD works well when assignment into treatment depends on a cutoff along a running variable
  - $\rightarrow$ Do incumbent politicians have an electoral advantage? (vote share)
  - $\rightarrow$ What is the effect of being drafted into the military? (birth year)
  - $\rightarrow$ Effect of national policies in ethnic identification in Africa? (distance to colonial borders)
- The source of the exogenous variation:
  - $\rightarrow$ Although many variables confound the relationships between X and Y, nothing should be too different *around the cutoff* between treatment and control groups (local randomization assumption)

Regression discontinuity design

Figure 2: RD Treatment Effect in Sharp RD Design
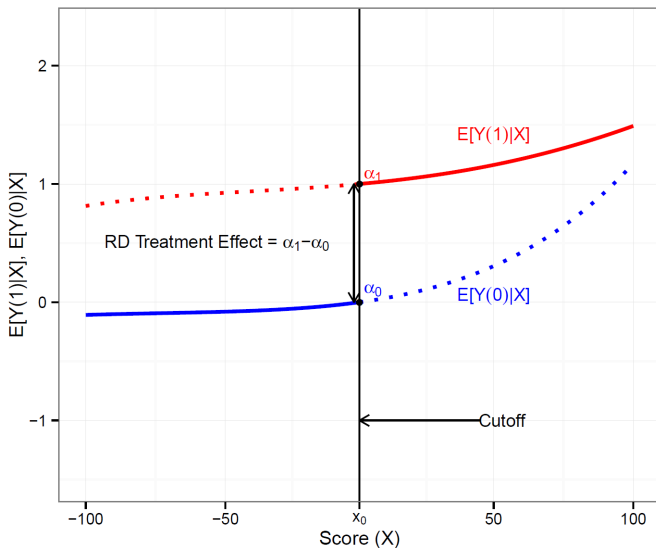
Regression discontinuity design

Figure 2: RD Treatment Effect in Sharp RD Design
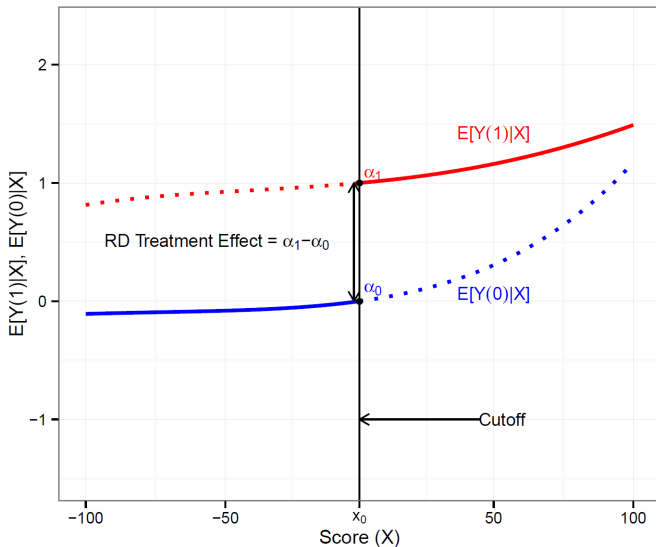
Regression discontinuity design

- Key assumption : other confounders also vary along the running variable, but are independent to the *jump over cutoff*
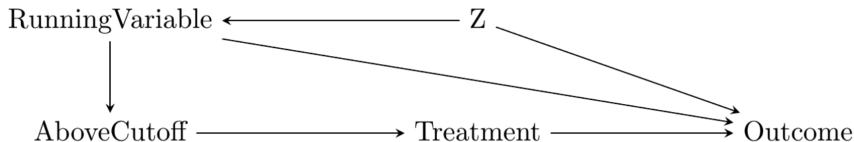
Huntington-Klein, *The Effect*, p.508

