The Statistics of Causal Inference

Elias Dinas
elias.dinas@politics.ox.ac.uk

Oxford Spring School in Advanced Research Methods
Oxford, 27-31 March, 2017

Outline

1. Asking Research Questions
2. The Potential Outcomes Framework
3. Matching
4. Instrumental Variables
5. Regression Discontinuity Design
6. Difference-in-Differences
7. Extensions
8. Lab Session: Applications

There are Questions & there are Questions

Descriptive

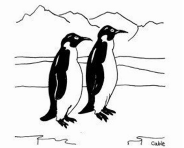
- Do PR systems redistribute more than FPTP systems?
- How popular has been Angela Merkel's decision to accept 1m refugees in Germany?
- Who votes for the radical right?

Explanatory (from Y to $X(s)$)

- Why do PR systems redistribute more than FPTP systems?
- Why are Americans less healthy than Canadians even if they spend a higher portion of their GDP on health?
- Why did the Russian Revolution happen in Russia and not in Britain?

Causal (from X to Y)

- What is the effect of the electoral system on redistribution?
- What is the effect of polluted water on cholera?
- What is the effect of democracy on economic development?



"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

There are Questions & there are Questions

Descriptive

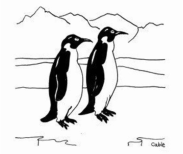
- Do PR systems redistribute more than FPTP systems?
- How popular has been Angela Merkel's decision to accept 1m refugees in Germany?
- Who votes for the radical right?

Explanatory (from Y to $X(s)$)

- Why do PR systems redistribute more than FPTP systems?
- Why are Americans less healthy than Canadians even if they spend a higher portion of their GDP on health?
- Why did the Russian Revolution happen in Russia and not in Britain?

Causal (from X to Y)

- What is the effect of the electoral system on redistribution?
- What is the effect of polluted water on cholera?
- What is the effect of democracy on economic development?



"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

Definitions

Causality

Refers to the relationship between events where one set of events (the effects) is a direct consequence of another set of events (the causes). (Hidalgo & Sekhon 2012)

Causal Inference

The process by which one can use data to make claims about causal relationships. (Hidalgo & Sekhon 2012)

Inferring causal relationships is a central task of science:

Examples

- What is the effect of peace-keeping missions on peace?
- What is the effect of class size on educational outcomes?
- What is the effect of church attendance on social capital?
- What is the effect of minimum wage on employment?
- What is the effect of schooling on earnings?

Establishing Criteria for Causal Inference

Some Background

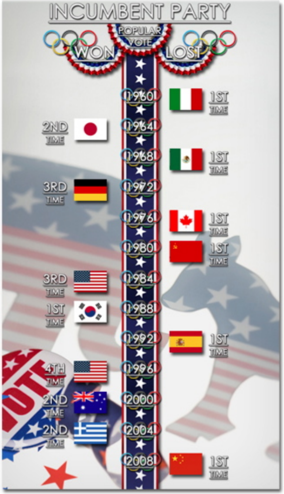
Hume: An Inquiry into Human Understanding (1751)

Regularity Models:

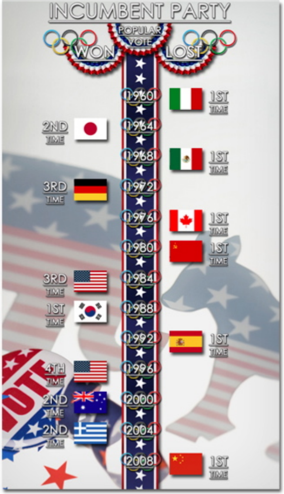
- Contiguity: Cause and effect must be contiguous in time and space
- Succession: Cause prior to event)
- Constant Conjunction: Constant union (association) between cause and effect

How could we know a flame caused heat? Only by calling "to mind their constant conjunction in all past instances. Without further ceremony, we call the one cause and the other effect, and infer the existence of one from that of the other."

Are then these relationships causal?



Are then these relationships causal?



Best Picture 1991:
Presidential Election 1992:

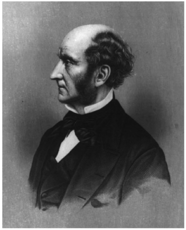
Incumbent Lost

Best Picture 1999:
Presidential Election 2000:
Incumbent Lost

Best Picture 2007:
Presidential Election 2008:

Incumbent Lost

A Counterfactual Logic



Background

John Stuart Mill: A System of Logic (1843)

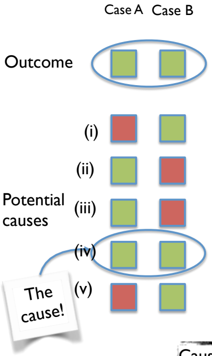
A series of “canons” for inductive inference. Main addition: “No plausible alternative explanation of the effect under study.” Logic leading to the two well-known methods of comparison (he came up with more than two):

- Method of Difference: Units similar in all respects but one manipulable treatment. Units differ in their outcomes.
- (Method of Indifference: Units dissimilar in all respects but one. Outcome same in both units.)

Mill's Methods

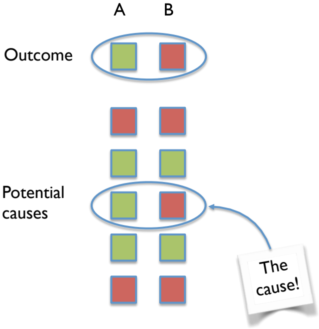
Suppose all of the potential causes can be enumerated and accurately measured. Then these two methods will *under certain conditions* tell us the cause of an outcome:

Method of agreement



		Case A	Case B
Outcome		🟩	🟩
Potential causes	(i)	🟥	🟩
	(ii)	🟩	🟥
	(iii)	🟩	🟥
	(iv)	🟩	🟩
	(v)	🟥	🟩

Method of difference



	A	B
Outcome	🟩	🟥
Potential causes	🟥	🟥
	🟩	🟩
	🟩	🟥
	🟩	🟩
	🟥	🟥

Causal inference from just two cases!!!

1

Mill & the Counterfactual Logic of Causality

Causal Inference as a counterfactual Problem

Rather than defining causality purely in reference to observable events, counterfactual models define causation in terms of a comparison of observable and unobservable events.

- We need to construct counterfactuals of the observed world as comparison units (Method of Difference).

> *In England, westerly winds blow during about twice as great a portion of the year as easterly. If, therefore, it rains only twice as often with a westerly as with an easterly wind, we have no reason to infer that any law of nature is concerned in the coincidence. If it rains more than twice as often, we may be sure that some law is concerned; either there is some cause in nature which, in this climate, tends to produce both rain and a westerly wind, or a westerly wind has itself some tendency to produce rain.* (Mill 1873:347-7)

$$H : P(rain|westerly\ wind, W) > P(rain|not\ westerly\ wind, W)$$

The problem lies in W. To see this, we need a proper framework of causal inference.

Putting Mill into Context

$$X\ (Years\ of\ Schooling) \xrightarrow{\hspace{4cm}} Y\ (Earnings)$$

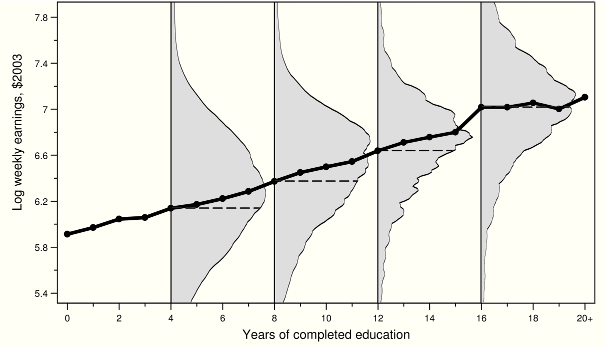


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

The Conditional Expectation Function I

Our first aim is to summarize the relationship between X & Y. To do that we find the average wage for each educational category.

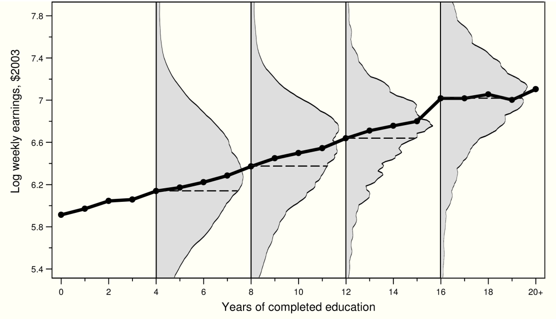


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

Why are we interested in the mean?

The Conditional Expectation Function I

Our first aim is to summarize the relationship between X & Y. To do that we find the average wage for each educational category.

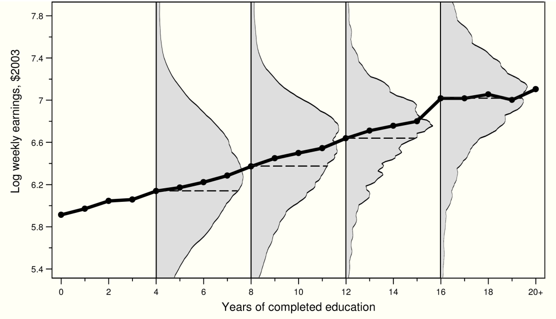


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

Why are we interested in the mean?

The Conditional Expectation Function II

Connecting the dots, we have our summary of the conditional expectation of Y, given X. We can repeat this for all education categories. We can succinctly write this as follows: $E(Y|X)$.

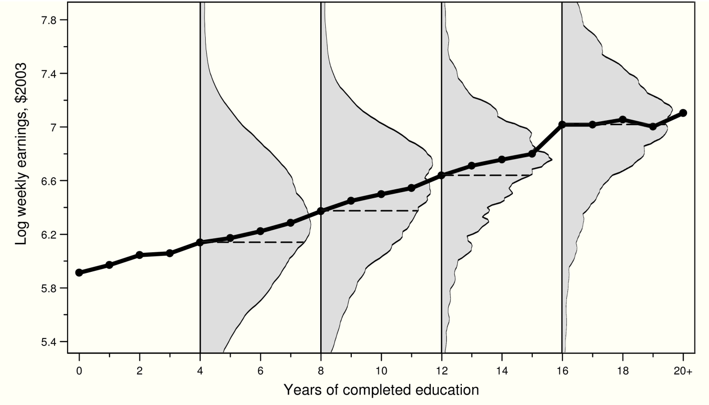


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

A Pause for Notation

- Informally, *E* stands for *best guess*.
  - This translates into the *population average* and is known as the *Expectation* of Y:

Definition of Expectations

Let Y_i denote a random variable, which takes values across observations $i = 1, \ldots, N$. $E[Y_i]$ is the population average of this variable. Two examples:

- If Y_i is generated by a random process (e.g. rolling a die), $E[Y_i]$ is the average in infinitely many repetitions of this process.
- If Y_i comes from a survey (as in this case), $E[Y_i]$ is the average obtained if everyone in the population from which the sample is drawn is enumerated.

  - The expectation of Y given . . . ?
  - ... given X.
  - This is denoted by the symbol "$|$".
- Taken altogether, the notation can be read as: $E(Y|X = x)$: The Conditional Expectation of Y given the value of X.

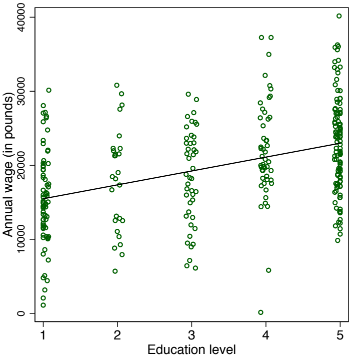
The Linear Equation

The easiest, most parsimonious although not always most adequate, way to summarize the conditional expectation of Y, given X, is to assume a common slope: a straight line: $E(Y|X) = \beta_0 + \beta_1 X$.

- So, collecting our own data or using data others have already collected, we fit a model of the following generic form:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

The Regression Prophecy



- Think of Y as *Salary* and X as *years or schooling*. What we are interested in is β_1. We would like to interpret this coefficient as the *average change produced in individuals' wages by one more year of schooling.*

The Key OLS (or any other regression) Assumption

OLS estimation and inference requires various assumptions but we are interested in one of them here. The one that would allow us to interpret β_1 as the average causal effect of education on economic well-being.

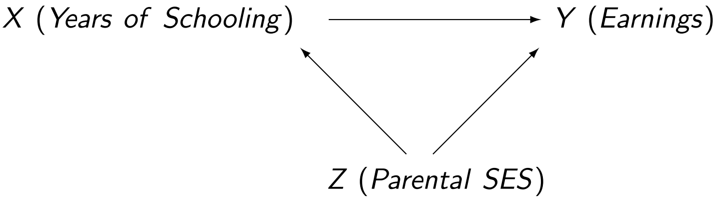
The zero-mean assumption

The error term, u_i, is centered around the same value (trivially 0) across all values of X: The explanatory variable (our regressor) and the error term are *independent*.

Any ideas whether this assumption is likely to hold in our example?

The selection Problem

For the moment the way to construct counterfactuals is to think about $Z's$ and add them into the equation. But we only include what we can observe. One example:



X (*Years of Schooling*)

Y (*Earnings*)

Z (*Parental SES*)

This is the selection problem, also known as: unobserved heterogeneity, omitted variable bias, spurious correlation, or endogeneity.

Addressing The Problem: Two Approaches

Model-Based Approaches

- Multiple Regression Analysis (adding X's in a regression framework)
- SEMs & Cross-Lagged Models

Model-Free Approaches

- Conditioning on Observables:
    - Matching
- Without Conditioning On Observables:
    - Randomized Experiments
    - Design-based methods that approach the experimental ideal

Addressing The Problem: Two Approaches

Model-Based Approaches

- Covariate Adjustment in a Regression Framework
- SEMs & Cross-Lagged Models

Model-Free Approaches

- Conditioning on Observables:
  - Matching
- Without Conditioning On Observables:
  - Randomized Experiments
  - Design-based methods that approach the experimental ideal

The Importance of Design

How to address it: Before collecting the data

> *We use conditional probabilities to learn about counterfactuals of interest – e.g. would Jane have voted if someone from the campaign had not gone to her home to encourage her to do so? One has to be careful to establish the relationship between the counterfactuals of interest and the conditional probabilities one has managed to estimate. Researchers too often forget that this relationship must be established by design and instead rely upon statistical models whose assumptions are almost never defended. Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive. This conclusion has implications for the kind of causal questions we are able to answer with some rigor. Clear, manipulable treatments and rigorous designs are essential.*
>
> (Sekhon, 2008:272)

Motivation

A turn into design-based identification strategies

> *Like a fashion trend traveling from New York to the heartland, soul-searching about causality has made its way from empirical research in economics to that in political science with the usual lag. Gone are the days when it is enough to have a nice theory, a conditional correlation and some rhetoric about the implausibility of competing explanations while implying but assiduously avoiding the 'c' word. Editors, reviewers, and search committees are beginning to look for more explicit and careful empirical treatments of causality . . .*

(Rodden 2007:1)

Motivation

A turn into design-based identification strategies

> *. . . researchers are expected to lay out a set of possible outcomes, or counterfactuals, generated by a set of determinants, and demonstrate that holding all possible determinants except one at a constant level, the manipulation of that determinants is associated with a specific change in outcome, which can be deemed a causal effect.*

(Rodden 2007:1)

The Potential Outcomes framework

Tom's Dilemma

- Take an Aspirin or Not? → Headache
- Do a Master's Degree or try to get a job? → Salary in ten-years time

Carthenia's Dilemma

- Presidential or Parliamentary? → Democratic Survival
- FPTP or PR? → Redistribution
- Bicameral or Unicameral? → Law production

The Intuition 1

The Starting Point

Causality is tied to an action, which you can call it also manipulation, treatment, intervention. This action is applied to a *unit*. Time is divided into what exists *before* and what is observed *after* this action. This means that we have a pair of units and time periods:

Tom

- Tom × Before the Aspirin
- Tom × After the Aspirin

Carthenia

- Carthenia × Before the electoral reform
- Carthenia × After the electoral reform

The Intuition 2

A Conventional Restriction

We typically focus on two actions, although everything we discuss can be generalized to more actions. Typically, these two actions are distinguished in the following way:

- Do A vs $Not - A$
- Do A vs B

More often than not, A is a more active treatment (take an Aspirin) and B (or not-A) is a more passive one (not taking an aspirin).

Counterfactual Logic

Only one action can be actually taken. But both of them are in principle equally likely. So, we can forget for the moment which one was actually taken and start thinking how the world/Tom/Carthenia would look like after either A or $Not - A(B)$.

Potential Outcomes

For each action, there is a corresponding outcome:

- How is Tom's headache with and without having previously taken an aspirin?
- How much does Carthenia redistribute with an FPTP vs a PR system?

Of course, Tom can only do one of the two things: either take or not take an aspirin. Similarly for Carthenia.

- Given two actions, there are two corresponding Potential Outcomes
- Call them Y_1 & Y_0.
- Y_1: The potential outcome realised if having taken the treatment (e.g. aspirin).
- Y_0: The potential outcome realised if not having taken the treatment (e.g. aspirin).

Defining a Causal Effect

A causal effect stems from a causal question:

- What's the effect of the aspirin on Tom's headache?
- What's the effect of PR on Carthenia's redistribution levels?

Given the potential outcomes, the answer is straightforward:

- Compare Tom's headache had he taken the apirin and had he not taken the aspiring: $Y_1 - Y_0$.
- Compare Carthenia's level of redistribution having Carthenia used a PR system and having Carthenia used a different electoral system: $Y_1 - Y_0$.

Causal Effect

For each Tom/Carthenia: $\tau = Y_{1,Tom/Carthenia} - Y_{0,Tom/Carthenia}$

Defining Causal Effects with Potential Outcomes

- The definition depends only on the potential outcomes, not on which of these is actually realized. We remain agnostic about whether A or B is realized.
- The causal effect comes from a comparison of the same unit in the same time point, after the action (either A or B) has been realized.

What we do not Do then:

- Compare Tom's before and after the headache.
- Compare Carthenia's redistribution levels with a country using a different electoral system, say Erstland.

The Problem

We cannot, of course, observe Tom both having and not having taken the aspirin. We cannot compare Carthenia under the same conditions and in the same time point having and not having used a PR system.

Or maybe we can?

The take-home point

For the estimation of causal effects, we will need to make different comparisons from the comparison made for their definitions.

We now turn into a more formal exposition of the counterfactual framework.

The Counterfactual Model

The Underlying Logic

- Causal inference as a missing data problem
- The framework applies in either quantitative or qualitative studies

The Ingredients

- Well-defined causal states to which all members of the population of interest could be exposed

Intellectual Background

- Fisher (designing experiments), and most importantly Neyman 1923; Rubin 1974: The Neyman-Rubin model

Defining the Potential Outcomes

Definition: Treatment

D_i : Indicator of treatment status for unit i

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases}$$

Definition: Observed Outcome

Y_i : Observed outcome variable of interest for unit i. (Realized after the treatment has been assigned)

Definition: Potential Outcomes

Y_{0i} and Y_{1i}: Potential Outcomes for unit i

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i} & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

Causality with Potential Outcomes

Definition: Causal Effect

Causal Effect of the treatment on the outcome for unit *i* is the difference between its two potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

Could be the ratio $(\frac{Y_{1i}}{Y_{0i}})$ or any other function of Y_{1i} & Y_{0i}, but we are predominantly interested in the difference between the two causal states because we assume linear individual-level effects.

The Fundamental Problem of Causal Inference

Definition (Holland, 1986)

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i.

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

	Y_{1i}	Y_{0i}
Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Control Group ($D = 0$)	Counterfactual	Observable as Y

The Fundamental Problem of Causal Inference

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

	Y_{1i}	Y_{0i}
Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Control Group ($D = 0$)	Counterfactual	Observable as Y

Holland revisited

Individuals contribute outcome information only for the treatment state in which they are observed. Realized outcomes contain only a portion of the information we need to directly calculate causal effects for all units.

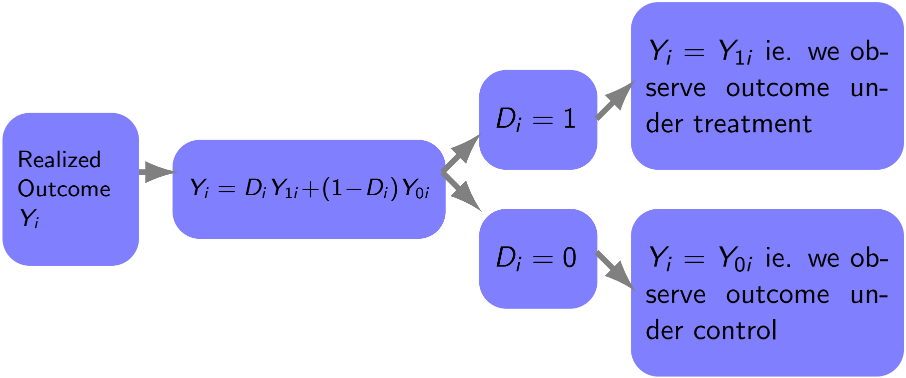
Heterogeneity in Treatment Effects

Problem

Causal inference is difficult because it involves missing data. How can we find $\tau = Y_{1i} - Y_{0i}$?

- A large amount of homogeneity could solve this problem:
  - (Y_{1i}, Y_{0i}) constant across individuals
  - (Y_{1i}, Y_{0i}) constant across time
- Unfortunately, often there is a large degree of heterogeneity in the individual responses to treatments in the social sciences.

From Potential Outcomes to Observed Outcomes



Realized Outcome Y_i

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

$D_i = 1$

$Y_i = Y_{1i}$ ie. we observe outcome under treatment

$D_i = 0$

$Y_i = Y_{0i}$ ie. we observe outcome under control

Question

Is there any assumption hidden here?