RDD estimates the LATE

(a) Raw Data
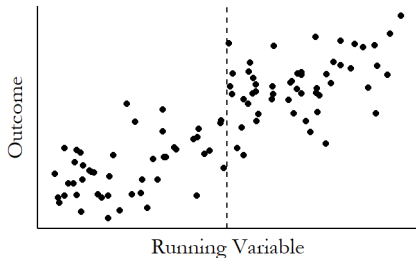
(b) Predict Values Near the Cutoff
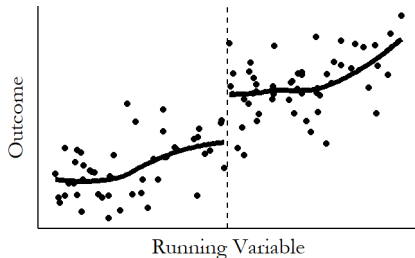
(c) Pick a Bandwidth
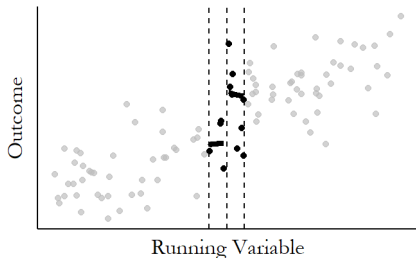
(d) Estimate Jump at the Cutoff
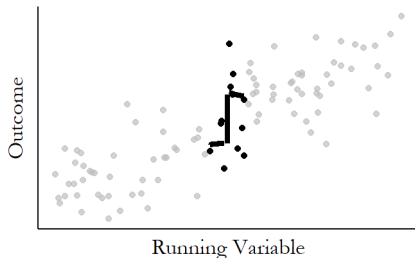
Threats in RDD

- Outcome → cutoff : what we study might actually affect cutoff (e.g. colonial borders and ethnic ID)
  - → could also be a case of confounding, cutoff ← Z → outcome
- Precise sorting : maybe something is *actually* affecting sorting around the cutoff (next example)
- Causal effect independent of treatment : being above/below cutoff affects Y independent of the treatment we're interesting in

Threats in RDD: precise sorting
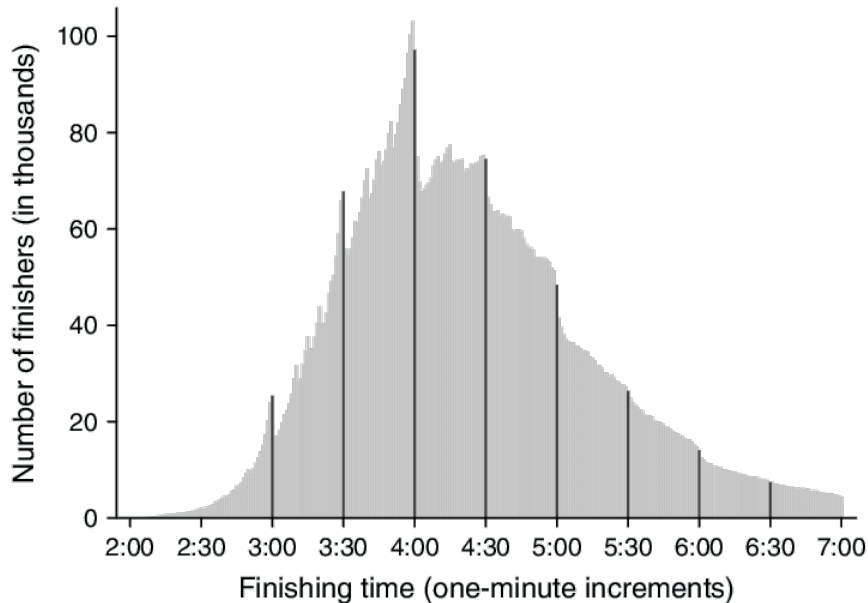
Instrumental variables
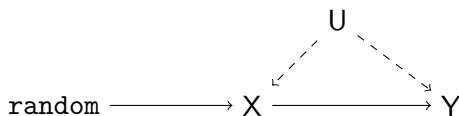
- Probably closest idea to a 'natural experiment'