Stable Unit Treatment Value Assumption (SUTVA)

An Assumption

The Observed outcome for each unit is realized as:

$$Y_i = D_i \cdot Y_{1i} + (1 - D_i) Y_{0i} \quad \text{so} \quad Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

This is SUTVA. It implies that potential outcomes for unit i are unaffected by treatment assignment for unit j

- Rules out inteference among units:
  - Speed Boarding on flight satisfaction or any situation in which treatment becomes less effective when more units take it.

Definition: Rubin 1985

SUTVA is simply the a priori asumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive.

Explosion of Potential Outcomes (SUTVA)

Let $\mathbf{D} = D_i, D_j$ be a vector of treatment assignments for two units, i ("Noah") and j ("Maria"). How many elements in $\mathbf{d}$?

$$\mathbf{D} = \{(D_i = 0, D_j = 0), (D_i = 1, D_j = 0), (D_i = 0, D_j = 1), (D_i = 1, D_j = 1)\}$$

How many potential outcomes for unit i?

$$Y_{1i}(\mathbf{D}) = \begin{cases} Y_{1i}(1,1) \\ Y_{0i}(1,0) \end{cases} \quad Y_{0i}(\mathbf{D}) = \begin{cases} Y_{0i}(0,1) \\ Y_{0i}(0,0) \end{cases}$$

$$\tau_i(\mathbf{D}) = \begin{cases} Y_{1i}(1,1) - Y_{0i}(0,0) \\ Y_{1i}(1,1) - Y_{0i}(0,1) \\ Y_{1i}(1,0) - Y_{0i}(0,0) \\ Y_{0i}(1,0) - Y_{0i}(0,1) \\ Y_{1i}(1,1) - Y_{1i}(1,0) \\ Y_{0i}(0,1) - Y_{0i}(0,0) \end{cases}$$

Explosion of Potential Outcomes (SUTVA)

Let $\mathbf{D} = D_i, D_j$ be a vector of treatment assignments for two units, i ("Noah") and j ("Maria"). How many elements in $\mathbf{d}$?

$$\mathbf{D} = \{(D_i = 0, D_j = 0), (D_i = 1, D_j = 0), (D_i = 0, D_j = 1), (D_i = 1, D_j = 1)\}$$

How many potential outcomes for unit i?

$$Y_{1i}(\mathbf{D}) = \begin{cases} Y_{1i}(1,1) \\ Y_{0i}(1,0) \end{cases} \quad Y_{0i}(\mathbf{D}) = \begin{cases} Y_{0i}(0,1) \\ Y_{0i}(0,0) \end{cases}$$

SUTVA implies $Y_{1i}(1,1) = Y_{1i}(1,0)$ and $Y_{0i}(0,1) = Y_{0i}(0,0)$ and allows us to define the effect for unit i $\tau_i = Y_{1i} - Y_{0i}$ independent of the treatment assignment process itself.

SUTVA is an exclusion restrictions. We rely on outside information to rule out the possibility of certain causal effects (eg. your taking the treatment has no effect on my outcome). Causal inference is impossible without such exclusion restrictions.

Quantities of Interest

Definition ATE

Average Treatment Effect:

$$\tau_{ATE} = E[Y_1 - Y_0]$$

Definition ATT

Average Treatment Effect of the Treated:

$$\tau_{ATT} = E[Y_1 - Y_0 | D = 1]$$

Definition ATC

Average Treatment Effect of the Controls:

$$\tau_{ATC} = E[Y_1 - Y_0 | D = 0]$$

These effects can be extended to specific population subgroups:

$$\tau_{ATE} = E[Y_1 - Y_0 | X = x] \quad \tau_{ATT} = E[Y_1 - Y_0 | D = 1, X = x]$$
$$\tau_{ATC} = E[Y_1 - Y_0 | D = 0, X = x]$$

Interpretation of Treatment Effects

Think of the following example: The effect of `Medicaid` on Americans' health status.

ATE

- The effect of `Medicaid` on the health status of the average American. How would Americans' health look like had there be no Affordable Care Act?

ATT

- The effect of the `Medicaid` on the health status for those who are actually enrolled. Would those who subscribed to `Medicaid` have a better or a worse health condition if they had not subscribed?

ATC

- The effect of the `Medicaid` on the health status for those who are not enrolled. Would those who have not subscribed to `Medicaid` have a better or a worse health condition if they had subscribed?

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	?	?	?
2	1	1	?	?	?
3	0	0	?	?	?
4	1	0	?	?	?

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	1	0	?	1	?

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	3	0	?
2	1	1	1	1	?
3	0	0	1	0	?
4	1	0	1	1	?

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0

What is the ATE?

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}]$$

An Example: ATE

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0
$E[Y_1]$			1.5		
$E[Y_0]$				0.5	
$E[Y_1 - Y_0]$					1

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = 1/4 \cdot (3 + 0 + 1 + 0) = 1$$

An Example: ATT

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATT}
1	3	1	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	1	0	1	1	0

What is the ATT?

$$\tau_{ATT} = E[Y_{1i} - Y_{0i} | D = 1]$$

An Example: ATT

Imagine a population of 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	τ_{ATE}
1	3	1	3	0	3
2	1	1	1	1	0
$E[Y_1\|D=1]$			2		
$E[Y_0\|D=1]$				0.5	
$E[Y_1 - Y_0\|D=1]$					1.5

$$\tau_{ATE} = E[Y_{1i} - Y_{0i}|D=1] = 1/2 \cdot (3+0) = 1.5$$

An Exercise for you: Try to estimate the ATC.

Selection Bias

The Problem

- Now let's reconsider the *fundamental problem of causal inference*.
- We only observe Y_{1i} if $D = 1$ and Y_{0i} if $D = 0$.
- So, the only thing we can do is compare the observed outcomes: $E[Y_i|D = 1] - E[Y_i|D = 0]$.
- Is that of any use? Does this represent any quantity of interest thus far?

Definition: Selection Bias

$$\underbrace{E[Y_i|D=1] - E[Y_i|D=0]}_{\text{Observed Differences in Outcome}} = \underbrace{E[Y_{1i}|D=1] - E[Y_{0i}|D=1]}_{\text{Average Treatment Effect of the Treated}}$$

$$+ \underbrace{E[Y_{0i}|D=1] - E[Y_{0i}|D=0]}_{\text{Selection Bias}}$$

Selection Bias

What this means

- It is unlikely that the last two terms will sum to zero in the previous equation
- Selection into treatment is associated with the potential outcomes

An Example: Selection Bias

Church attendance is often found to be correlated with turnout. Does this mean it induces turnout? Church goers differ in various respects from people who do not attend religious ceremonies, etc they have higher sense of civic duty. Thus, turnout for them would already be higher than for non-church goers even if individuals who attend religious ceremonies did not do so.

$$E[Y_{0i}|D=1] - E[Y_{0i}|D=0] > 0$$

Selection Bias: An Exercise

Observing treated and untreated units

- Do hospitals make us healthier?
- Do economic perceptions predict vote choice?
- Does perceive efficiency in government predict voter preferences?
- Do job training programs help the disadvantaged?

A common pattern: The problem lies in the baseline status. Under no treatment, some units would be already more likely to register the outcome of interest than others. They select themselves into the treatment.

Selection Bias in a Regression Framework

Remember the assumption about how potential outcomes are linked to treatment status:

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

easily rewritten as:

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

The expression is useful because remember: $\tau_i = Y_{1i} - Y_{0i}$. Now, assume that the treatment effect is constant for all i, so that: $\tau = Y_{1i} - Y_{0i}$ is a constant. With constant treatment effects, we can rewrite this equation as:

$$Y_i = \underbrace{\alpha}_{E(Y_{0i})} + \underbrace{\tau}_{Y_{1i} - Y_{0i}} D_i + \underbrace{u_i}_{Y_{0i} - E(Y_{0i})}$$

Evaluating the conditional expectation of this equation with treatment switched off and on:

$$E[Y_i | D = 1] = \alpha + \tau + E[u_i | D = 1]$$

$$E[Y_i | D = 0] = \alpha + E[u_i | D = 0]$$

So that:

$$E[Y_i | D = 1] - E[Y_i | D = 0] = \underbrace{\tau}_{\text{ATE}} +$$

$$\underbrace{E[u_i | D_i = 1] - E[u_i | D_i = 0]}_{\text{selection bias}}$$

Selection Bias in a Regression Framework

Remember the assumption about how potential outcomes are linked to treatment status:

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

It can be easily rewritten as:

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

The expression is useful because remember: $\tau_i = Y_{1i} - Y_{0i}$. Now, assume that the treatment effect is constant for all i, so that: $\tau = Y_{1i} - Y_{0i}$ is a constant. With constant treatment effects, we can rewrite this equation as:

$$Y_i = \underbrace{\alpha}_{E(Y_{0i})} + \underbrace{\tau}_{Y_{1i} - Y_{0i}} D_i + \underbrace{u_i}_{Y_{0i} - E(Y_{0i})}$$

Evaluating the conditional expectation of this equation with treatment switched off and on:

$$E[Y_i | D = 1] = \alpha + \tau + E[u_i | D = 1]$$
$$E[Y_i | D = 0] = \alpha + E[u_i | D = 0]$$

So that:

$$E[Y_i | D = 1] - E[Y_i | D = 0] = \underbrace{\tau}_{\text{ATE}} +$$

Selection bias = violation of conditional mean assumption →

$$\underbrace{E[u_i | D_i = 1] - E[u_i | D_i = 0]}_{\text{selection bias}}$$

Selection Bias in a Regression Framework

Remember the assumption about how potential outcomes are linked to treatment status:

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

It can be easily rewritten as:

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

The expression is useful because remember: $\tau_i = Y_{1i} - Y_{0i}$. Now, assume that the treatment effect is constant for all i, so that: $\tau = Y_{1i} - Y_{0i}$ is a constant. With constant treatment effects, we can rewrite this equation as:

$$Y_i = \underbrace{\alpha}_{E(Y_{0i})} + \underbrace{\tau}_{Y_{1i} - Y_{0i}} D_i + \underbrace{u_i}_{Y_{0i} - E(Y_{0i})}$$

Evaluating the conditional expectation of this equation with treatment switched off and on:

$$E[Y_i|D = 1] = \alpha + \tau + E[u_i|D = 1]$$

$$E[Y_i|D = 0] = \alpha + E[u_i|D = 0]$$

So that:

$$E[Y_i|D = 1] - E[Y_i|D = 0] = \underbrace{\tau}_{\text{ATE}} +$$

Selection bias = correlation of regressor with the error term →

$$\underbrace{E[u_i|D_i = 1] - E[u_i|D_i = 0]}_{\text{selection bias}}$$

Selection Bias in a Regression Framework

Since $u_i = Y_{0i} - E(Y_{0i})$, we can rewritte the last equation as:

$$E[u_i|D_i = 1] - E[u_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

This correlation reflects the difference in the baseline: the fact that potential outcomes differ between treated and untreated units in their no-treatment condition.

- To see what this means think about the hospital example.
- This is why we call this problem selection bias.

The Assignment Mechanism

How to proceed: Focus on the assignment mechanism

Definition: The Assignment Mechanism

Assignment mechanism is the procedure that determines which units are selected into the treatment. The following mechanisms are available, arranged from more to less satisactory:

- Random Assignment
- Selection on Observables
- Selection on Unobservables

The Assignment Mechanism

How to proceed: Focus on the assignment mechanism

Definition: The Assignment Mechanism

Assignment mechanism is the procedure that determines which units are selected into the treatment. The following mechanisms are available, arranged from more to less satisactory:

- Random Assignment
  - We have manipulated who gets the treatment.
- Selection on Observables
  - Selection into the treatment disappears once we take into account some factors that we think may be relevant.
- Selection on Unobservables
  - We do not know how units ended up treated or untreated.

The Experimental Ideal

Random Assignment solves the selection problem.

This is our problem:

$$E[Y_{0i}|D=1] - E[Y_{0i}|D=0] \neq 0$$

Think how experiments work: you assign some subjects to some stimulus A and some other subjects to no stimulus or some other stimulus B. You use no information in creating the groups: you remain agnostic about the assingment process –How is that possible?

- Why is this so useful?
- Does it remind you of some other process?

Randomization

Intuition: Similar to the idea of Random Sampling
Imagine you have a population of interest. You can draw as many samples from this population as you want. Each one would converge to the population average. So, you can treat them as exchangeable.
Take two samples, J and K, from the same population. Given random sampling properties:

$$E[Y_{0i}|i \in J] = E[Y_{0i}|i \in K] = E[Y_{0i}]$$

Now, replace J with $C = Control$ and K with $T = Treatment$:

$$E[Y_{0i}|i \in C] = E[Y_{0i}|i \in T] = E[Y_{0i}]$$

Now, take again our term of selection bias:

$$E[Y_{0i}|D=1] - E[Y_{0i}|D=0] = E[Y_{0i}|i \in T] - E[Y_{0i}|i \in C] = 0$$

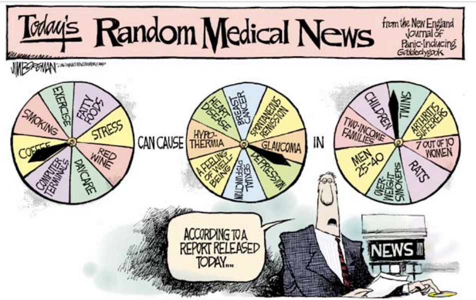
Experimental vs Observational Studies

Definition: Experimental Study

An experimental study is an empiric investigation of the effects of exposure to different treatment regimes, in which the investigator can control the assignment of treatment to subjects.

- When researchers are in control of the assignment mechanism, they do so by employing random assignment.
- This means that control and treatment units are exchangeable
  - We can substitute any unit i from the treatment group with any unit j from the control group without fearing a change in the outcome of interest. All that matters is treatment intake: $E[Y_{ij}|D_i = 1] = E[Y_{ij}|D_i = 0] = E[Y_{ij}|D_i = j]$
- This amounts to independence between potential outcomes and treatment assignment: $Y_{1i}, Y_{0i} \perp D_i$

Does Vitamin C reduce risk of Asthma?



- Observational Studies: Yes!
- Randomized Trials with Placebo Controls: No!
- Lancet 2003: comparing among individuals with the same age, gender, blood pressure, diabetes, and smoking, those with higher vitamin C levels have lower levels of obesity, lower levels of alcohol consumption, are less likely to grow up in working class, etc.

Experimental vs Observational Studies

Definition: Observational Study

An observational study is an empiric investigation of the effects of exposure to different treatment regimes, in which the investigator cannot control the assignment of treatment to subjects.

- This means that control and treatment units are not automatically exchangeable.
  - Does this mean we can only work with experimental data?
- Of course Not. This is why we add controls to our regression models. To adjust for the observed covariates.
  - What about the unobservables?
- Well, we hope they are also balanced.
  - - Is that all?
- No.

Causal Inference with Observational Studies

> *The planner of an observational study should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation.*
> (Cochran 1965)

- The main issue here is to make treatment intake as independent as possible to potential outcomes.
- Doing so, ideally, would allow you to draw causal inferences without assuming that the treatment intake has been decided only on the basis of the observables.

Three (plus one) ways to approach the experimental ideal.

- Find an instance in which the treatment has been assigned in `a haphazard way` that is as good as randomly assigned: Instrumental Variables.
- Find instance where the treatment assignment process is determined or affected by a known but not easily manipulable threshold: Regression Discontinuity Design
- Use time creatively: Difference-in-Differences

Assuming Comparability based on Observables

- Find an instance in which treatment assignment depends on a set of observable covariates: Matching

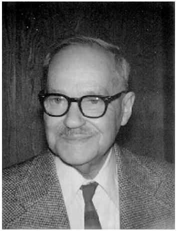
The Good, the Bad & the Ugly

Treatments, Covariates, Outcomes

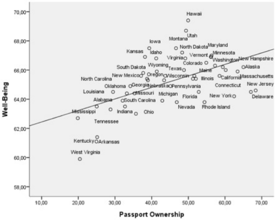
- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes
- Well-Designed Observational Study: Treatments are well-defined, control status is also well defined, clear distinction between covariates and outcomes.
- Poorer Observational Study: Hard to say when treatment began or what the treatment really is. Distinction between covariates and outcomes is blurred, so problems that arise in experiments seem to be avoided but are in fact just ignored

From Neyman to Rubin

We can only learn from experiments



Observational Research



Emphasis on the Design

Matching

Definitions

Covariates

A covariate is a variable that is predetermined with respect to treatment D_i: $X_0 = X_1$, i.e. its value does not depend on the value of D_i.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary

Outcomes

Those variables, Y, that are (possibly) not predetermined are called outcomes (for some individual i , $Y_{0,i} \neq Y_{1i}$)

In general, one should not condition on outcomes, because this may induce `post-treatment bias`.

Methods of Assignment

Three Main Properties:

- Individualistic: Whether unit i receives treatment is only dependent upon X_i; it is not dependent upon X_j, where X denotes pre-treatment characteristics of units.
- Probabilistic: No unit is *a priori* excluded from taking the treatment: Probability of being treated is strictly between zero and one for all units.
- Unconfoundedness: The treatment is free from the potential outcomes: $P(D_i = 1) \perp Y_0, Y_1$

The first two assumptions are typically assumed to hold. The main focus is on the third assumption: unconfoundedness.

Removing Bias by Conditioning

We need to satisfy: $P(D_i = 1) \perp Y_0, Y_1$

Experiments

- Randomization ensures unconfoundedness without selection on observables: $P(D_i = 1) \perp Y_0, Y_1$ which translates into:
- $E(Y_1|D = 1) = E(Y_1|D = 0)$ &
- $E(Y_0|D = 1) = E(Y_0|D = 0)$

Typical Observational Studies

- Unconfoundedness can be assumed to hold only after conditioning on a set of pre-treatment variables: $P(D_i = 1|X_i) \perp Y_0, Y_1$ which translates into:
- $E(Y_1|D = 1, X) = E(Y_1|D = 0, X)$ &
- $E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$

Conditioning on Observables

- Regression
- Subclassification
- Matching

Balancing

- All studies have a common goal: to `balance` the distributions of covariates for units which are treated and units untreated.
- They differ in how they try to achieve `balance`. In some instances estimation of causal effects happens (seemingly) simultaneously with the attempt to maximise balance on the observables (e.g. regression).
- In others, these two steps are clearly distinguished, with the design stage being the first stage in which balance is attempted and estimation follows after balance is achieved (e.g. matching).

Conditioning on Observables

- Regression
- Subclassification
- Matching

Balancing

- All studies have a common goal: to `balance` the distributions of covariates for units which are treated and units untreated.
- They differ in how they try to achieve `balance`. In some instances estimation of causal effects happens (seemingly) simultaneously with the attempt to maximise balance on the observables (e.g. regression).
- In others, these two steps are clearly distinguished, with the design stage being the stage in which balance is attempted and estimation follows after balance is achieved (e.g. matching).

A Quick Example about Sub-classification

TABLE 1

DEATH RATES PER 1,000 PERSON-YEARS

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

(Cochran 1968)

Is Pipe smoking more hazardous than cigarette smoking?

A Quick Example about Sub-classification

Table 2

Mean Ages, Years

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

(Cochran 1968)

Is Pipe smoking more hazardous than cigarette smoking?

A Quick Example about Sub-classification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution.

How can we do this?
— One solution is subclassification:

How does it work

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups
- average within age subgroup death rates using fixed weights (eg. number of cigarette smokers)

A Quick Example about Sub-classification

Adjust death rates by age:

	Death Rates 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

A Quick Example about Sub-classification

Adjust death rates by age:

	Death Rates 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?

Answer

$15 \times (11/40) + 35 \times (13/40) + 50 \times (16/40) = 35.5$

A Quick Example about Sub-classification

Adjust death rates by age:

	Death Rates 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 if they had the same age distribution as Non-smokers?

Answer

$15 \times (29/40) + 35 \times (9/40) + 50 \times (2/40) = 21.2$

A Quick Example about Sub-classification

Adjust death rates by age:

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

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

Problem

- The Curse of Dimensionality: We will see this later as well but as you add covariates, number of cells increases exponentially, hence empty cells.

Matching

The Underlying Logic

Think of matching as a way to address the missing data problem, by "imputing" missing observations for potential outcomes, using observed outcomes from units chosen on the basis of information about a set of $X's$, which—we believe—drive subjects into their treatment status.

So, if X denotes a set of pre-treatment characteristics for subjects, matching is based on the following assumption:

Unconfoundedness

$Y_1, Y_0 \perp D|X$, which is equivalent to saying that:

- within each cell defined by X treatment is random;
- the selection into treatment depends on observables X up to a random factor.

Matching: A Running Example

MPs for Sale?

> *"We are not supposed to be an assembly of gentlemen who have no interests of any kind and no association of any kind. That is ridiculous. That may apply in Heaven, but not, happily, here."*

— Winston Churchill
in the House of Commons in 1947

Research Question

What is the effect of serving in Parliament on politicians' wealth? (Eggers & Hainmueller, 2009)

The Set Up

Definition: Treatment

D_i : Indicator of treatment status for politician i

$$D_i = \begin{cases} 1 & \text{if } i \text{ was elected into Parliament} \\ 0 & \text{if } i \text{ was not elected into Parliament.} \end{cases}$$

Definition: Observed Outcome

Y_i : Observed wealth at death for politician i

Definition: Potential Outcomes

Y_{0i} and Y_{1i}: Potential Outcomes for politician i

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential wealth for } i \text{ if } i \text{ was elected in Parliament} \\ Y_{0i} & \text{Potential wealth for } i \text{ if } i \text{ was not elected in Parliament} \end{cases}$$

What we Observe

An Example with 10 candidates:

	Won?
Candidate 1	Yes
Candidate 2	Yes
Candidate 3	No
Candidate 4	No
Candidate 5	No
Candidate 6	Yes
Candidate 7	No
Candidate 8	No
Candidate 9	Yes
Candidate 10	Yes

What we Observe

An Example with 10 candidates:

	Won?
Candidate 1	Yes
Candidate 2	Yes
Candidate 3	No
Candidate 4	No
Candidate 5	No
Candidate 6	Yes
Candidate 7	No
Candidate 8	No
Candidate 9	Yes
Candidate 10	Yes

What we Observe

An Example with 10 candidates:

	D_i
Candidate 1	1
Candidate 2	1
Candidate 3	0
Candidate 4	0
Candidate 5	0
Candidate 6	1
Candidate 7	0
Candidate 8	0
Candidate 9	1
Candidate 10	1

What we Observe

An Example with 10 candidates:

	D_i
Candidate 1	1
Candidate 2	1
Candidate 3	0
Candidate 4	0
Candidate 5	0
Candidate 6	1
Candidate 7	0
Candidate 8	0
Candidate 9	1
Candidate 10	1

Quantities of Interest

$E(Y_{1i} - Y_{0i})$ — (ATE)
$E(Y_{1i} - Y_{0i} | D = 1)$ — (ATT)
$E(Y_{1i} - Y_{0i} | D = 0)$ — (ATC)

What do we Observe?

$E(Y_{1i}|D=1) - E(Y_{0i}|D=0)$

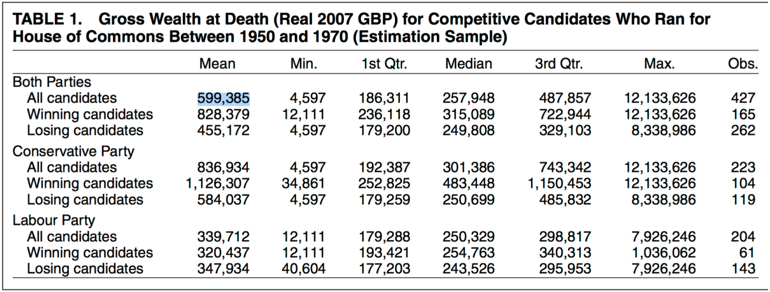


TABLE 1. Gross Wealth at Death (Real 2007 GBP) for Competitive Candidates Who Ran for House of Commons Between 1950 and 1970 (Estimation Sample)

	Mean	Min.	1st Qtr.	Median	3rd Qtr.	Max.	Obs.
Both Parties							
All candidates	599,385	4,597	186,311	257,948	487,857	12,133,626	427
Winning candidates	828,379	12,111	236,118	315,089	722,944	12,133,626	165
Losing candidates	455,172	4,597	179,200	249,808	329,103	8,338,986	262
Conservative Party							
All candidates	836,934	4,597	192,387	301,386	743,342	12,133,626	223
Winning candidates	1,126,307	34,861	252,825	483,448	1,150,453	12,133,626	104
Losing candidates	584,037	4,597	179,259	250,699	485,832	8,338,986	119
Labour Party							
All candidates	339,712	12,111	179,288	250,329	298,817	7,926,246	204
Winning candidates	320,437	12,111	193,421	254,763	340,313	1,036,062	61
Losing candidates	347,934	40,604	177,203	243,526	295,953	7,926,246	143

(Eggers & Hainmueller 2009)

Back to the Problem

$E(Y_{1i}|D=1) - E(Y_{0i}|D=0)$ leads us to the problem of selection bias that we have already seen.

Randomization would solve the problem, but you cannot randomize who gets elected and who does not.

Instead, we try to find direct comparisons: `matches` for each treated unit.

Exact Matching

An Example with 10 candidates:

	Observed Outcome: Wealth at Death	D_i	Male?
Candidate 1	855,557	1	1
Candidate 2	912,331	1	1
Candidate 3	566,271	0	1
Candidate 4	319,838	0	1
Candidate 5	612,233	0	0
Candidate 6	601,222	1	0
Candidate 7	485,709	0	1
Candidate 8	102,509	0	1
Candidate 9	991,511	1	1
Candidate 10	757,972	1	1

What do we do next?

Step 1: Split Observations According to Covariate Values

	Observed Outcome: Wealth at Death	D_i	Male?
Candidate 1	855,557	1	1
Candidate 2	912,331	1	1
Candidate 3	566,271	0	1
Candidate 4	319,838	0	1
Candidate 7	485,709	0	1
Candidate 8	102,509	0	1
Candidate 9	991,511	1	1
Candidate 10	757,972	1	1

	Observed Outcome: Wealth at Death	D_i	Male?
Candidate 5	612,233	0	0
Candidate 6	601,222	1	0

Step 2: Obtain Counterfactuals for Observed Outcome

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 1	855,557	?	1	1
Candidate 2	912,331	?	1	1
Candidate 9	991,511	?	1	1
Candidate 10	757,972	?	1	1
Candidate 3	?	566,271	0	1
Candidate 4	?	319,838	0	1
Candidate 7	?	485,709	0	1
Candidate 8	?	102,509	0	1

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 5	?	612,233	0	0
Candidate 6	601,222	?	1	0

Step 2: Obtain Counterfactuals for Observed Outcome

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 1	855,557	?	1	1
Candidate 2	912,331	?	1	1
Candidate 9	991,511	?	1	1
Candidate 10	757,972	?	1	1
Candidate 3	?	566,271	0	1
Candidate 4	?	319,838	0	1
Candidate 7	?	485,709	0	1
Candidate 8	?	102,509	0	1

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 5	601,222	612,233	0	0
Candidate 6	601,222	612,233	1	0

Imputing the Missing Outcomes

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 1	855,557	368581.75	1	1
Candidate 2	912,331		1	1
Candidate 9	991,511		1	1
Candidate 10	757,972		1	1
Candidate 3	879342.75	566,271	0	1
Candidate 4		319,838	0	1
Candidate 7		485,709	0	1
Candidate 8		102,509	0	1

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Male?
Candidate 5	601,222	612,233	0	0
Candidate 6	601,222	612,233	1	0

Step 3: Estimate Treatment Effects

ATT

- $E[Y_{1i} - Y_{0i}|D = 1, X = 1]$ for each $i = 1, 2, 9, 10$ &
- $E[Y_{1,6} - Y_{0,6}|D = 1, X = 0]$ &
- And take weighted average

ATC

- $E[Y_{1i} - Y_{0i}|D = 0, X = 1]$ for each $i = 3, 4, 7, 8$ &
- $E[Y_{1,5} - Y_{0,5}|D = 0, X = 0]$ &
- And take the weighted average.

ATE

- $E[Y_{1i} - Y_{0i}|X = 1]$ for each $i = 1, 2, 3, 4, 7, 8, 9, 10$ &
- $E[Y_{1i} - Y_{0i}|X = 0]$ for each $i = 5, 6$ &
- And take the weighted average.

Exact Matching with a Continuous Variable?

	Observed Outcome: Wealth at Death	D_i	Age
Candidate 1	855,557	1	54
Candidate 2	912,331	1	28
Candidate 3	566,271	0	73
Candidate 4	319,838	0	79
Candidate 5	612,233	0	28
Candidate 6	601,222	1	65
Candidate 7	485,709	0	47
Candidate 8	102,509	0	54
Candidate 9	991,511	1	33
Candidate 10	757,972	1	40

What do we do next?

Exact Matching with a Continuous Variable?

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Age
Candidate 1	855,557	?	1	54
Candidate 2	912,331	?	1	28
Candidate 6	601,222	?	1	65
Candidate 9	991,511	?	1	33
Candidate 10	757,972	?	1	40
Candidate 3	?	566,271	0	73
Candidate 4	?	319,838	0	79
Candidate 5	?	612,233	0	28
Candidate 7	?	485,709	0	47
Candidate 8	?	102,509	0	54

Exact Matching

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Age
Candidate 1	855,557	102,509	1	54
Candidate 2	912,331	612,223	1	28
Candidate 6	601,222	566,271	1	65
Candidate 9	991,511	612,233	1	33
Candidate 10	757,972	612,233	1	40
Candidate 3	?	566,271	0	73
Candidate 4	?	319,838	0	79
Candidate 5	?	612,233	0	28
Candidate 7	?	485,709	0	47
Candidate 8	?	102,509	0	54

Question

What is the ATT?

Exact Matching

	Potential Outcome Under Treatment	Potential Outcome Under Control	D_i	Age
Candidate 1	855,557	102,509	1	54
Candidate 2	912,331	612,223	1	28
Candidate 6	601,222	566,271	1	65
Candidate 9	991,511	612,233	1	33
Candidate 10	757,972	612,233	1	40
Candidate 3	?	566,271	0	73
Candidate 4	?	319,838	0	79
Candidate 5	?	612,233	0	28
Candidate 7	?	485,709	0	47
Candidate 8	?	102,509	0	54

Question

What is the ATT?

Answer

$1/5 \times [(855,557 - 102,509) + (912,331 - 612,223) + (601,222 - 566,271) + (991,511 - 612,233) + (757,972 - 612,233)] = 322624.8$

Adding Covariates

	D_i	Male?	Oxbridge?
Candidate 1	1	1	0
Candidate 2	1	1	1
Candidate 3	0	1	0
Candidate 4	0	1	0
Candidate 5	0	0	0
Candidate 6	1	0	1
Candidate 7	0	1	0
Candidate 8	0	1	1
Candidate 9	1	1	1
Candidate 10	1	1	0

Rearrange with respect to values of X_1

	D_i	Male?	Oxbridge?
Candidate 1	1	1	0
Candidate 2	1	1	1
Candidate 3	0	1	0
Candidate 4	0	1	0
Candidate 7	0	1	0
Candidate 8	0	1	1
Candidate 9	1	1	1
Candidate 10	1	1	0

	D_i	Male?	Oxbridge?
Candidate 5	0	0	0
Candidate 6	1	0	1

Is there a match for Candidate 6?

Is Candidate 5 good enough for any match?

So, we are left with:

	D_i	Male?	Oxbridge?
Candidate 1	1	1	0
Candidate 10	1	1	0
Candidate 3	0	1	0
Candidate 4	0	1	0
Candidate 7	0	1	0

	D_i	Male?	Oxbridge?
Candidate 2	1	1	1
Candidate 9	1	1	1
Candidate 8	0	1	1

Adding More Covariates

	D_i	Male?	Oxbridge?	Aristocrat?	Public schooling?
Candidate 1	1	1	0	0	1
Candidate 2	1	1	1	1	1
Candidate 6	1	0	1	1	1
Candidate 9	1	1	1	1	1
Candidate 10	1	1	0	1	1
Candidate 3	0	1	0	0	0
Candidate 4	0	1	0	0	0
Candidate 5	0	0	0	0	0
Candidate 7	0	1	0	0	0
Candidate 8	0	1	1	0	1

What are we looking for?

- Units differing in their D_i values while in the same time:
- Having the exact same values in all other columns ($X's$)
- Any chance?

Adding the Observed Outcome

	Gross Wealth at Death	D_i	Male?	Oxbridge?	Aristocrat?	Public schooling?
Candidate 1	855,557	1	1	0	0	1
Candidate 2	912,331	1	1	1	1	1
Candidate 6	601,222	1	0	1	1	1
Candidate 9	991,511	1	1	1	1	1
Candidate 10	757,972	1	1	0	1	1
Candidate 3	566,271	0	1	0	0	0
Candidate 4	319,838	0	1	0	0	0
Candidate 5	612,233	0	0	0	0	0
Candidate 7	485,709	0	1	0	0	0
Candidate 8	102,509	0	1	1	0	1

How would estimate the effect of D_i using regression?

- $Wealth_i = \beta_0 + \tau_M P + \beta_1 Male + \beta_2 Oxbridge + \beta_3 Aristocrat + \beta_4 PublicSchooling + u_i$
- Would you get an estimate for τ?
- $Wealth_i = 433816 + 857644 MP + 82929 Male + 104596 Oxbridge - 97585 Aristocrat - 518832 PublicSchooling + u_i$
- How would you interpret it? What does that tell you about regression?

Lesson 1: Extrapolations

- The “else equal” principle is often satisfied only through extrapolations beyond the range of the available data.
- Such extrapolations are in turn based on assumptions, which are typically untestable and ‘invisible’ within the regression framework.
- Matching, thus makes the stage of making units similar with regard to covariates more transparent.
- Imagine candidate 7 also went to a public school; thus, comparing Candidate 1 and 7 would provide an estimate of ATE.
- Even so, from $n = 10$ we would have gone down to $n = 2$.
- Again this is also the case with regression: extrapolations require attaching greater weight to most similar units.

Lesson 2: Dimensionality

- In the original study, there are more than 400 observations available.
- Yes, but many more covariates are taken into account
- As the number of covariates used to "match" units increases, it becomes exponentially more difficult to find perfect matches.
- Exact matching fails in finite samples if the dimensionality of X is large: not enough information. Far too demanding for the vast majority of research questions and data available.
- With more than one continuous variable, it is also sub-optimal (Abadie & Imbens, 2006).

The Curse of Dimensionality

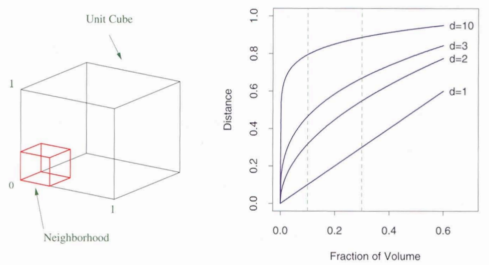


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Matching in Multidimensional Space

“Else being similar”

With many X’s and typically also with continuous X’s, estimation of ATT is based on the detection of the closest possible control unit to match every treated unit:

$\hat{\tau} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)})$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the closest value to X_i among the untreated observations.

Problem

- How to decide which control is closest?

Defining Closensess

Think of it as a distance metric. Let $X_i = (X_{i1}, X_{i2}, \dots, X_{ik})$ and $X_j = (X_{j1}, X_{j2}, \dots, X_{jk})$ be covariate vectors for i and j. We want to find ways to link rows according to their similarities in their values in each of these vectors. This is done with distance metrics.

Distance Metrics

A potentially useful distance metric:

Mahalanobis distance

$MD(X_i, X_j) = \sqrt{(X_i - X_j)'\Sigma^{-1}(X_i - X_j)}$
where Σ is the Variance-Covariance matrix. For an exact match, $MD(X_i, X_j) = 0$.

- Appropriate distance measure if each covariate has an elliptic distribution whose shape is common between treatment and control groups (Mitchell and Krzanowski 1985, 1989).
- In finite samples, Mahalanobis distance will not be optimal.

Other distance metrics can be used, e.g.

- `Stata`'s `nmatch` uses a slightly different matrix.
- Genetic matching uses a yet different metric.

The Propensity Score

Another way to reduce dimensionality: `match` on the Propensity Score

Definition

The probabity to receive treatment (also known as the selection probability) conditional on the set of pre-treatment covariates: $p(X) = P(D = 1|X)$

Identification Assumptions

1. $(Y_1, Y_0) \perp D|X$ (Selection on Observables)
2. $0 < Pr(D = 1|X) < 1$ (common support)

Propensity Score Properties

Balancing: Balancing of pre-treatment variables given the propensity score: $D \perp X|p(X)$

Unconfoundedness: If $Y_1, Y_0 \perp D|X$, then $Y_1, Y_0 \perp D|p(X)$.

How to Estimate the Propensity Score

- Regress D_i on the set of $X's$ using a logit or probit function to estimate the score.
- Take the predicted values of D_i. These predicted values represent the probability of being assigned to treatment, given X(the Propensity Score).
- Choose closest control on $p(X_i)$ (Call this the Nearest Neighbor (NN))
  - For ATT, match every X_{Ti} with the nearest X_{Ci}
  - For ATC, match every X_{Ci} with the nearest X_{Ti}
  - For ATE, match every X_{Ti} with the nearest X_{Ci} & match all units with the nearest X_{Ti}.
- Test for balance: If not satisfactory: redo with more X's or by changing matching criteria.
- Repeat until balance is satisfactory:
  Estimate PrScore → Check Balance → Re-Estimate → Check Balance

Extending the NN matching

Matching with Replacement

- If nearest control unit is already used, use it again.
- Drop unmatched controlled units

Radius or Caliper Matching

- Q: What if the NN is not a very good "clone" of the treated unit?
- A: Can use a caliper: For each treated unit find all the control units whose score differs by less than a given tolerance r chosen by the researcher. Allow for replacement of control units. If none exists, drop the unit.

Many-to-One:

- Kernel: Choose all of them and use a (weighted) average
- Caliper: All control units within the radius r from p_i are matched to unit i. Again, use (weighted) average of neighbours within r.

The Propensity Score Paradox

Matching Vs Regression

- More transparent & less model-driven. A non-parametric way (thus less assumptions involved) of ensuring balance on observables.

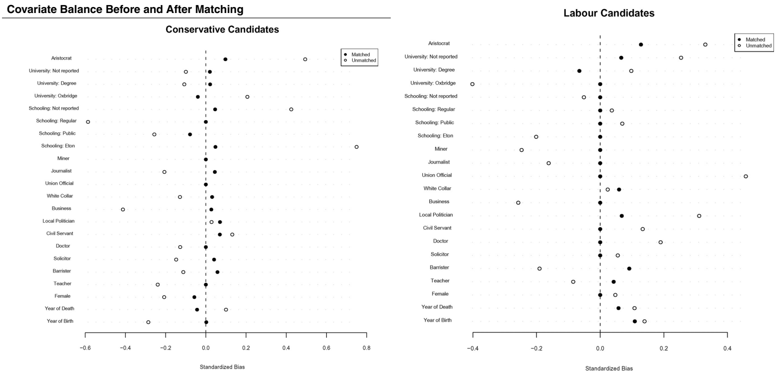
Problem

- To reduce Dimensionality use the Propensity Score, which however is essentially a regression model, thus suffers from above-mentioned problems.

The Key Difference

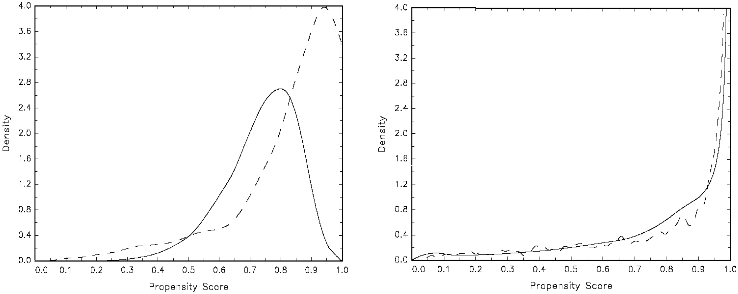
- The Propensity Score is an `ancillary statistic`: You can attempt as many estimations as you want, without losing degrees of freedom.
- This is because you never work with the outcome. You remain completely agnostic about the outcome.

How to Check for Balance: Standardised Bias



(Eggers & Hainmueller 2009)

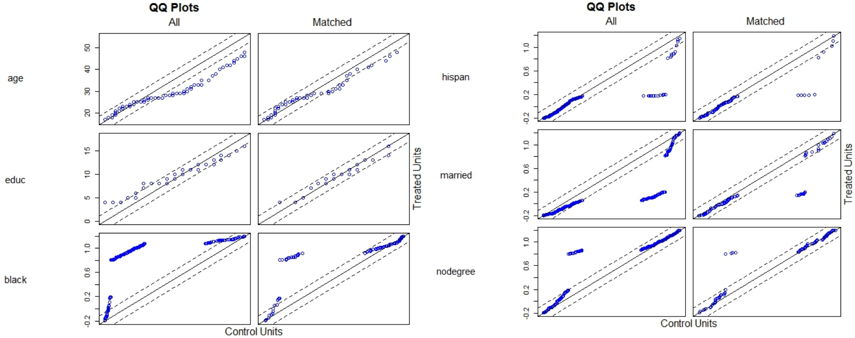
How to Check for Balance: Compare Distributions of Important X's



Non-balanced Matching **Balanced Matching**

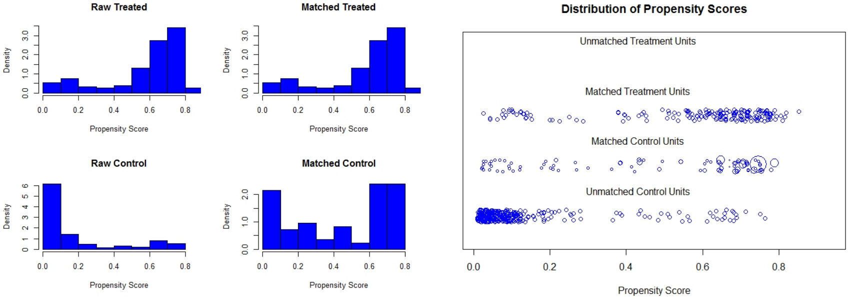
(Dashed curve is treatment, solid curve is control, from Shaikh and coworkers, 2005,)

How to Check for Balance: QQ-Plots



(Packmohr)

How to Check for Balance: Histograms & Scatterplots



(Packmohr)

Balance Tests

Measuring Balance

T-tests

- T-tests for the difference of means. Well-known but the null is the groups are the same, so need to adjust significance levels. p-values conventionally understood to signal balance (e.g., 0.10) are often too low to produce reliable estimates.

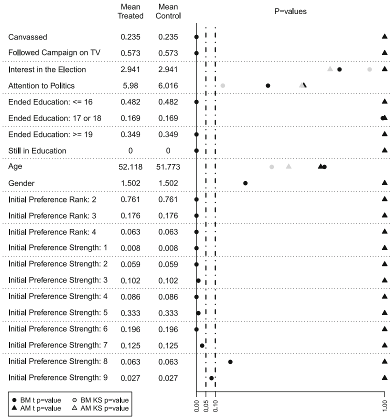
Equivalence Tests

- Use the null that the groups are different. Power of detecting similarity increases as sample size increases.
- Less well known. Make the tests to assess the covariates different from the tests to assess the outcome.