Instrumental variables

- Find an exogenous source of variation in the treatment variable
  - → e.g., we want to know the effect of protests on government action, and we use *rainfall* as an instrument
- Isolate that variation and use it to identify the causal effect
  - → two-stage least squares, or 2SLS

Instrumental variables

- Find an exogenous source of variation in the treatment variable
  - $\rightarrow$ e.g., we want to know the effect of protests on government action, and we use *rainfall* as an instrument
- Isolate that variation and use it to identify the causal effect
  - $\rightarrow$ two-stage least squares, or 2SLS

Assumptions:

- Relevance : the instrument explains at least some part of the treatment variable
- Validity or exclusion restriction : no back door paths between the instrument and the outcome

Matching

- Main idea behind matching: adding control variables is *not the only way* to control / close back doors
- When you use a subset selecting on a variable, you are controlling for that variable
  - → e.g. if you analyze *income* ← *education* and only use data from individuals who grew up in big cities, you are already controlling for urban/rural
- Matching builds on this, and it is essentially about keeping in the same only comparable observations, or *matched pairs*
- Example of $Y \leftarrow$ *Treatment* matching on X

Matching

- Normally you match treated units on a few control units (so you would get the **ATT**)
- But in some cases you do the opposite, because you have more treated units that control (and you get the **ATC**)
    - → Example from Ukraine survey paper
- Two basic approaches:
    - → Distance matching: you select observations that have similar values on the matching variables
    - → Propensity score matching: we estimate the probability of being treated based on a set of variables, and then control/subset for that propensity score

Overall inference strategy

- In ideal experimental setting, we actually don't need any sophisticated statistics
  - → We can just compare the *mean* of the treatment group and the *mean* of the control group, that's it

Overall inference strategy

- In ideal experimental setting, we actually don't need any sophisticated statistics
  - $\rightarrow$ We can just compare the *mean* of the treatment group and the *mean* of the control group, that's it
- In observational data, these two types of tools (controlling + exploiting exogenous variation) are often used in combination
- But **always remember** these methods **depend** on a **causal model**
- Let's look at an example (from Huntington-Klein *The Effect*, kind of)

Controlling **and** exploting exogeneity

- Q: **Does pollution determine transport choice?**
- And say we are going to observe variation across *days*
    - $\rightarrow$ Outcome: car driving
    - $\rightarrow$ Treatment: pollution levels
- Clear problem of endogeneity, no? (driving in t_{-1}, economic activity, etc)

Z (unobserved stuff)

Pollution   Transport choice
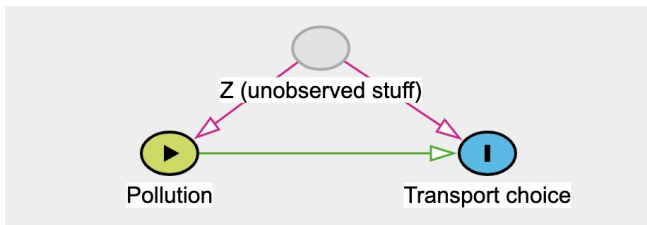
Chasing Clean Air: Pollution-Induced Travels in China

Shuai Chen, Yuyu Chen, Ziteng Lei, Jie-Sheng Tan-Soo

Abstract: This study uses "big" data to empirically investigate a highly effective, but underexamined way of reducing one's exposure toward air pollution—short-term travel. We determine subscribers' locations using mobile phones' signals and thereby establish linkages between air pollution and short-term population movements between cities in China. Using an instrumental variable based on daily variation in wind directions and pollution levels from distant upwind locations, we find that a one-unit increase in the origin city's air quality index (AQI) over the destination city's AQI increases short-term population flow from origin to destination by 0.15%. Further analyses provide richer characterizations of the decision-making process behind travel movements. Our findings add to the evidence base by examining in detail an understudied behavioral response toward air pollution.

Controlling **and** exploting exogeneity

- But it's not enough with wind direction, no?

Controlling **and** exploting exogeneity

- But it's not enough with wind direction, no?
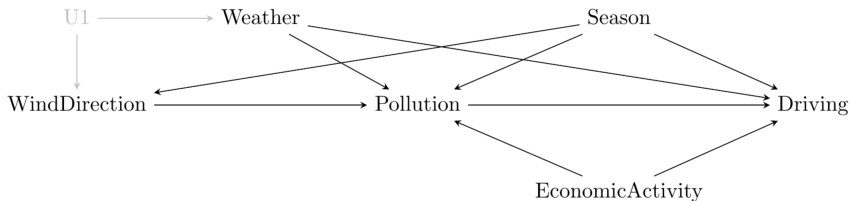
(And city, in this case)

Controlling **and** exploting exogeneity

- So in this case, we need to do two things:
1. Exploit exogenous variation
   - $\rightarrow$ In this case, using an instrumental variables approach, which isolates the variation in pollution which is explained by variation in wind direction

Controlling **and** exploting exogeneity

- So in this case, we need to do two things:
1. Exploit exogenous variation
   - → In this case, using an instrumental variables approach, which isolates the variation in pollution which is explained by variation in wind direction
2. *And* control for weather, season, and city
   - → Using a regression with control variables, or matching, or fixed effects

Roadmap

Intro and overview

Methods in brief

Paper discussion

Re-cap and essay guidelines

Paper discussion

Do TJ policies cause backlash? Evidence from street name changes in Spain

Francisco Villamil[1] ⑥ **and Laia Balcells[2]**

Abstract
Memories of old conflicts often shape domestic politics long after these conflicts end. Contemporary debates about past civil wars and/or repressive regimes in different parts of the world suggest that these are sensitive topics that might increase political polarization, particularly when transitional justice policies are implemented and political parties mobilize discontentment with such policies. One such policy recently debated in Spain is removing public symbols linked to a past civil war and subsequent authoritarian regime (i.e., Francoism). However, the empirical evidence on its impact is still limited. This article attempts to fill this gap by examining the political consequences of street renaming. Using a difference-in-differences approach, we show that the removal of Francoist street names has contributed to an increase of electoral support for a new far-right party, Vox, mainly at the expense of a traditional right-wing conservative party, PP. Our results suggest that revisiting the past can cause a backlash among those ideologically aligned with the perpetrator, and that some political parties can capitalize on this.

Keywords
Transitional justice, voting, conflict memories, Spain

Roadmap

Re-cap and final essay

- **Groups?** Send me an email **before Oct 15th**

Re-cap and final essay

1. Problem/topic
   → unique to group, general issue

2. Stories, arguments about mechanisms

3. Research question

4. Proper theory, concepts and operationalization

5. Measurement, unit of analyses, data sources, etc

6. Inference strategy

7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
   $\rightarrow$ testing an argument, testing 2 against each other
3. Research question
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
   $\rightarrow$ operationalize the story you have in mind
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
4. Proper theory, concepts and operationalization
   - $\rightarrow$ develop a proper argument with view to empirical test
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
   $\rightarrow$ ideally think about more than one strategy
6. Inference strategy
7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
   $\rightarrow$ depends on previous steps: one per 'arm' or multiple strategies
7. Results & interpretation

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
7. Results & interpretation
   $\rightarrow$ what do we really learn about the initial topic? next steps?

Re-cap and final essay

1. Problem/topic
2. Stories, arguments about mechanisms
3. Research question
4. Proper theory, concepts and operationalization
5. Measurement, unit of analyses, data sources, etc
6. Inference strategy
7. Results & interpretation

- **General guideline:** maybe it's not realist you do it, but it needs to be feasible (data collection, etc)

Re-cap and final essay

- Logistics:
    - → 10-15 slots
    - → 10min presentation + 10min feedback
    - → If less groups, we can increase time
- **Questions?**