Kolmogorov-Smirnov distributional tests

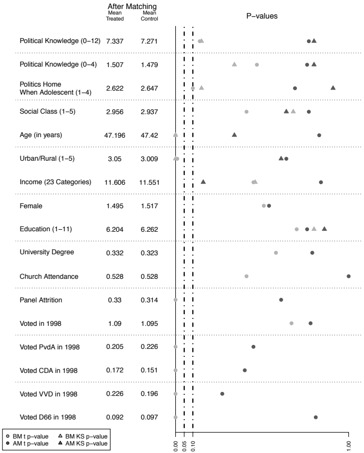
- The test statistic is the maximum distance between the empirical CDFs of the treatment and control distributions.
- Good to detect imbalance beyond simply means comparisons.

How to Check for Balance: T-Test & KS-test



	Mean Treated	Mean Control
Canvassed	0.235	0.235
Followed Campaign on TV	0.573	0.573
Interest in the Election	2.941	2.941
Attention to Politics	5.98	6.016
Ended Education: <= 16	0.482	0.482
Ended Education: 17 or 18	0.169	0.169
Ended Education: >= 19	0.349	0.349
Still in Education	0	0
Age	52.118	51.773
Gender	1.502	1.502
Initial Preference Rank: 2	0.761	0.761
Initial Preference Rank: 3	0.176	0.176
Initial Preference Rank: 4	0.063	0.063
Initial Preference Strength: 1	0.008	0.008
Initial Preference Strength: 2	0.059	0.059
Initial Preference Strength: 3	0.102	0.102
Initial Preference Strength: 4	0.086	0.086
Initial Preference Strength: 5	0.333	0.333
Initial Preference Strength: 6	0.196	0.196
Initial Preference Strength: 7	0.125	0.125
Initial Preference Strength: 8	0.063	0.063
Initial Preference Strength: 9	0.027	0.027

Fig. 3 Improved balance from genetic matching in the case of treatment versus control group 1

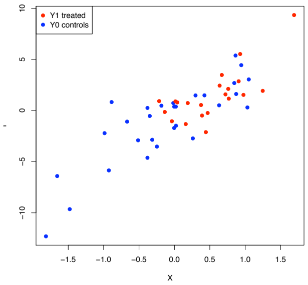


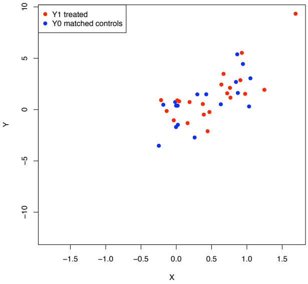
	After Matching	
	Mean Treated	Mean Control
Political Knowledge (0–12)	7.337	7.271
Political Knowledge (0–4)	1.507	1.479
Politics Home When Adolescent (1–4)	2.622	2.647
Social Class (1–5)	2.956	2.937
Age (in years)	47.196	47.42
Urban/Rural (1–5)	3.05	3.009
Income (23 Categories)	11.606	11.551
Female	1.495	1.517
Education (1–11)	6.204	6.262
University Degree	0.332	0.323
Church Attendance	0.528	0.528
Panel Attrition	0.33	0.314
Voted in 1998	1.09	1.095
Voted PvdA in 1998	0.205	0.226
Voted CDA in 1998	0.172	0.151
Voted VVD in 1998	0.226	0.196
Voted D66 in 1998	0.092	0.097

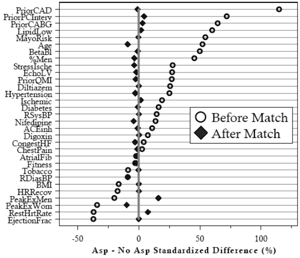
(Bølstad et al. 2013)

(Dinas et al. 2015)

Before and After Matching







Covariate Balance for Aspirin Study (Love, 2004)

Equal Percent Bias Reduction (EPBR)

- If your controls are all normally distributed (more precisely, follow and elliptic distribution) and your sample size is large enough, then matching on Mahalanobis distance has the Equal Percent Bias Reduction (EPBR) property.
- This means that matching will not make balance on any covariate worse.
- It does not mean that your estimated treatment effect will be less biased. It could be more biased if you do not have the right X's.

Genetic Matching

The problem is where to draw the line. In principle, given assumptions, EPBC tells us that we can infinitely reduce bias. All we need to do is iterate the procedure with different matching algorithms until we are sure we found the best match. But this is often too time consuming and sometimes requires a seemingly infinite number of comparisons.

A Solution: Genetic Matching

- A combination of propensity score and Mahalanobis matching.
- An automated iteration process that bring optimal matches, given X, in terms of bias reduction.
- Greatly improved when including the $p(Xi)$ as a covariate.
- Follows Rubin and Rosenbaum in that in addition to propensity score matching, it matches on individual covariates by minimizing the MD of X to obtain balance on X.
- Has been shown to recover experimental benchmarks

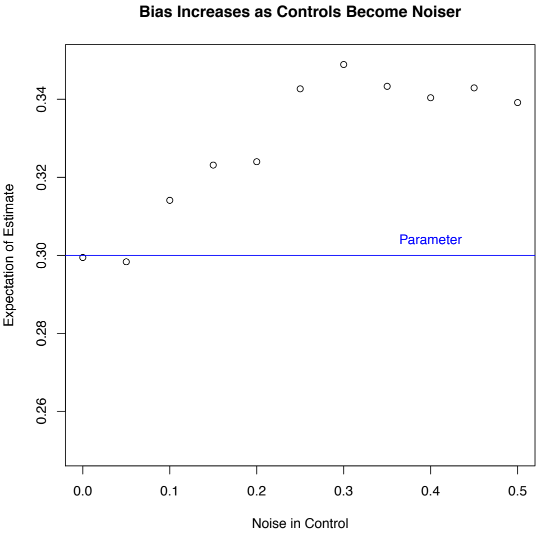
Facts about Matching

- Matching does not mean that you will get better balance on the covariates that you match on.
- Getting better balance on your controls does not mean that your bias will decrease
- Matching is susceptible to attenuation bias

Example of Better Balance Increasing Bias

- Imagine that there is a summer program that high school students can take. Enrollees tend to be (1) harder working and (2) younger.
- At the end of the following school year, there are 20 awards given to students. Awards tend to be given to students who are (1) harder working and (2) older.
- We want to estimate the effect of the summer program on the likelihood of winning an award. Imagine there is no effect. We match on age.
- Before matching, the treatment group will have younger students and harder working students, so the bias will partly cancel.
- After matching on age, the treatment group will tend to have harder workers, but not younger students. So the hard work bias will no longer be mitigated by the age bias.

Example of Attenuation Bias



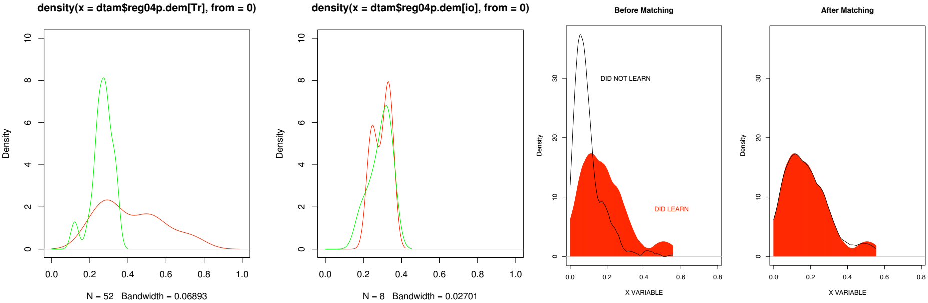
Potential Problems with Matching

What can go wrong?

1. There might not be support in the data.
2. There might be support, but you chose the wrong X's.
3. You might have support and the right X's, but your formula for the propensity score is wrong (if you are doing propensity score matching) or your controls do not each follow an elliptic distribution (if you are doing Mahalanobis distance matching).
4. Everything else worked, but there was noise in your controls.
5. Everything worked perfectly, but people will still be skeptical or think that you p-hacked.

Genetic Matching solves Problem (3) only.

Examples from Genetic Matching

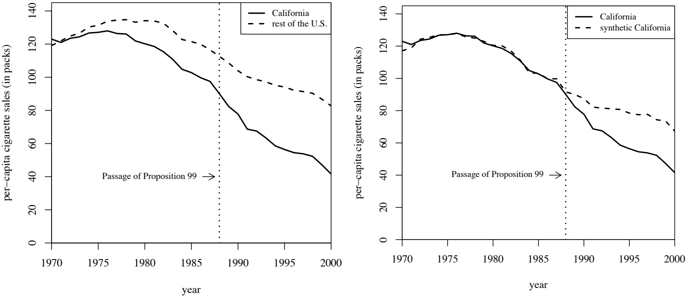


Toolbox

- Always establish balance before you even look at the Y
- Look for balance not only at the characteristics included in matching but higher polynomials and on other covariates. Balance should extend beyond X if X is correctly specified.
- Do not simply think of matching as an alternative or final resort when design-based identification is not provided. Conversely, use it when there is some design that allows you to make the conditional-on-observables assumption more credibly.

Numerous Extensions

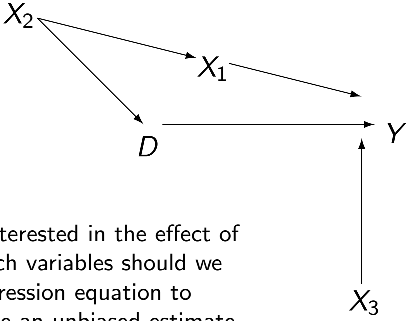
- E.g. Entropy Balancing (Hainmueller)
- Synthetic Control (Abadie) (combining re-weighting and time-variant data)
  - Particularly useful for small-n comparative studies (using countries or regions as units of analysis)
  - Eg. Two independent studies presenting in the same panel the same New Zealand case using synthetic control.
  - If there are good data, can be powerful:



Instrumental Variables

Instrumental Variables

An Introduction to Causal Diagrams (Judea Pearl)

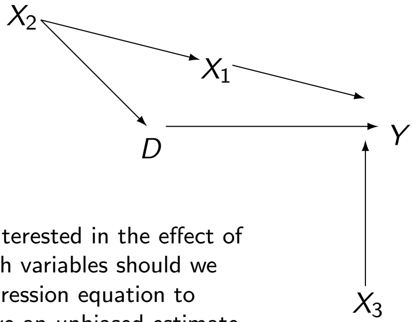


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

$$Y = \alpha + \tau D + \beta_1 X_1 + \beta_2 X_2 + u_i$$

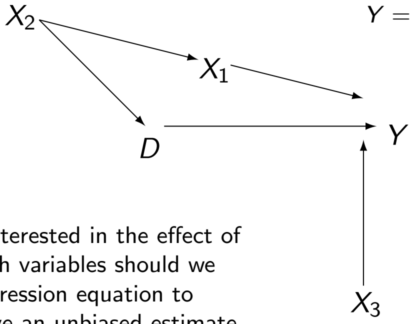


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

$$Y = \alpha + \tau D + \beta_1 X_1 + \beta_2 X_2 + u_i$$
$$Y = \alpha + \tau' D + \beta_1 X_1 + u_i$$

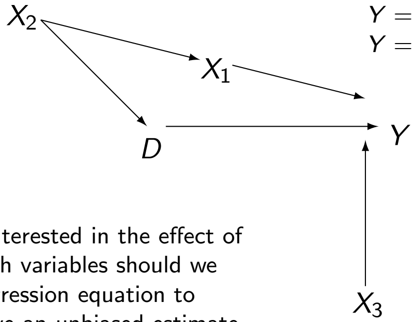


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

$$Y = \alpha + \tau D + \beta_1 X_1 + \beta_2 X_2 + u_i$$
$$Y = \alpha + \tau' D + \beta_1 X_1 + u_i$$
$$Y = \alpha + \tau'' D + \beta_1 X_2 + u_i$$

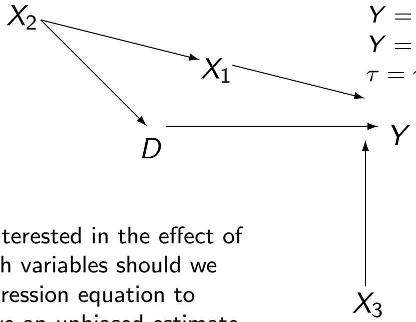


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

$$Y = \alpha + \tau D + \beta_1 X_1 + \beta_2 X_2 + u_i$$
$$Y = \alpha + \tau' D + \beta_1 X_1 + u_i$$
$$Y = \alpha + \tau'' D + \beta_1 X_2 + u_i$$
$$\tau = \tau' = \tau''$$

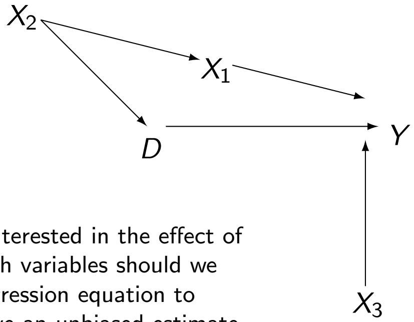


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

What about X_3?

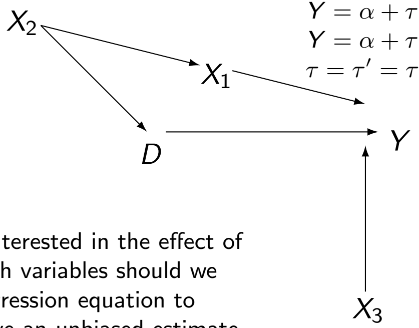


Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

An Introduction to Causal Diagrams

$$
\begin{aligned}
&Y = \alpha + \tau D + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_i \\
&Y = \alpha + \tau' D + \beta_1 X_1 + \beta_2 X_3 + u_i \\
&Y = \alpha + \tau'' D + \beta_1 X_2 + \beta_2 X_3 + u_i \\
&Y = \alpha + \tau''' D + \beta_1 X_1 + u_i \\
&Y = \alpha + \tau'''' D + \beta_1 X_2 + u_i \\
&\tau = \tau' = \tau'' = \tau''' = \tau''''
\end{aligned}
$$



Imagine we are interested in the effect of D on Y. For which variables should we control in our regression equation to make sure we have an unbiased estimate of the ATE of D on Y? You can assume this is the correct model and that all Xs are pre-treatment.

Instrumental Variables

Our Running Example

We are interested in whether watching Western TV in communist regimes affected attitudes towards the political regime.

- Let's denote as $D = 1$ the treatment state, i.e. watching Western TV, and as $D = 0$ the control state, i.e. not watching Western TV.
- By now we know that Y_1 denotes support for communism if unit i watches Western TV and Y_0 if i does not watch Western TV.

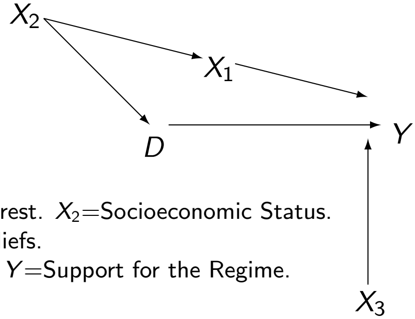
We also know the problem with the following regression equation:

$$Y_i = \alpha + \tau D_i + u_i$$

People who watched western TV would probably differ in their attitudes towards Communism from people who did not watch western TV even if they were not watching western TV.

Instrumental Variables

Putting this idea into context

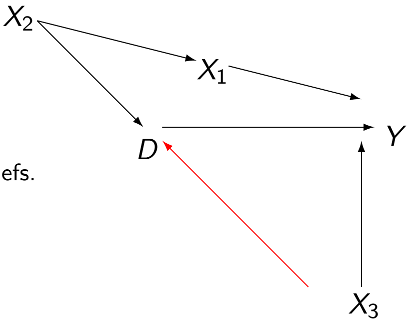


X_1=Political Interest. X_2=Socioeconomic Status.
X_3=Religious Beliefs.
D=Western TV. Y=Support for the Regime.

Back to Reality

Do we really know this is the correct model?

- Eg. How to justify the absence of this arrow?

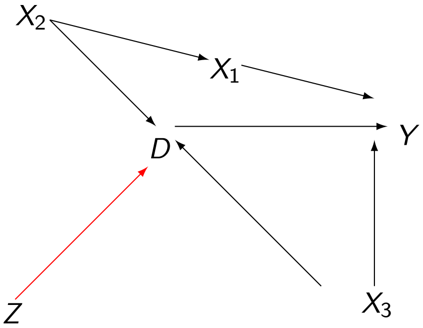


X_3=Religious Beliefs.

Instrumental Variables

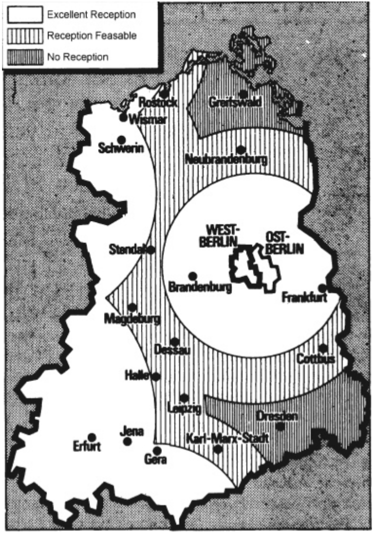
What if...?

$Z =$? What could Z be? Why would it be helpful?



Z=Access to Western TV

Hainmueller
and Kern (2009)



The Intuition

Why is this variation helpful?

The whole logic is based on the idea that we need to go back one step: D is itself a random variable, which means that even if units select their values in this state, there may be some other [prior] state that predicts $D = d$ and can be regarded as having been randomly assigned. If we find such a variable, we call it an instrumental variable, or simply an instrument. Here, we will refer to instruments as Z.

- We can treat D_i the same way as Y_i:
- D_0:Treatment Status under the state $Z = 0$
- D_1:Treatment Status under the state $Z = 1$
  - In this example, D_0 refers to the probability of watching TV for someone who lived in an area with only limited or no access to western TV (e.g. Dresden) & D_1 refers to the probability of watching TV if you have access to it.
- Z does not need to deterministically predict D.
- We are fine with finding people who live in Dresden watching Western TV and/or people not residing in Dresden not watching Western German TV, insofar as the overall proportion of those watching TV in $Z = 1$ is higher than in $Z = 0$.

The Reduced Form

What is the regression equation now?

$$Y_i = \alpha + \theta Z + u_i$$

- What is the quantity captured by θ?
- $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$
- Have we seen this quantity before?

The Intent-To-Treat Effect

ITT $\equiv$ Support towards the regime conditional on access to Western TV (we remain agnostic with regard to whether people actually watched Western TV).

Compliance

A Precondition for the IV estimation

$$D_i = \alpha + \lambda Z + u_i$$

- λ should be $\neq 0$.
- For this to happen, people in Dresden should actually watch on average W. TV less than other people
- Think of it as a lab experiment
- Imagine we give to some people a treatment but we do not know if they comply to the assignment
  - Some people assigned to take the treatment (not watch TV) may actually take it: compliers
  - Some people assigned not to take the treatment (watch TV) may not take it: non-compliers

Classification, Z and D

		$Z = 0$	
		$D_0 = 0$	$D_0 = 1$
$Z = 1$	$D_1 = 1$	Complier	Always Taker
	$D_1 = 0$	Never Taker	Defier

Classification, Z and D

		$Z = 0$	
		$D_0 = 0$	$D_0 = 1$
$Z = 1$	$D_1 = 1$	Complier	Always Taker
	$D_1 = 0$	Never Taker	Defier

Classification, Z and D

		$Z = 0$	
		$D_0 = 0$	$D_0 = 1$
$Z = 1$	$D_1 = 1$	Watched TV only if not in Dresden	Watched TV
	$D_1 = 0$	Did not Watch TV	Watched TV only if in Dresden

Potential Outcomes Model for Instrumental Variables

Following Angrist, Imbens and Rubin (1996), we can define:

Definition

- Compliers: $D_1 > D_0$ $(D_0 = 0 \ \& \ D_1 = 1)$
- Always-Takers: $D_1 = D_0 = 1$
- Never-Takers: $D_1 = D_0 = 0$
- Defiers: $D_1 < D_0(D_0 = 1 \ \& \ D_1 = 0)$

Problem

Only one of the potential treatment indicators (D_0, D_1) is observed, so we cannot identify which group any particular individual belongs to.

Who are the compliers? (Examples...)

Estimation

The Wald Estimator

Imagine that we only have compliers, i.e. people who only watched TV if they were not in Dresden, so that $D_{1i} = 1$ and $D_{0i} = 0 \; \forall \; i \in$ N. Would the following equation tell us something about the effect of W.TV on Communist support?

$$Y_i = \alpha + \theta Z + u_i$$

- If $D_i = Z_i$ then $\theta = \tau = E[Y|Western TV] - E[Y|NoWestern TV]$.
- But we know that this is not true. Moreover, we know how many watched TV among those in Dresden and those not in Dresden. So, we could adjust for this. How?

The Wald Estimator

$\tau_{IV} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$. Does this remind you of anything? Have a look at the nominator and the denominator separately.

The Wald Estimator

The Nominator

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

- This is θ : $Y_i = \alpha + \theta Z + u_i$
- What is θ? (What is a regression coefficient?)

$$\theta = \frac{Cov(Z, Y)}{Var(Z)}$$

The Denominator

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$

- This is λ : $D_i = \alpha + \lambda Z + u_i$
- What is λ? (What is a regression coefficient?)

$$\lambda = \frac{Cov(Z, D)}{Var(Z)}$$

The Wald Estimator

Taken together:

$$\text{Wald Estimator} = \frac{Cov(Z, Y)}{Cov(Z, D)}$$

Does this give us the ATE or the ATT?

The IV Assumptions

1. Ignorability

- Think about Dresden and Western TV
- What qualities should Dresden have to be regarded a good instrument for Y?
- Hint: We started all this because watching TV was not randomly assigned.

Ignorability: Definition

The Instrument is independent of potential outcomes and treatments:

$$[Y_0, Y_1, D_0, D_1 \perp Z]$$

Do you think this is satisfied in our example?

The IV Assumptions

2. First Stage

- Why did we choose Dresden and not any other region?

First Stage: Definition

The probability of being treated needs to be different between those assigned the treatment and those not assigned the treatment: $E(D_1 - D_0)) \neq 0$

This is quite easy to test. How?

$$D_i = \alpha + \lambda Z + u_i$$

All we ask is that $\lambda \neq 0$

The IV Assumptions

3. Exclusion
   - The instrument does not affect the outcome in any other way than through the treatment
   - In our case: People in Dresden should be otherwise exchangeable with people elsewhere.

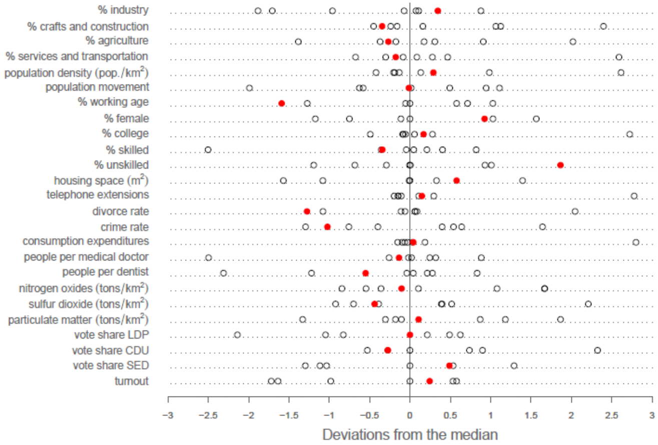
Exclusion: Definition

Define as Y_{dz} the potential outcome of unit i with $Z = z$ and $D = d$. So, for instance, Y_{10} is an `always taker`. Then, exclusion is:
$P(Y_{d1} = Y_{d0}) = 1$.

Do you think this is satisfied in our example?

The IV Assumptions

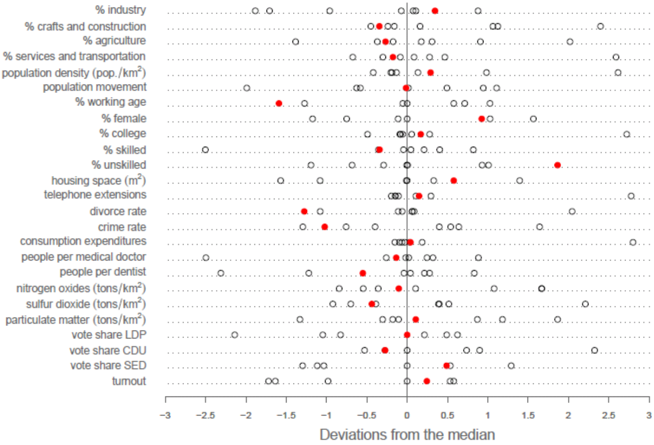
Exclusion vs Ignorability



Is this evidence for exclusion or for ignorability?

The IV Assumptions

Exlusion vs Ignorability



Exclusion

The IV Assumptions

4. Monotonicity

- No one is encouraged to watch TV because of not having access to it.
- In our typology, monotonicity assures there are no defiers and that there are at least some compliers

Monotonicity: Definition

No one does the opposite to his/her assignment, no matter what the assignment is: $D_{1i} \geq D_{0i} \; \forall \; i$

How the Assumptions Work

1. Independence/Ignorability
   - Allows us to identify the causal effect for each group: If Z is randomly assigned, both the ITT and first stage are causally identified.

3. Exclusion
   - Exclusion restrictions ensure the causal effect is zero for always- and never-takers. It is only non-zero for compliers and defiers. Allows to attribute correlation between Z and Y to the effect of D alone; assumption is not testable.

4. Monotonicity
   - Ensures there are no defiers and there exists at least one complier

2. First stage
   - Ensures estimation is possible, i.e. there is a link from Z to D.

The IV Estimand=LATE

If Assumptions 1-4 hold: the IV estimand is the Local Average Treatment Effect

LATE: Definition

If Assumptions 1-4 hold, then:

$$\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}].$$

When 1-4 are met, we can get the LATE by looking at how well Z

predicts Y and adjusting for the fact that Z is not the same as D

How informative is the LATE?

LATE

Better LATE than Nothing

- The LATE refers to units whose value in D depends on their value in Z. This means, these people can be manipulated by the instrument, i.e. the instrument can determine their status, they react to it they are responsive

- The LATE cannot say anything about peple not affected by the assignment to the treatment: those who do not change treatment status as a result of their being assigned to do so.

- The "L" part of the LATE does not have actual geographical meaning. It simply denotes the units that are responsive to the treatment assignment mechanism.

Adding Covariates

In many instances all these four assumptions do not hold just like that. They hold only if we include controls. For example, you need to include all indicators shown in the previous graph where Dresden was significantly different from other regions. Thus:

- Ignorability $[Y_0, Y_1, D_0, D_1 \perp Z|X]$
- Exclusion $P(Y_{1d} = Y_{0d}|X) = 1$
- First Stage $P(D_1 = 1|X) \leq P(D_0 = 1|X)$
- Monotonicity $P(D_1 \geq D_0|X) = 1$

How to use the IV-Estimand in practice in the presence of controls?

Two-Stage-Least-Squares Estimator

Consider an outcome Y, a treatment D that is not randomly assigned,k covariates X, and an instrument Z, for which Assumptions I-IV are met.

Estimation

First Stage

Fit the following equation into the data:
$D_i = \alpha + \theta Z_i + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + v_i$, and obtain the $\widehat{D}$.

Second Stage

Regress Y on $\hat{D}$ and $X's$:
$Y_i = \alpha + \tau\widehat{D} + +\beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$
τ provides the LATE.

Two-Stage-Least-Squares Estimator

Toolkit

- Regress your D on your Z plus all your pre-treatment controls that you think help you satisfy ignorability and exclusion (or even first stage).
- RULE: Do this with OLS no matter what your D_i is.
- Take the predicted values and use them the way you would use D. There are your $\widehat{D}$.
- Regress Y on $\widehat{D}$ plus all your pre-treatment controls that you included in the first stage. Nothing less and nothing more.
- You need to adjust the standard errors for the first-stage predictions
- Any statistical software does this very easily and gives you the correct standard errors.
- This is what the Two-Stage-Least-Squares estimator is: The IV estimand with controls.

Extensions

Bias of 2SLS and the LARF estimator

When including covariates 2SLS can be biased, even if everything goes well. What to Do:

- Nothing
- Use a more difficult-to-implement estimator (with problematic finite-sample properties): The Local Average Response Function estimator, proposed by Abadie
- You need to use, R, Matlab or an ado file in STATA.
- Corrects the problem
- Most often trivial differences between 2SLS and LARF

Examples

Angrist

- Question: Does being a veteran pay off financially?
- Problem: People choose whether to go to the army
- Solution: Using the Lottery Draft

Turnout

- Question: Is there a cost in voting?
- Problem: People choose whether to vote
- Solution: Using rainfalls on the day of the election

Acemoglu

- Question: Do institutions matter for growth?
- Problem: Growth and institutions not randomly assigned, e.g. Putnam etc.
- Solution: Using variation in mortality rates in colonies, 1700-1900s.

The Regression Discontinuity Design

Regression Discontinuity Design

Background

- RDD is a fairly old idea (Thistlethwaite and Campbell, 1960) but this design experienced a renaissance in recent years.
- Assignment to treatment and control is not random, but we know the assignment rule determining how people are assigned or select into treatment
- Widely applicable in a rule-based world (adiminstrative programs, elections, etc.)
- High internal validity: In their validation study Shadish, Clark, and Steiner (2010) identify RDD as one of the few observational study designs that can accurately reproduce experimental benchmarks

What is it?

Definition

- Imagine a binary treatment D that is completely determined by the value of a predictor X_i being on either side of a fixed cutoff point c:

$$D_i = 1(X > C) \quad so = \begin{cases} D_1 = 1 & \text{if } X_i > c \\ D_0 = 0 & \text{if } X_i < c \end{cases}$$

- X, called the forcing variable, may be correlated with the outcomes Y so comparing treated and untreated units does not provide causal estimates
- Design arises often from administrative decisions, where the allocation of units to a program is partly limited for reasons of resource constraints, and sharp rules rather than discretion by administrators is used for allocation.

An Example

Thistlethwaite and Campbell, 1960

- Question: What is the effect of scholarship on later income?
- Problem: Scholarships go to the best students
- Soultion: Scholarships are given on the basies of whether or not a student's test score exceeds some threshold c

Why is this useful?

- Treatment D is scholarship
- Forcing Variable X is SAT score with cuttoff c
- Outcome Y is subsequent earnings
- Y_0 denotes potential earnings without the scholarship. Y_1 denotes potential earnings with the scholarship.
- Y_1 and Y_0 are correlated with X: on average, students with higher SAT scores obtain higher earnings

Two types of RDD

Sharp Regression Discontinuity Design

- Treatment is deterministically assigned conditional on whether the unit is below or above the treatment
- Examples: Scholarships, taxation schemes etc.

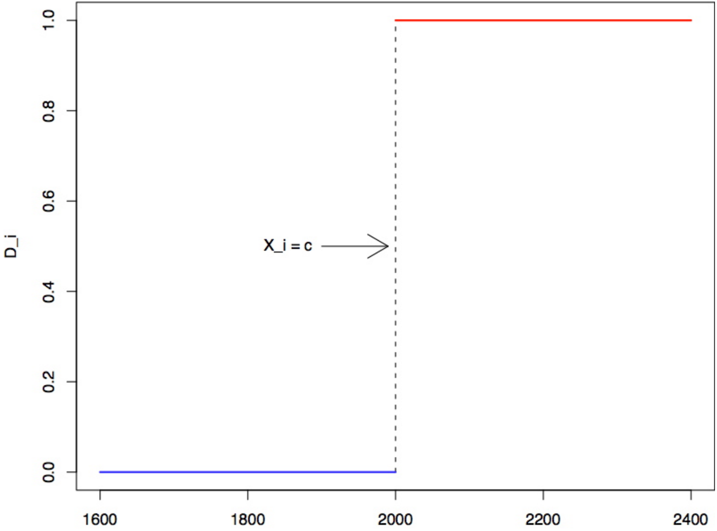
Fuzzy Regression Discontinuity Design

- Treatment is probabilistically assigned conditional on whether the unit is below or above the treatment
- Examples: Electoral thresholds, class size etc.

We will mainly focus on the sharp RDD

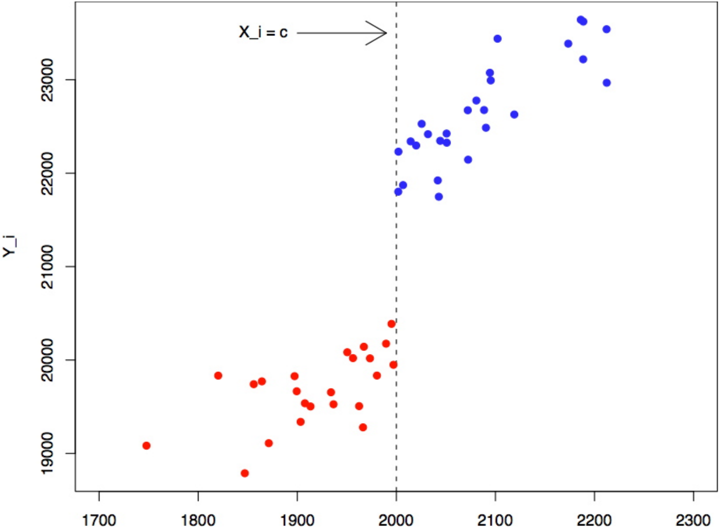
Sharp RDD: An Illustration

How the treatment is assigned



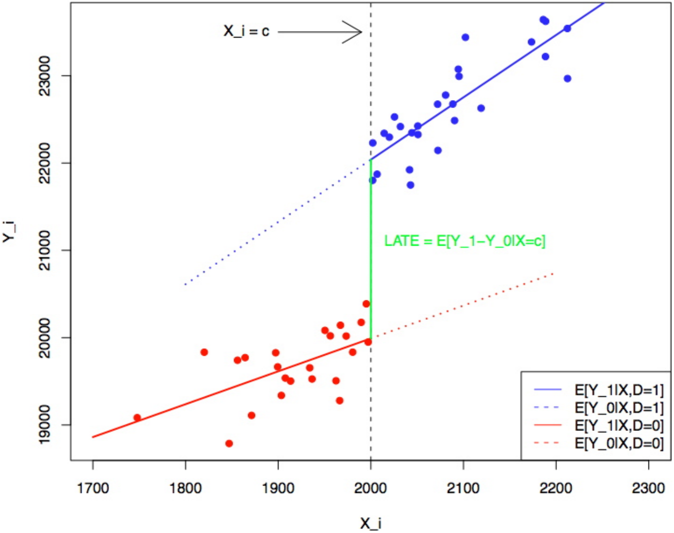
An Illustration

The Effect on the Outcome



An Illustration

The Effect on the Outcome



Identification

It practically necessitates one (important) assumption

Continuity: Definition

$E[Y_1|X, D]$ and $E[Y_0|X, D]$ are continuous in X around the threshold $X = c$ (to compensate for lack of common support)

Continuous density of the forcing variable

If this assumption holds, then

$\tau_{SRDD} = E[Y_1 - Y_0|X = c] = E[Y_1|X = c] - E[Y_0|X = c]$

Without further assumptions τ_{SRDD} is only identified at the threshold

Continuity: What it means in Practice

The density in the area of the threshold is smooth. If GPA is between 0 and 20, and if scholarships are given above 18.5, whether one will have 18.4 or 18.6 is practically coincidental.

An Example: Eggers 2010

Duverger's Law

> *A majority vote on one ballot is conducive to a two-party system. Proportional representation is conducive to a multiparty system.*

- We expect the number of parties to increase when going from a majority to a proportional electoral system

The Problem

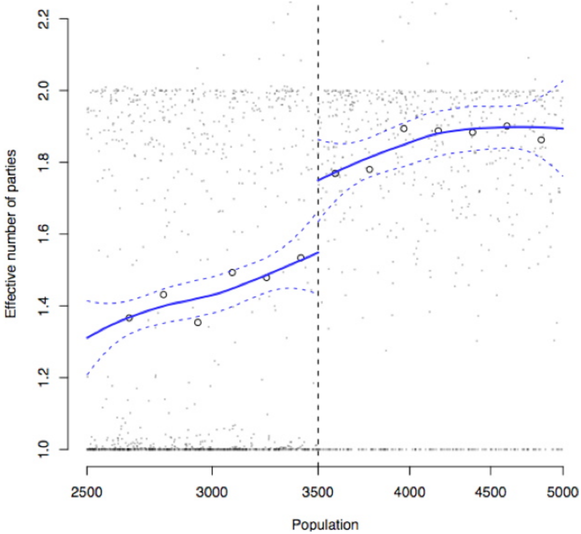
Countries decide what electoral system they have and this is in part determined by their social, linguistic etc. divisions which also give room to parties, eg. the Netherlands etc.

The Solution

In French municipalities, the electoral rule used to elect the municipal council depends on the city's population:
Cities with fewer than 3,500 people elect their councils with plurality
Cities with a population of 3,500 or more use a PR rule

An Illustration

Duverger's Law



Another Example: Lee 2006

Party Incumbent Advantage

- An ongoing debate about whether incumbents in the American Congress have an advantage of reelection
- Discussion motivated by the lack of variation in winning patterns

The Problem

Do they win because they have won already? And then why did they win at first place? Maybe they are just more efficient?

The Solution

With so many elections and states you can just see whether those who just won at t_0 did better in election t_1 than those who just (marginally) lost.

The Set-Up

Party Incumbent Advantage

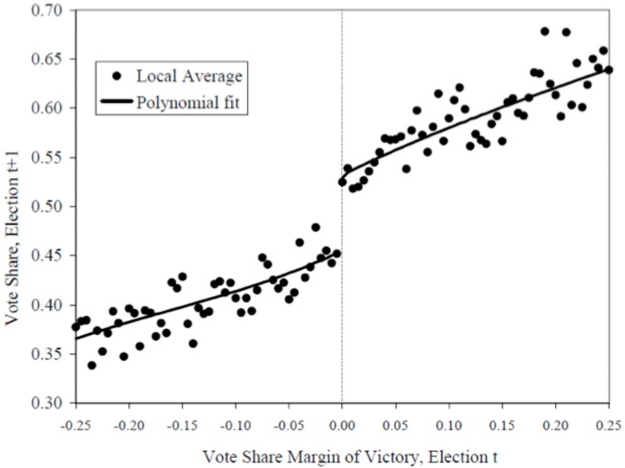
- What is the effect of incumbency status on vote shares?
- Let i indicate congressional districts, j indicate parties and t indicate elections, d indicate incumbency status.
- V_{ditj} is the vote share of j in i at t as incumbent $d = 1$ or non-incumbent $d = 0$
- Party incumbency effect: $V_{1itj} - V_{0itj}$
- Margin of victory for party j: $Z_{itj} = V_{itj} - V_{itk}$ where k indicates the strongest opposition party.
- Party incumbency status is then assigned as:

$$D_{ij,t+1} = 1[Z_{itz} > 0] \quad so \quad D_i = \begin{cases} D_{ij,t+1} = 1 & \text{if} \quad Z_{itj} > 0 \\ D_{ij,t+1} = 0 & \text{if} \quad Z_{itj} < 0 \end{cases}$$

An Illustration

Incumbency Advantage

Figure IVa: Democrat Party's Vote Share in Election t+1, by Margin of Victory in Election t: local averages and parametric fit



Estimation

The Intuition

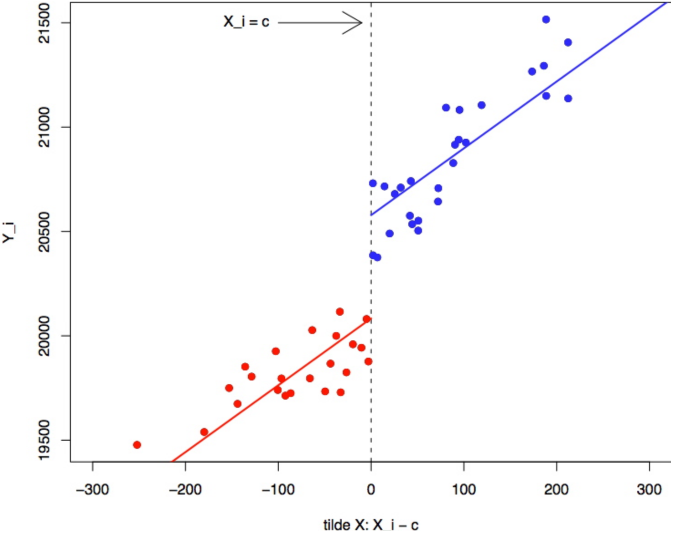
- We are interested in minimally extrapolating at the point of the discontinuity
- To do this we need an interaction term

$$Y_i = \alpha + \beta_1(X_i - c) + \tau D_i + \beta_2(X_i - c) \cdot D_i + u_i$$

- We are interested in τ. Why?
- Because it is the effect of D_i evaluated at the point of the discontinuity
- Is this enough?

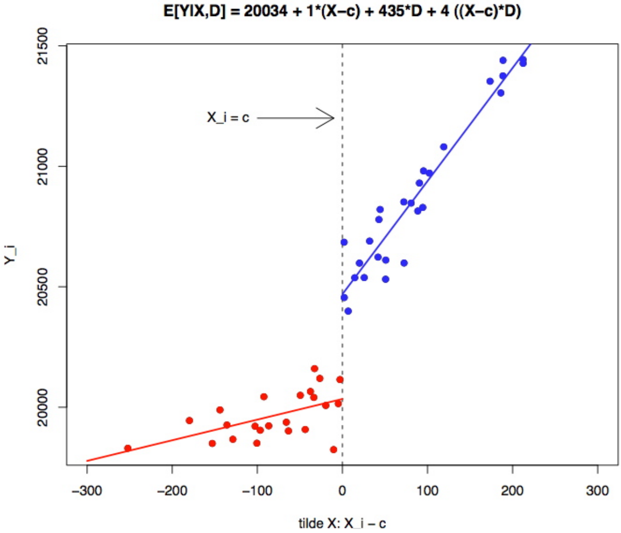
An Illustration

Linear Slopes



An Illustration

Different Slopes



Why does the RD solve the selection problem?

Back to our selection bias formula

Assume there is a homogenous treatment effect, τ and a confounder X:

$$Y_i = \alpha + \tau D + \beta X + u_i$$

- Then:

$$\begin{aligned} E[Y_i|D_i=1]-E[Y_i|D_i=0] &= \\ [\alpha+\tau+\beta E(X_i|D_i=1)+E[u_i|D_i=1]- \\ -[\alpha+0+\beta E[X_i|D_i=0]+E[u_i|D_i=0] &= \\ \tau+\beta[E(X_i|D_i=1)-E(X_i|D_i=0)] &= \\ \text{true effect} + \text{Bias} \end{aligned}$$

Since we assume that X varies smoothly with respect to Z (i.e. there are no jumps at the treatment threshold z_0):

$$\begin{aligned} E[Y_i|D_i=1]-E[Y_i|D_i=0] &= \\ = \tau+\beta[(x^*+\delta)-x^*] &= \\ = \tau+\beta^*\cdot\delta &\approx \\ &\approx \tau \end{aligned}$$

Estimation

Global Polynomial Function

- Better to be on the safe side
- Avoid Non-linearities falsely taken for discontinuities
- Add many polynomials

$$
\begin{aligned}
Y_i = \alpha + \beta_1(X_i - c) + \tau D_i + \beta_2(X_i - c) \cdot D_i \\
+\beta_3(X_i - c)^2 + \beta_4(X_i - c)^2 \cdot D_i \\
+\beta_5(X_i - c)^3 + \beta_6(X_i - c)^3 \cdot D_i \\
+\beta_7(X_i - c)^4 + \beta_8(X_i - c)^4 \cdot D_i \\
+u_i
\end{aligned}
$$

- We are still interested in τ.

The equation can be augmented by including control variables

Estimation

Local Linear Regression

- We are interested in minimally extrapolating at the point of the discontinuity
- To do this we need an interaction term
$$Y_i = \alpha + \beta 1(X_i - c) + \tau D_i + \beta 2(X_i - c) \cdot D_i + u_i$$
- We are interested in τ. What has changed?
- This regression is estimated linearly in the area around the threshold. Not in all the range of the forcing variable. A kernel regression applied locally around the cutoff point
- Make sure your software uses a triangular kernel

Again, the equation can be augmented by including control variables

Bandwidth

Cross-validation

- Bandwidth: The area below and above the threhold
- How far to open the window?
  - Try different windows and make sure you get effects even when you have closed it quite a lot (despite increasing uncertainty)

Optimal Bandwidth

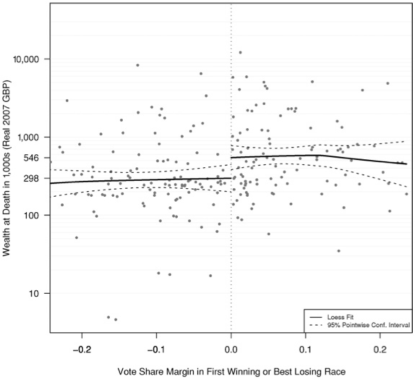
- A data-driven, flexible algorithm to optimally choose the bandwidth
- Proposed by Kalyanaraman and Imbens
- Available in Stata and R.
- In there, use a local regression to estimate the effects

An Application

Do MPs make more money from politics?

FIGURE 4. Regression Discontinuity Design: Effect of Serving in House of Commons on Wealth at Death

Conservative Candidates



Eggers and Hainmueller 2009: MPs for Sale?

Concerns

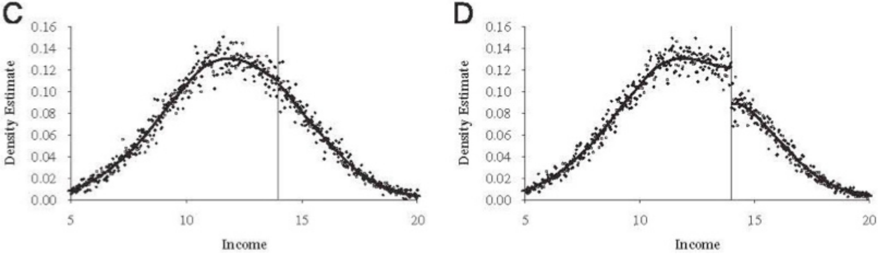
1. Sorting

- Subjects are aware of these interventions and they manipulate their position just below or above the treatment
- How would we know?
- Graph the density function of the forcing variable
- Use density as the outcome and test for a gap at the point of the discontinuity
- Use McCrary density test (available in R, RDD package)
    - The null is no sorting

An Illustration

Evidence for Sorting

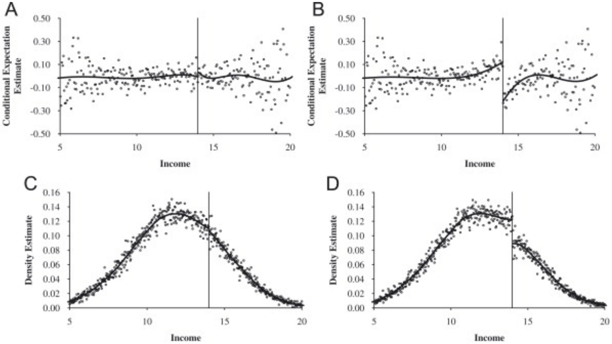
Example: Beneficial job training program offered to agents with income $< c$. Concern, people will withhold labor to lower their income below the cut-off to gain access to the program.



How to test for it

Applying the RD in the forcing variable

Example: Beneficial job training program offered to agents with income $< c$. Concern, people will withhold labor to lower their income below the cut-off to gain access to the program.



Concerns

2. Other Jumps

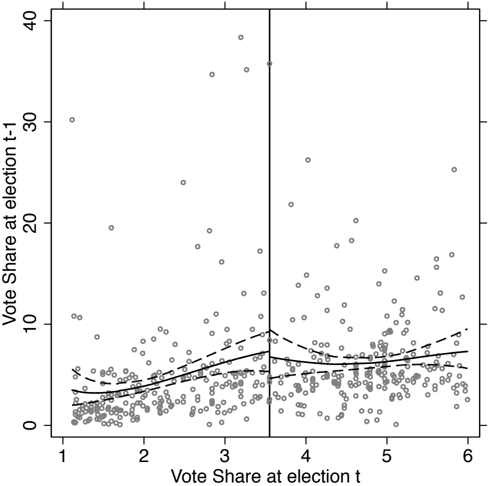
- A good check is to look for within group jumps.
- split each group, below or above the threshold, in the median (to maximize power) and treat this point as the cutoff.
- Run the RDD
- You hope not to find a jump. You use only all observations with $X < c$ and separately another test with all observations $X \geq c$

3. Placebo Outcomes

- Use placebo outcomes as dependent variables and again hope for zero-effects

A typical Placebo Outcome

Use Y_{t-1} as the outcome of interest



Fuzzy RD

The Intuition

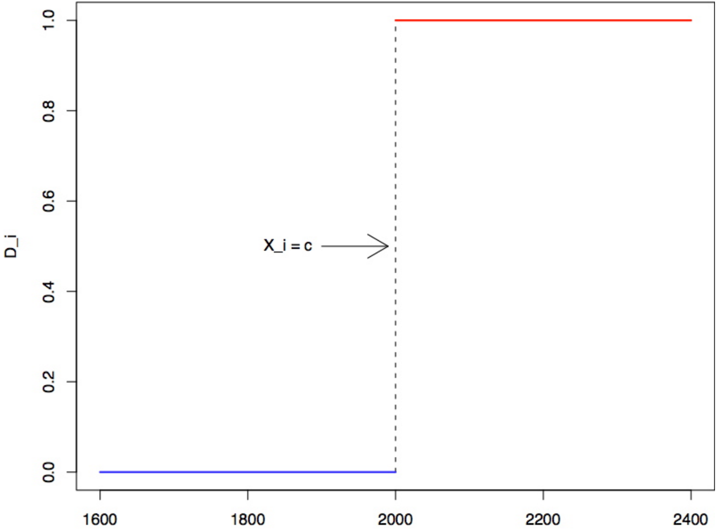
- Sometimes crossing the threshold increases the probability of taking the treating but does not guarantee treatment intake.
- This is still useful if we combine the ideas of IVs and RDD
- We use the forcing variable, as an instrument for treatment:

$$P(D_i = 1|x_i) = \begin{cases} g_1(x_i) & \text{if} \quad x_i \geq c \\ g_0(x_i) & \text{if} \quad x_i < c \end{cases}, \text{where} \quad g_1(c) \neq g_0(c).$$

The functions $g_0(x_i)$ and $g_1(x_i)$ can be anything as long as they differ (and the more the better) at c. If $g_1(c) > g_0(c)$, then $x_i \geq c$ makes treatment more likely (and vice versa).

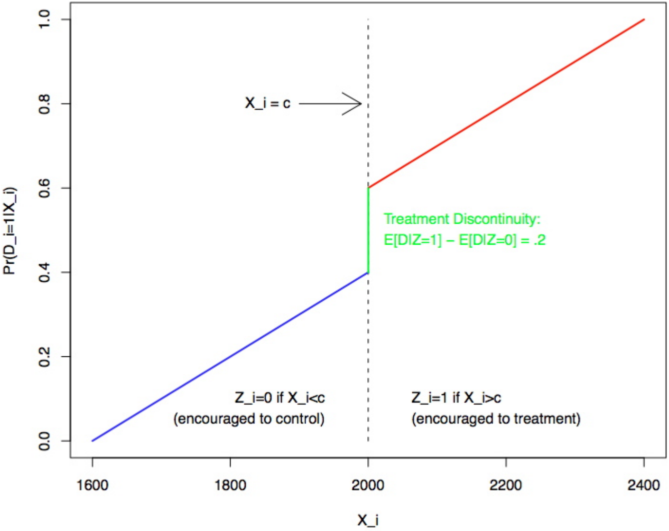
An Illustration

Remember the Sharp RD: How the treatment is assigned



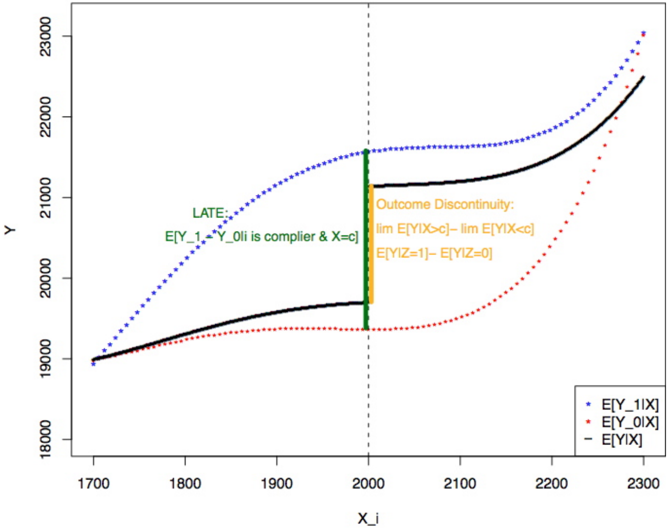
An Illustration: $E[D|X]$

Compare with the fuzzy RD: How the treatment is assigned



An Illustration: $E[Y|X]$

The Outcome graph remains similar



Identification

Identification Assumption

- Binary Instrument Z with $Z = 1\{X > c\}$
- Find the optimal bandwidth and focus on the observations within this bandwidth, where $E(Y|D, X)$ jumps, so that $X \approx c$ and thus $E(X|Z = 1) - E(X|Z = 0) \approx 0$
- All IV assumptions hold (ignorability, exclusion, monotonocity etc.)

Thus, for a neighborhood Δ, centered around c:

The Fuzzy RD Estimand

$$\tau_{FRDD} = E[Y_1 - Y_0 | X = c \quad \text{and } i \text{ is a complier}]$$

$$= \lim_{\Delta \to 0} \frac{E[Y_i | c < x_i < c + \Delta] - E[Y_i | c - \Delta < x_i < c]}{E[D_i | c < x_i < c + \Delta] - E[D_i | c - \Delta < x_i < c]}$$

Estimation

Intuition

Z is the instrument. $Z = 1$ if you pass the threshold and $Z = 0$ if not.

D is the treatment. $D = 1$ if you take the treatment (e.g. scholarship) and $D = 0$ if not.

In this setup, we apply the IV estimation.

- Step 1: Run the first-stage regression:
  $D = \alpha + \beta_1 Z + \beta_2 (X - c) + \beta_3 Z \cdot (X - c) + u_i$
- Take the $\widehat{D_i}$
- Regress Y on $\widehat{D}, X$ as if it were a sharp RD:
- $Y = \alpha + \beta_1 (X - c) + \tau \widehat{D} + \beta_2 \widehat{D} \cdot (X - c) + u_i$
- Focus on τ
- Specification can be more flexible by adding *(p)* polynomials $(X^2, X^3, ..., X^p)$ and their interactions with $\widehat{D}$.
- Check First Stage
- If you want to include controls, do so, but in both stages, the same as with a normal IV.

Final Checklist

If you want to use an RDD:

- Find a discontinuity that determines or affects who gets the treatment
- Estimate the treatment effect using:
  - a global polynomial equation
  - a local linear regression
- In the second case, first find the optimal bandwidth
- Use different bandwidths (preferably smaller) to check the robustness of the results
- Check for sorting
- Check for within-group jumps
- Check for discontinuities in placebo outcomes

In a Fuzzy RDD

- Do all the above plus:
- Graph $E[D|X]$, the probability of taking the treatment given the forcing variable

Appendix: The RD & Randomization Inference

RD as a Local Randomized Experiment

The Intuition

If variation in the treatment near the threshold is approximately randomized, then it follows that all "baseline characteristics" –all those variables determined prior to the realization of the assignment variable– should have the same distribution just above and just below the cutoff (Lee and Lemieux, 2009)

If this happens, then you can treat units below and above the cutoff point as exhangeable. This gives options for difference-of-means estimation. But two problems remain:

- What decision rule to use to cut the window now?
- What inference strategy to use, after having been left with a very small portion of the data?

Adressing the two Questions

Choosing the Window

- We test for sequentially nested windows (from small to large) whether any covariate imbalances between treated units and untreated units are greater than what we would expect to occur from sampling variability alone (Cattaneo, Frandsen and Titiunik 2013).

- Can we reject the sharp null hypothesis that there was no effect of the covariates taken together on the treatment assignment of any unit?

What Inference Strategy?

- Using Randomization Inference (a quick guide to this follows . . .)

Randomization Inference

The idea is based on one specific hypothesis, the **Sharp Null**:

- A hypothesis of no effect *for all units*
- A unit labeled as "treated" will have the exact same outcome as a unit labeled as "control."
- Under the null, the units' responses are fixed and the only random element is the meaningless rotation of labels.
- Intellectual Background: Fisher & "The Lady Teasting Tea"

Another Example

How many ways are there to choose 4 cups out of 8 cups?
(without replacement)

Pick the First One

How many choices do you have?

Pick the First One

How many choices did you have?

8

Pick the Second One

How many choices did you have?

Pick the Second One

How many choices did you have?

8×7

Pick the Third One

How many choices did you have?

$8 \times 7 \times 6$

Pick the Fourth One

How many choices did you have?

$$8 \quad \times \quad 7 \quad \times \quad 6 \quad \times \quad 5$$

Did we miss something out?

$$8 \times 7 \times 6 \times 5 = 1,680.$$

Do we really have 1,680 ways of choosing 4 cups?

$$8 \qquad \times \qquad 7 \qquad \times \qquad 6 \qquad \times \qquad 5$$

Did we miss something out?

What if I had started the other way round?

Did we miss something out?

What if I had started the other way round?

Did we miss something out?

What if I had started the other way round?

Did we miss something out?

What if I had started the other way round?

Would it matter for the experiment?

- Stopping at 1,680 would have considered this selection different to the previous one that ended with the first four cups being chosen. And it would have considered any different order that would end up with the same cups as different.
- This means we need to take into account in how many ways we could have taken the cups we did.

How many ways are there to arrange four cups?

Pick One:

How many ways are there to arrange four cups?

Pick One:

4

How many ways are there to arrange four cups?

Repeating the procedure . . .

$$4 \times 3 \times 2 \times 1$$

How many ways are there to arrange four cups?

Repeating the procedure . . .

24 Ways

Fisher's Exact Test

Estimating the Probability of H_0

- So, we have in total 1,680/24=70 ways in which we can choose randomly 4 cups out of 8
- 70 means the total number of possible outcomes is 70. This is always our denominator.
- Out of these 70 possible permutations, how many would include 4 correct choices and no incorrect choices?
- In other words, out of these 70 possible ways of choosing 4 out of 8 cups, how many could be: 4 correct and 0 incorrectly chosen cups?
- ...
- What is the probability of choosing this by accident? $1/70 = 0.014$. This is the significance level for testing the null hypothesis that the lady has no ability to discriminate
- Could we believe then the lady's claim?

What about the other probabilities?

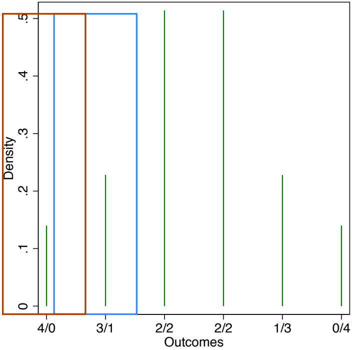
The Lady has ...

- 1 way of making 4 correct and 0 incorrect choices
- 16 ways of making 3 correct and 1 incorrect choices
- 36 ways of making 2 correct and 2 incorrect choices
- 16 ways of making 1 correct and 3 incorrect choices
- 1 way of making 0 incorrect and 4 correct choices
- If we add up all the possible outcomes, how many outcomes will we get?
- 70

H_0: Associating Outcomes with Probabilities

Cumulative Probability:

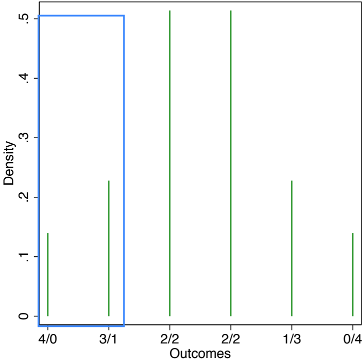
- The probability of finding 3 cups is: 16/70
- Do we stop here?
- The probability of finding 4 cups is: 1/70
- The probability of finding at least 3 cups is: 1.7=0.24



H_0: Associating Outcomes with Probabilities

Cumulative Probability:

- The probability of finding 3 cups is: $16/70$
- Do we stop here?
- The probability of finding 4 cups is: $1/70$
- The probability of finding at least 3 cups is: 1.7=0.24

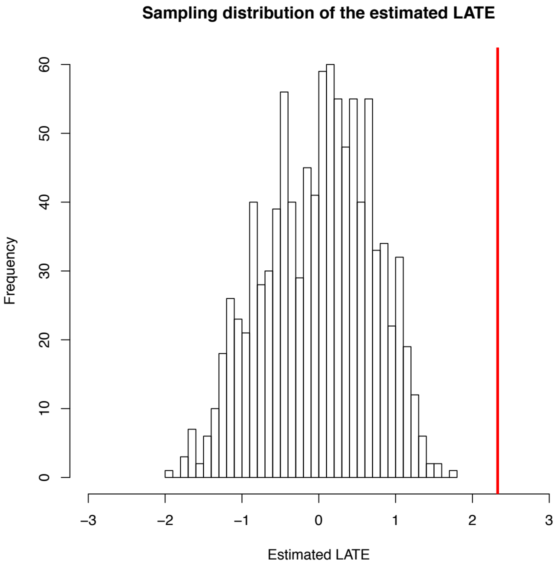


Fisher's Test: An Assessment

- The test is distribution (and model) free. The only probability used in this experiment is the probability created by the experimenter.
- If the randomisation model is not correct the test level will still be correct
- But in this test the problem is that we are not very sensitive. With only 8 cups and fixed margins even one mistake makes us fail to reject the null, gives a p-value of .24
- The key point here is that this is not a feature of the test, but rather of the design. Had we allowed for a different and more sensitive design, we would have more permutations and thus we could still reject the null even if the lady had made one mistake: Binomial probability test (`bitest` in *STATA*).
- This test is Fisher's exact test: exact in that you get the exact probability of observing what you get from the data by chance.
- Randomisation becomes the key idea in statistics with this experiment.
- This is the formal introduction of experiments in social sciences.

Back to the RD

- Step 1: Calculate differences-in-means between treated and untreated observations (within the chosen window for which local randomization assumption holds).
- Step 2: Simulate large number of hypothetical local randomization outcomes around the threshold under assumption that the LATE is zero for all units.
- Step 3: Compare differences-in-means estimate from data to the mean of all differences-in-means estimates over all hypothetical local randomization outcomes under assumption of no treatment effect for any observation.
- Remember: If the sharp null hypothesis was true and the LATE was 0 for all units, then we can just randomly reassign the units to fall below or above the threshold and estimate the LATE from every reassignment.
- If we do that 10000 times, we get the sampling distribution of LATEs under the assumption of no treatment effect for any unit.


Sampling distribution of the estimated LATE
Frequency
0
10
20
30
40
50
60
−3
−2
−1
0
1
2
3
Estimated LATE

Constructing CIs

- Impose further assumption: No treatment effect heterogeneity in window directly surrounding the threshold.
- Impute missing “treated” values by adding the estimated LATE to the observed untreated values.
- Impute missing “untreated” values by substracting the estimated LATE from the observed treated values.
- List the estimated LATEs from each local randomization in ascending order: The 2.5th percentile marks the bottom of the 95% confidence interval, and the estimate at the 97.5th percentile marks the top.

Difference-in-Differences

Intuition

Creative use of variation both over time and across space:

Fixed effects estimation of aggregate data

Imagine we want to evaluate the efficiency of a new program aiming at making the allocation of benefits within municipalities more effective. The project is applied to some municipalities but not others.

Scenario 1

- We observe funding allocation only after the policy has been implemented.
- Compare the chosen municipalities with those not chosen
- What is the problem here?
- How do we know if the two groups are *exchangeable*?

Intuition

Creative use of variation both over time and across space:

Fixed effects estimation of aggregate data

Imagine we want to evaluate the efficiency of a new program aiming at making the allocation of benefits within municipalities more effective. The project is applied to some municipalities but not others.

Scenario 2

- Imagine we observe only the chosen municipalities, both before and after the policy introduction
- Compare their outcomes before and after the policy introduction
- What is the problem here?
- How do we know any change is because of the intervention and not because of general national macroeconomic trends?

Intuition

Creative use of variation both over time and across space:

Fixed effects estimation of aggregate data

Imagine we want to evaluate the efficiency of a new program aiming at making the allocation of benefits within municipalities more effective. The project is applied to some municipalities but not others.

Scenario 3

- Imagine we observe both treated and untreated municipalities, both before and after the intervention
- Compare Treated with untreated before and after the intervention:

$$\tau = [\mathit{Treated}_{t=1} - \mathit{Untreated}_{t=1}] - [\mathit{Treated}_{t=0} - \mathit{Untreated}_{t=0}]$$

- Imagine $\tau > 0$. Could it be because of the national economy?
- Could it be because the treated municipalities are generally different from the non-treated municipalities?

An Example

Voting After Bombing

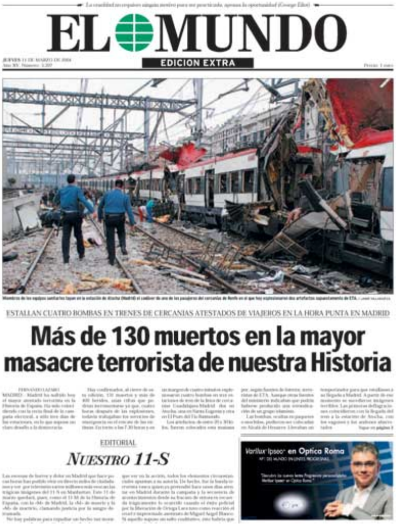
The Spanish 2004 general seemed to be an comfortable victory for the incumbent Conservative Party (PP). But things did not go as expected:

- On March 11, three days before election, Islamic terrorists deposited nine backpacks full of explosive in several trains arriving in a central train station in Madrid (Atocha). 191 people were killed and 1,500 wounded.
- Spain had been declared as one of al Qaeda's targets, after the decision of the Spanish government in March 2003 to join the US in the war against Iraq.
- Although PP targeted ETA, it was soon realized that the attack had been executed by an islamic terrorist group.

Did the Attack affect the outcome of the 2004 election? Would PP have won in the absence of the Attack?

Atocha

Atocha



La crueldad no requiere ningún motivo para ser practicada, apenas la oportunidad (George Eliot)

EL MUNDO

JUEVES 11 DE MARZO DE 2004

EDICION EXTRA

Precio: 1 euro

ESTALLAN CUATRO BOMBAS EN TRENES DE CERCANIAS ATESTADOS DE VIAJEROS EN LA HORA PUNTA EN MADRID

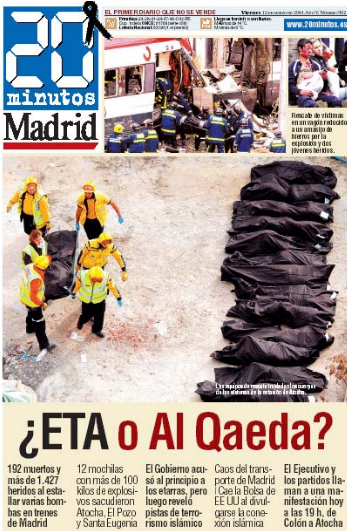
Más de 130 muertos en la mayor masacre terrorista de nuestra Historia

FERNANDO LÁZARO

EDITORIAL

NUESTRO 11-S

Varilux Ipseo en Optica Roma



20 minutos Madrid

EL PRIMER DIARIO QUE NO SE VENDE

Viernes 12 de marzo de 2004

www.20minutos.es

Rescate de víctimas en un vagón reducido a un amasijo de hierros por la explosión y dos jóvenes heridos.

Los equipos de rescate trasladan los cuerpos de las víctimas de la estación de Atocha.

¿ETA o Al Qaeda?

192 muertos y más de 1.427 heridos al estallar varias bombas en trenes de Madrid

12 mochilas con más de 100 kilos de explosivos sacudieron Atocha, El Pozo y Santa Eugenia

El Gobierno acusó al principio a los etarras, pero luego reveló pistas de terrorismo islámico

Caos del transporte de Madrid | Cae la Bolsa de EE UU al divulgarse la conexión islámica

El Ejecutivo y los partidos llaman a una manifestación hoy a las 19 h, de Colón a Atocha

The setup

The Question

Did the terrorist attack determine the winner of the 2004 Spanish election? [The implications of this question are important, since for example the first thing the socialist party did after it got elected was to withdraw, as promised, from Iraq].

The Problem

We cannot know what people would have voted without the event. We cannot observe them. Hence, most studies have rested upon post-election surveys, where people are asked their views about the attack. We are missing the counterfactual: $E(Y_0|D=1)$ (Voting without knowing about Atocha.

The Solution

Absent citizens (voters) who live or reside abroad could vote in the local Spanish consulate either in person or by mail by 7, March 2004. Compare the 14 March votes with the absentee votes cast before the event.

Is this comparison enough?

The setup

The Question

Did the terrorist attack determine the winner of the 2004 Spanish election?

The Problem

We cannot know what people would have voted without the event. We are missing the counterfactual: $E(Y_0|D=1)$ (Voting without knowing about Atocha.)

The Solution

Absent citizens (voters) who live or reside abroad could vote in the local Spanish consulate either in person or by mail by 7, March 2004. Compare the 14 March votes with the absentee votes cast before the event.

No, this is why we will use the Difference-In-Differences estimator:

$$[PP_{Nationals2004} - PP_{Absent2004}] - [PP_{Nationals2000} - PP_{Absent2000}]$$

The Framework

Definitions (The simplest case)

Two groups

- $D = 1$ Treated Units
- $D = 0$ Control Units

Two periods

- $T = 0$ Pre-treatment Period
- $T = 1$ Post-treatment Period

Potential Outcomes $Y_d(t)$

- $Y_{1i}(t)$ Potential outcome i attains in period t if treated between t_0 and t_1
- $Y_{0i}(t)$ Potential outcome i attains in period t if *not* treated between t_0 and t_1

The Framework

Definitions (The simplest case)

Two groups

- $D = 1$ Voting knowing about Atocha
- $D = 0$ Voting without knowing about Atocha

Two periods

- $T = 0$ Before Atocha
- $T = 1$ After Atocha

Potential Outcomes $Y_d(t)$

- $Y_{1i}(t)$ Vote choice if i casting a ballot after March 11.
- $Y_{0i}(t)$ Vote choice if i casting a ballot before March 11.

The DD Estimand

Defining The Causal Effect

$\tau_{it} = Y_{1i}(t) - Y_{0i}(t)$
But we are interested in what happens in t_1, because the treatment is only realized after t_0. Thus, we would like to use the classic ATT estimand, formulated as follows:
$\tau_{ATT} = E[Y_1(1) - Y_0(1)|D = 1]$

The problem is D occurs only after $t = 0$ ($D_i = D_i(1)$ and $Y_i(0) = Y_{0i}(0)$) we have $Y_i(1) = Y_{0i}(1) \cdot (1 - D_i) + Y_{1i}(1) \cdot D_i$

The Potential Outcomes

Let's now summarize what we observe:

	Post-Period(t_1)	Pre-Period(t_0)
Treated $D = 1$	$E[Y_1(1)\|D = 1]$	$E[Y_0(0)\|D = 1]$
Control $D = 0$	$E[Y_0(1)\|D = 0]$	$E[Y_0(0)\|D = 0]$

The Problem

We cannot observe $E[Y_{0i}(1)|D = 1]$: What is the average post-period outcome for those treated in the absence of the treatment?

The Potential Outcomes

Defining The Causal Effect

$$\tau_{it} = Y_{1i}(t) - Y_{0i}(t)$$

Let's now summarize what we observe:

	Post-Period(t_1)	Pre-Period(t_0)
Treated $D = 1$	$E[Y_1(1)\|D = 1]$	$E[Y_0(0)\|D = 1]$
Control $D = 0$	$E[Y_0(1)\|D = 0]$	$E[Y_0(0)\|D = 0]$

Control strategy: Treated: Before and After

- Use $E[Y(1)|D = 1] - E[Y(0)|D = 1]$
- Assumes $E[Y_0(1)|D = 1] = E[Y_0(0)|D = 1]$

The Potential Outcomes

Defining The Causal Effect

$\tau_{it} = Y_{1i}(t) - Y_{0i}(t)$

Let's now summarize what we observe:

	Post-Period(t_1)	Pre-Period(t_0)
Treated $D = 1$	$E[Y_1(1)\|D = 1]$	$E[Y_0(0)\|D = 1]$
Control $D = 0$	$E[Y_0(1)\|D = 0]$	$E[Y_0(0)\|D = 0]$

Control strategy: Treated vs. Control in Post-Period

- Use $E[Y(1)|D = 1] - E[Y(1)|D = 0]$
- Assumes $E[Y_0(1)|D = 1] = E[Y_0(1)|D = 0]$

The Potential Outcomes

Defining The Causal Effect

$\tau_{it} = Y_{1i}(t) - Y_{0i}(t)$

Let's now summarize what we observe:

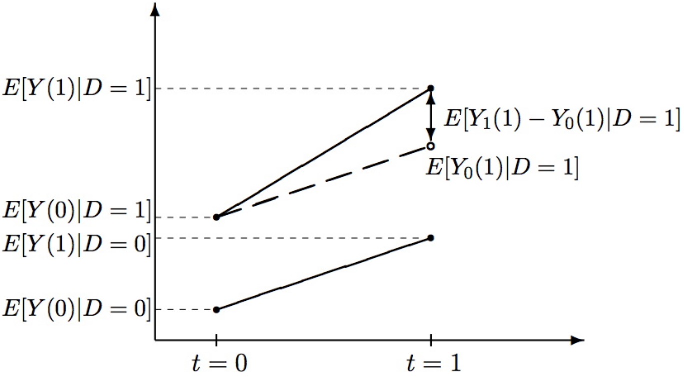
	Post-Period(t_1)	Pre-Period(t_0)
Treated $D = 1$	$E[Y_1(1)\|D = 1]$	$E[Y_0(0)\|D = 1]$
Control $D = 0$	$E[Y_0(1)\|D = 0]$	$E[Y_0(0)\|D = 0]$

Control strategy: Difference-in-Differences

$$\{E[Y(1)|D = 1] - E[Y(1)|D = 0]\} - \{E[Y(0)|D = 1] - E[Y(0)|D = 0]\}$$

- Assumes $E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0]$

A Graphical Illustration



Parallel Trends

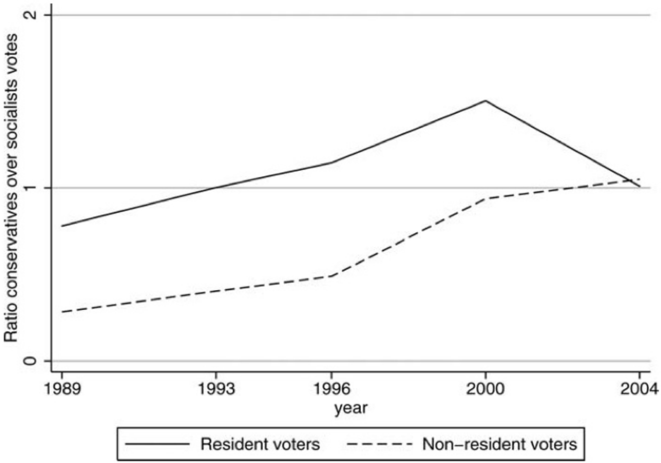
Identification Assumption

$$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0]$$

Intuition In the absence of the event/intervention etc., the treated and the control group would change from t_0 to t_1 to the same rate.
What does that mean for our example?

Parallel Trends: An Illustration

FIGURE 2.—AVERAGE RATIO ACROSS PROVINCES OF VOTES OF CONSERVATIVES OVER SOCIALISTS



Montalvo, 2011

Estimation

The Model

- The heart of DD lies in the additive structure for potential outcomes in the no-treatment state:

$$E[Y_{st}(0)|s,t] = \gamma_s + \lambda_t$$

- In the absence of the attack, vote for the Spanish Conservatives at any province $s = S$ at any time $t = T$ is determined by the sum of a time-invariant province effect and a year effect that is common across provinces
- Let $D_{s,t}$ denote a dummy for votes of those living in Spain in 2004. Assume that $E[Y_{1st} - Y_{0st}|s,t]$ is constant, τ.
- Then:

$$Y_{ist} = \gamma_s + \lambda_t + \tau D_{st} + u_{st}$$

Estimation

Deriving the ATT

$$\begin{aligned} &E[Y_{ist}|s = Absentees, t = 2004] - \\ &E[Y_{ist}|s = Absentees, t = 2000] = \\ &\lambda_{2004} - \lambda_{2000} \end{aligned}$$

$$\begin{aligned} &E[Y_{ist}|s = Residents, t = 2004] - \\ &E[Y_{ist}|s = Residents, t = 2000] = \\ &\lambda_{2004} - \lambda_{2000} + \tau \end{aligned}$$

So, the difference-in-differences estimator is:

$$\begin{aligned} &\{E[Y_{ist}|s = Residents, t = 2004] - \\ &E[Y_{ist}|s = Residents, t = 2000]\} - \\ &\{E[Y_{ist}|s = Absentees, t = 2004] - \\ &E[Y_{ist}|s = Absentees, t = 2000]\} = \tau \end{aligned}$$

The Regression Analogy

Two time points two groups

$$Y_{st} = \alpha + \gamma Resident + \lambda Year_{2004} + \tau D_{st} + u_{st}$$

An equivalent formulation could be:

$$Y_{st} = \alpha + \gamma Resident + \lambda Year_{2004} + \tau Resident \cdot Year_{2004} + u_{st}$$

Two time points many groups

$$Y_{st} = \alpha + \gamma_s + \lambda Year_{2004} + \tau D_{st} + u_{st}$$

Many time points many groups

$$Y_{st} = \alpha + \gamma_s + \lambda_t + \tau D_{st} + u_{st}$$

Practical Issues

- DD needs not be confined to two time points. Infromation about the pre-intervention state is very useful for the evaluation of the parallel trends assumption.
- How can we relax the parallel trends assumption?
- Include an (s)-specific trend: not always feasible with few time point, needs at least four time points.
- Extension: include the inclusion of two control groups instead of one: Difference-in-Differences-in-Differences.

Applications

Numerous

- Pischke looks at a policy change in German schools (but not in Bavaria) to identify the effect of schooling on income (Pichke 2007)
- Card & Kruger exploit an increase in minimum wage in New Jersey, comparing it with other states, in a DD framework as a way to see whether increase in minimum wage increases unemployment
- Bechtel and Hainmueller exploit the Elbe floods right before the 2002 German Federal election to examine whether voters vote change their votes to express their gratitude to the incumbent when receiving beneficial politics (Bechtel and Hainmueller 2011)
- Plenty of other examples are available, this is a very popular and easy-to-implement design.

Extensions

Multiple Control Groups

The Problem

Observational studies lack random assignment of units to control/treatment groups. Addressing this problem often requires conditioning on a list of observables. But are these covariates suffiicient? How would we know?

A solution

Use a second control group: In the best of circumstances, this test is consistent and unbiased, and its power exceeds the probability of falsely detecting a treatment effect." (Rosenbaum, 1987)

When does it work?

One should select two control groups to systematically vary an unobserved covariate?that is, to select groups known to differ substantially on the covariate, even though the individual values of the covariate are unknown.If bias due to the unobserved covariate is responsible for the differing outcomes in treated and control groups, then this should be apparent, because the control groups should differ from each other (Lu & Rosenbaum, 2004).

Synthetic Control Method

Combine Matching with Dif-in-Difs: Re-weight control units to construct a synthetic control unit, such that the weighted vector of observed covariates matches the observed vector.

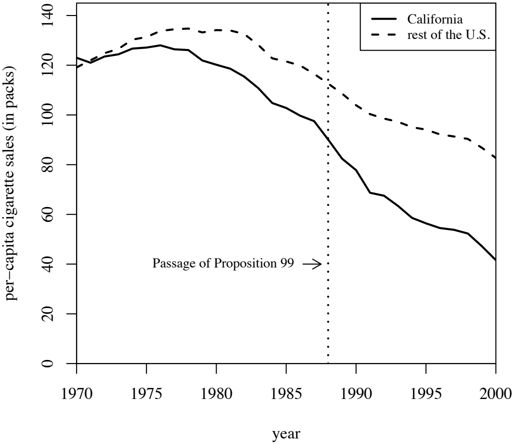
Software

Granted a list of covariates, both `Stata` & `R` easily implement the method.

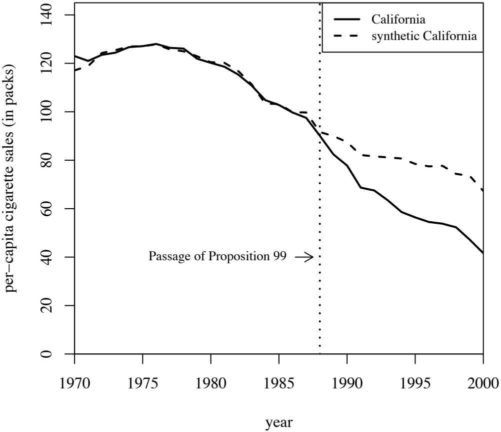
When to use it

When you have a dif-in-difs with one main control unit (comparative low-n time-T studies).

An Example: California's Proposition 99



Cigarette Consumption: CA Vs Synthetic CA

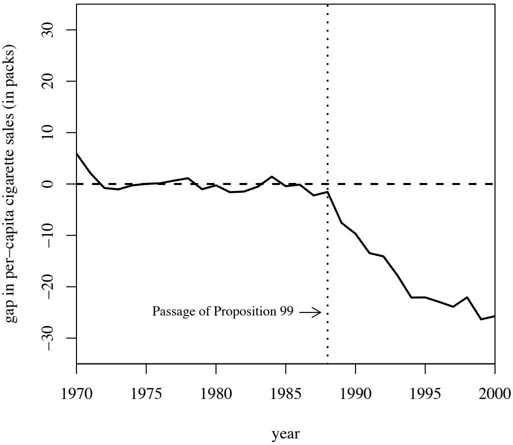


Predictor Means

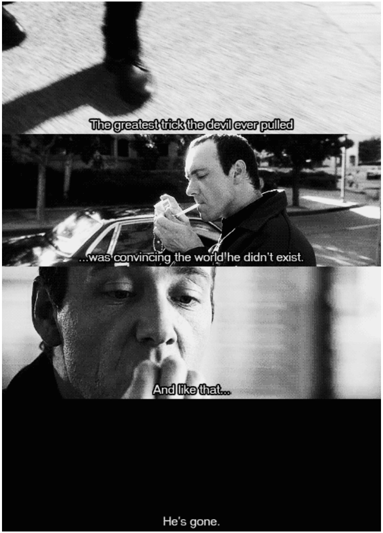
	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Smoking GAP: CA vs Synthetic CA



Don't let them fool you

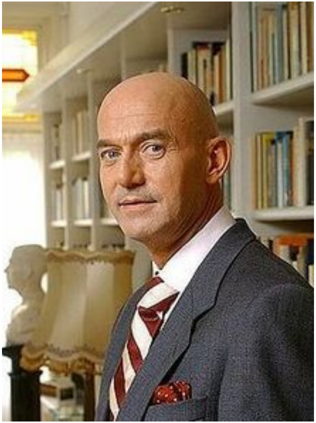

The greatest trick the devil ever pulled
...was convincing the world he didn't exist.
And like that...
He's gone.

Lab Session: Applications

Matching

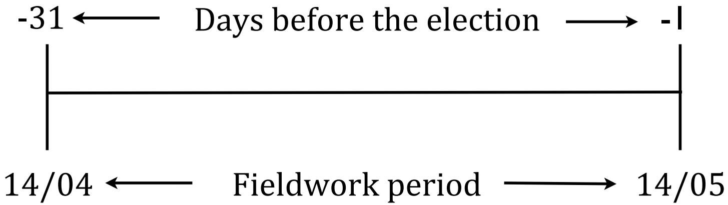
Pim Fortuyn

- Sociologist
- Ex-member of the socialist party
- “Islam a backward culture”
- “If it were legally possible I would close the borders for Muslim immigrants”
- List Pim Fortuyn in the 2002 Dutch Election



The Design

The 2002 pre-election DPES

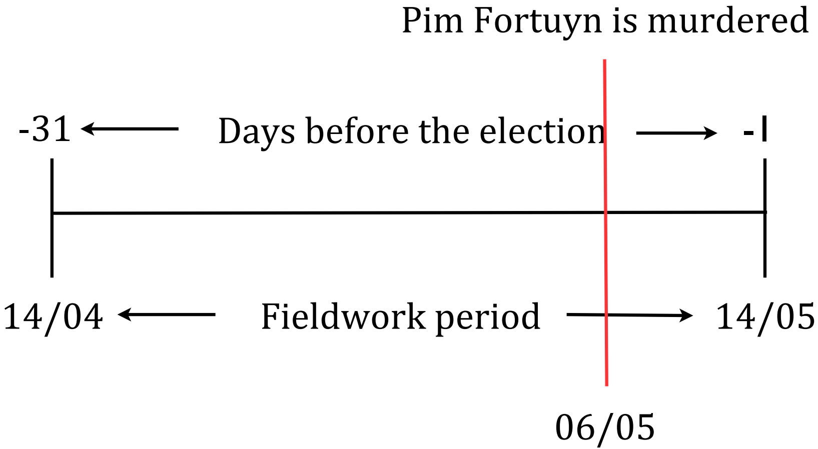


-31 ← Days before the election → -1

14/04 ← Fieldwork period → 14/05

The Design

The 2002 pre-election DPES



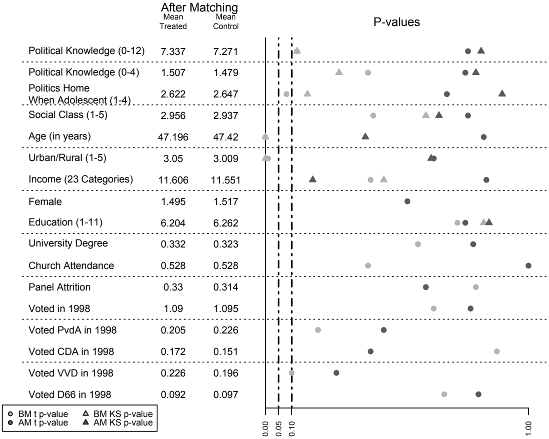
Pim Fortuyn is murdered

-31 ← Days before the election → -1

14/04 ← Fieldwork period → 14/05

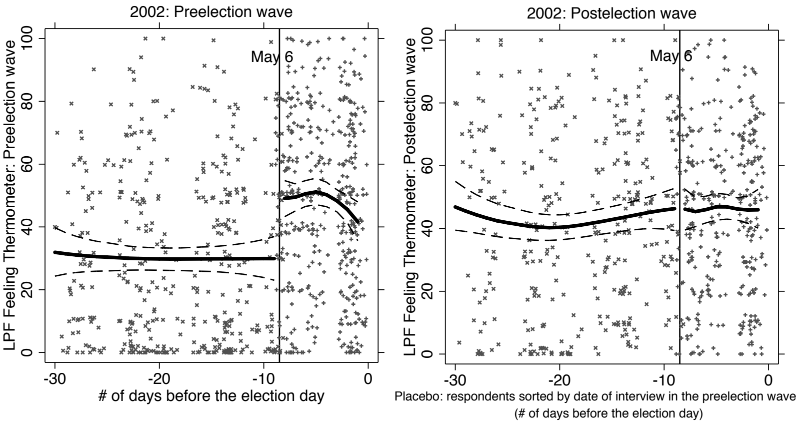
06/05

Matching



	After Matching	
	Mean Treated	Mean Control
Political Knowledge (0-12)	7.337	7.271
Political Knowledge (0-4)	1.507	1.479
Politics Home When Adolescent (1-4)	2.622	2.647
Social Class (1-5)	2.956	2.937
Age (in years)	47.196	47.42
Urban/Rural (1-5)	3.05	3.009
Income (23 Categories)	11.606	11.551
Female	1.495	1.517
Education (1-11)	6.204	6.262
University Degree	0.332	0.323
Church Attendance	0.528	0.528
Panel Attrition	0.33	0.314
Voted in 1998	1.09	1.095
Voted PvdA in 1998	0.205	0.226
Voted CDA in 1998	0.172	0.151
Voted VVD in 1998	0.226	0.196
Voted D66 in 1998	0.092	0.097

Visualization



Instrumental Variables

The Question

Why does Party ID strengthen with age?
Hypothesis: Partisan predispositions are reinforced by voting for one's preferred party. Voting entails a choice over a set of alternatives, which induces rationalization and provides a signal of group identity.

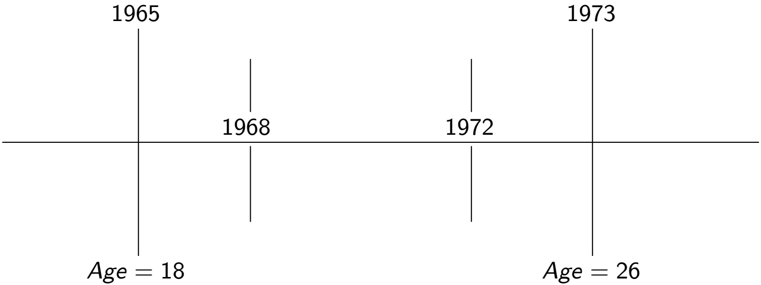
The Problem

> *[I]ndividuals choose whether to participate [and how to vote] in elections, making it difficult to separate the effects of early voting experiences from the forces that caused individuals to participate and make their party choices at first place. (Meredith, 2009:2)*

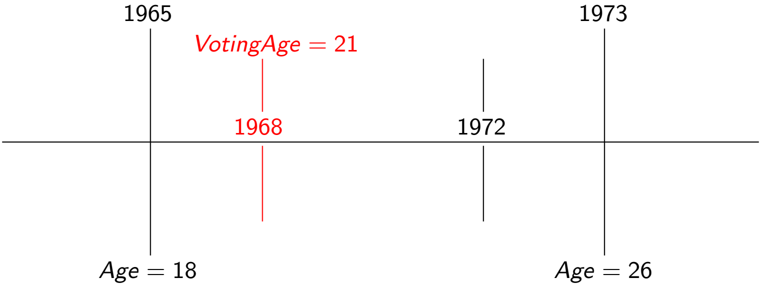
Data

The Youh-Parent Socialization Panel study, Four waves: 1965 - 1973 - 1982 - 1997

The Identification Strategy



The Identification Strategy



A Graphical Illustration

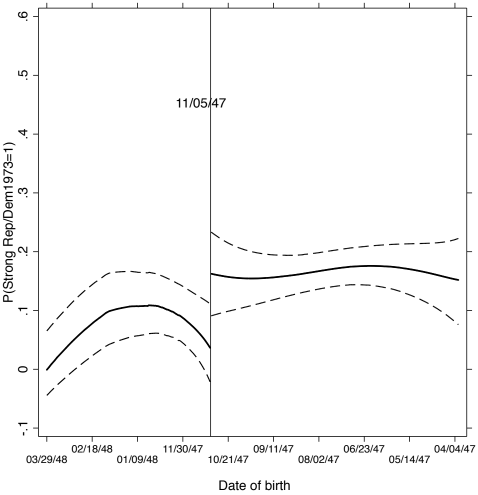


Figure: Proportion of strong partisans, by date of birth

A Graphical Illustration

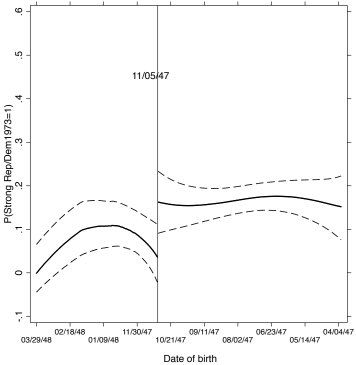


Figure: Proportion of strong partisans, by date of birth

$$ITT_{Vote68} = 16.8 - 9.5 = 7.3\%$$

$$ATT_{Vote68} = \frac{16.8 - 9.5}{0.63} = 11.58\%$$

1968 eligibility and strength of partisanship

	$(\text{Strength of PID})_{73}$
Eligible_{68}	.302 (.151)
n	782

Y:	$(\text{Strength of Partisanship})_{73}$
0:	Independent
1:	Leaner
2:	Weak Partisan
3:	Strong Partisan

A Placebo Test: Y=PID$_{65}$

	(Strength of PID)$_{65}$
Eligible$_{68}$	.058 (.287)
n	894

Y:	(Strength of Partisanship)$_{65}$
0:	Independent
1:	Leaner
2:	Weak Partisan
3:	Strong Partisan

Adjusting for one-sided non-compliance

	(Strength of PID)$_{73}$
$\widehat{Vote}_{68}$	.319 (.173)
n	774

Y:Strength of Partisanship$_{73}$ (0-3 scale):
Entries denote ATT estimates using the Wald Estimator

Adding covariates

	(Strength of PID)$_{73}$
$\widehat{Vote}_{68}$	.321 (.161)
n	774

Y:Strength of Partisanship$_{73}$(0-3 scale):
Entries denote ATT estimates using the 2SLS Estimator
Indicative controls: parental education, political interest and PID

Are the effects due to the age gap?

	(Strength of PID)$_{73}$	
Older Eligibles$_{68}$	.064	-.095
	(.287)	(.069)
(Strength of PID)$_{65}$		.520
		(.085)
n	894	894

Y:Strength of Partisanship$_{73}$ (0-3 scale):
Eligibles split into two equally-sized groups.
Young group treated as non-elibiles.

Lab Session: Applications (RDD)

Regression Discontinuity Design

The Question

Why do some small parties persist whereas others die?
Hypothesis: Parliamentary representation comes with organisational and other benefits which help small parties survive and perform better in future elections.

The Problem

> *Parliamentary representation is not randomly assigned. Parties that make it to the parliament are also likely to differ from parties that do not make it in various respects*

Solution

Use the discontinuities generated by the presence of electoral thresholds of representation in PR systems

$E(Y|X)$

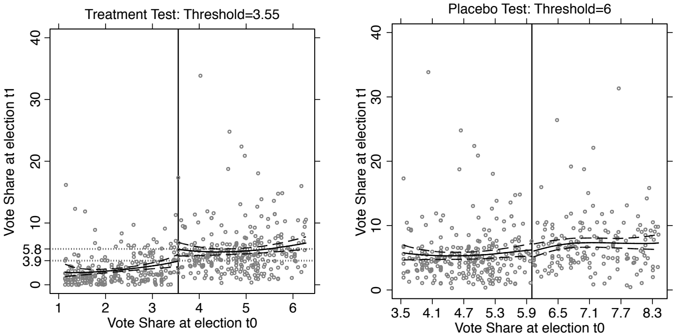


Figure: Vote share at t_1 given vote share at t_0

$E(Y|X)$: No Outliers

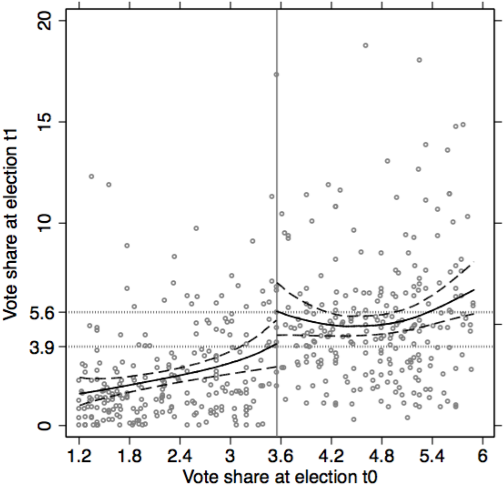


Figure: Vote share at t_1 given vote share at t_0

$E(Y|X)$: Without Israel & the Netherlands

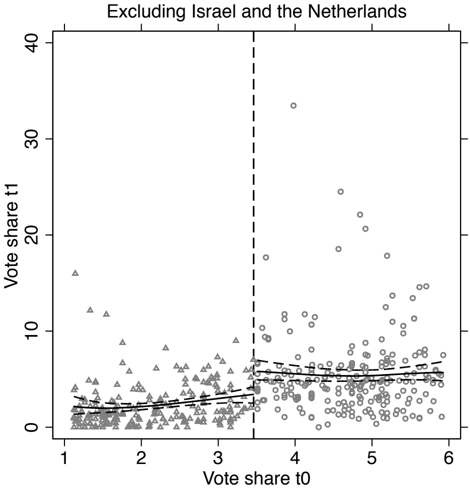


Figure: Vote share at t_1 given vote share at t_0

$E(D|X)$

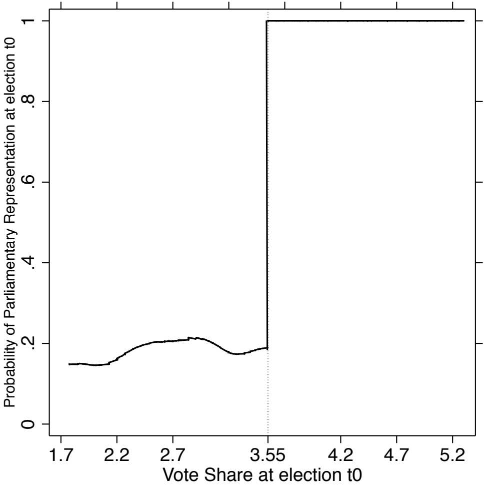


Figure: $P(D = 1)$ at t_0 given vote share at t_0

Placebo Test

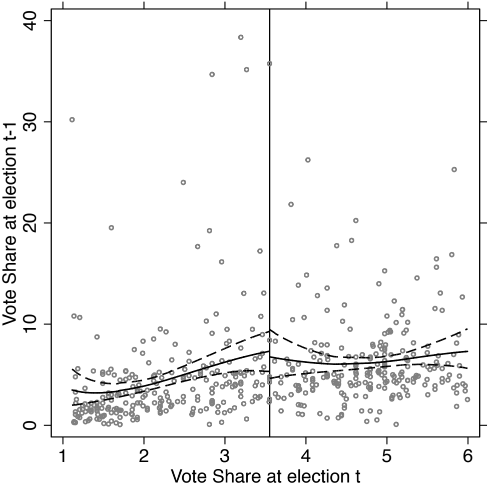


Figure: Vote share at t_{-1} given vote share at t_0

Sorting: Visualisation (1)

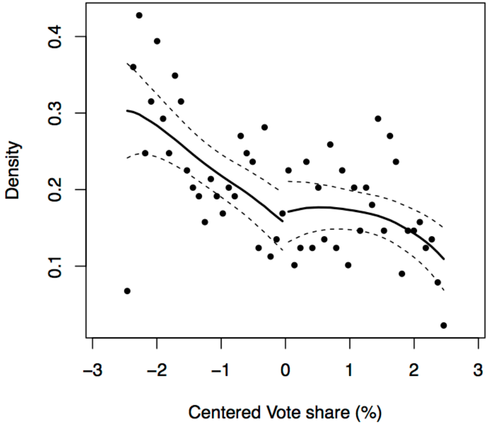


Figure: Density of X at t_0 given vote share at t_0

Sorting: Visualisation (2)

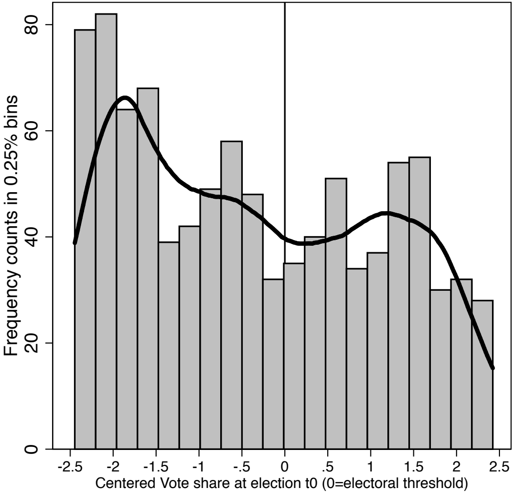


Figure: Density of X at t_0 given vote share at t_0

Lab Session: Applications (Dif-in-Difs)

The DD Running Example

The Question

Are governments rewarded when they deliver good policies?
Hypothesis: Eficient managerial administration has long-term electoral returns.

The Problem

Retrospective voting is based on memory of governmental performance as well as partisan priors, which bias voters' perceptions. Governments choose their policies on the basis of macroeconomic and electoral constraints. All these factors would bias the etimate of reported retrospective voting.

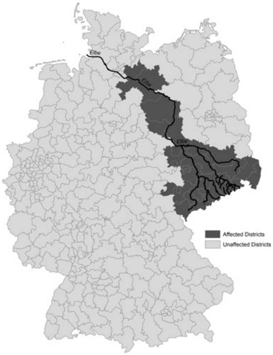
Solution

Use a natural disaster where the incumbent did a good job in managing the crisis. Use dif-in-difs to establish comparison both over time and across space (unless the whole population is equally affected by the disaster)

Schröeder's Machine God

Bechtel and Hainmueller 2011

FIGURE 1 Affected versus Unaffected Electoral Districts in the 2002 Election



Note: The map shows the boundaries of the 299 electoral districts in the 2002 German federal election. Directly flood-affected districts (i.e., *Flooded* = 1) are shaded dark gray; unaffected districts are shaded light gray. A district was coded as affected if it experienced at least one of the following events: stabilization or breach of levees, flood warning, overtopping of levee, flooding, evacuation warning, or evacuation. Source: Own computation based on flood report by the International Commission for the Protection of the Elbe River (2002).

The Model

The Intuition

Three steps:

- The floods affected part of Germany. Compare between affected and non-affected.
- Elections in Germany took place before and after the floods. Compare between before and after.
- Combine the first two comparisons:

Assumption

The parallel trends assumption:
$E[Y_{0i,2002} - Y_{0i,1998} | D_i = 1] = E[Y_{0i,2002} - Y_{0i,1998} | D_i = 0]$

Estimation

ATT

Define the ATT as
$\alpha = \{E[Y_{i,2002}|D_i = 1] - E[Yi, 1998|D_i = 1]\} - \{E[Y_{i,2002}|D_i = 0] - E[Yi, 1998|D = 0]\}$. In a regression framework:
$Y_{it} = \gamma_i + \delta_t + \alpha D_{it} + X^o_{it}\beta + u_{it}$

The Results

Short- and long-term results & the placebo.

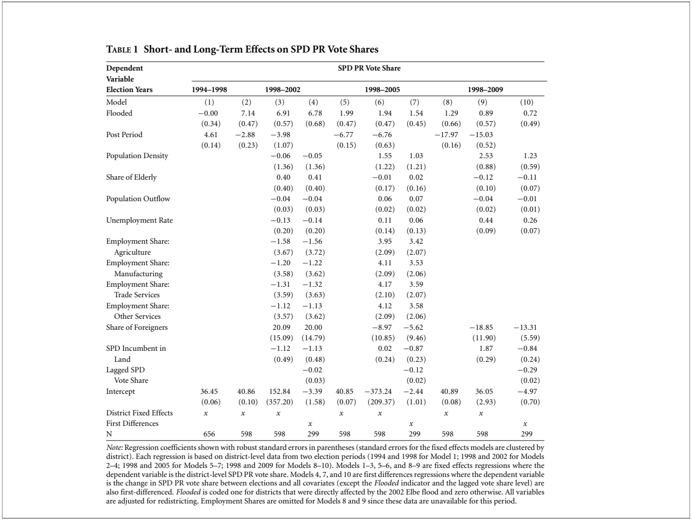


Table 1 Short- and Long-Term Effects on SPD PR Vote Shares

Dependent Variable					**SPD PR Vote Share**					
Election Years	**1994–1998**		**1998–2002**			**1998–2005**			**1998–2009**	
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Flooded	−0.00	7.14	6.91	6.78	1.99	1.94	1.54	1.29	0.89	0.72
	(0.34)	(0.47)	(0.57)	(0.68)	(0.47)	(0.47)	(0.45)	(0.66)	(0.57)	(0.49)
Post Period	4.61	−2.88	−3.98		−6.77	−6.76		−17.97	−15.03	
	(0.14)	(0.23)	(1.07)		(0.15)	(0.63)		(0.16)	(0.52)	
Population Density			−0.06	−0.05		1.55	1.03		2.53	1.23
			(1.36)	(1.36)		(1.22)	(1.21)		(0.88)	(0.59)
Share of Elderly			0.40	0.41		−0.01	0.02		−0.12	−0.11
			(0.40)	(0.40)		(0.17)	(0.16)		(0.10)	(0.07)
Population Outflow			−0.04	−0.04		0.06	0.07		−0.04	−0.01
			(0.03)	(0.03)		(0.02)	(0.02)		(0.02)	(0.01)
Unemployment Rate			−0.13	−0.14		0.11	0.06		0.44	0.26
			(0.20)	(0.20)		(0.14)	(0.13)		(0.09)	(0.07)
Employment Share:			−1.58	−1.56		3.95	3.42			
Agriculture			(3.67)	(3.72)		(2.09)	(2.07)			
Employment Share:			−1.20	−1.22		4.11	3.53			
Manufacturing			(3.58)	(3.62)		(2.09)	(2.06)			
Employment Share:			−1.31	−1.32		4.17	3.59			
Trade Services			(3.59)	(3.63)		(2.10)	(2.07)			
Employment Share:			−1.12	−1.13		4.12	3.58			
Other Services			(3.57)	(3.62)		(2.09)	(2.06)			
Share of Foreigners			20.09	20.00		−8.97	−5.62		−18.85	−13.31
			(15.09)	(14.79)		(10.85)	(9.46)		(11.90)	(5.59)
SPD Incumbent in			−1.12	−1.13		0.02	−0.87		1.87	−0.84
Land			(0.49)	(0.48)		(0.24)	(0.23)		(0.29)	(0.24)
Lagged SPD				−0.02			−0.12			−0.29
Vote Share				(0.03)			(0.02)			(0.02)
Intercept	36.45	40.86	152.84	−3.39	40.85	−373.24	−2.44	40.89	36.05	−4.97
	(0.06)	(0.10)	(357.20)	(1.58)	(0.07)	(209.37)	(1.01)	(0.08)	(2.93)	(0.70)
District Fixed Effects	*x*	*x*	*x*		*x*	*x*		*x*	*x*	
First Differences				*x*			*x*			*x*
N	656	598	598	299	598	598	299	598	598	299

Note: Regression coefficients shown with robust standard errors in parentheses (standard errors for the fixed effects models are clustered by district). Each regression is based on district-level data from two election periods (1994 and 1998 for Model 1; 1998 and 2002 for Models 2–4; 1998 and 2005 for Models 5–7; 1998 and 2009 for Models 8–10). Models 1–3, 5–6, and 8–9 are fixed effects regressions where the dependent variable is the district-level SPD PR vote share. Models 4, 7, and 10 are first differences regressions where the dependent variable is the change in SPD PR vote share between elections and all covariates (except the *Flooded* indicator and the lagged vote share level) are also first-differenced. *Flooded* is coded one for districts that were directly affected by the 2002 Elbe flood and zero otherwise. All variables are adjusted for redistricting. Employment Shares are omitted for Models 8 and 9 since these data are unavailable for this period.

Wednesday, 11 June 14

The Parallel Trends Assumption?

Using Opinion Poll Data

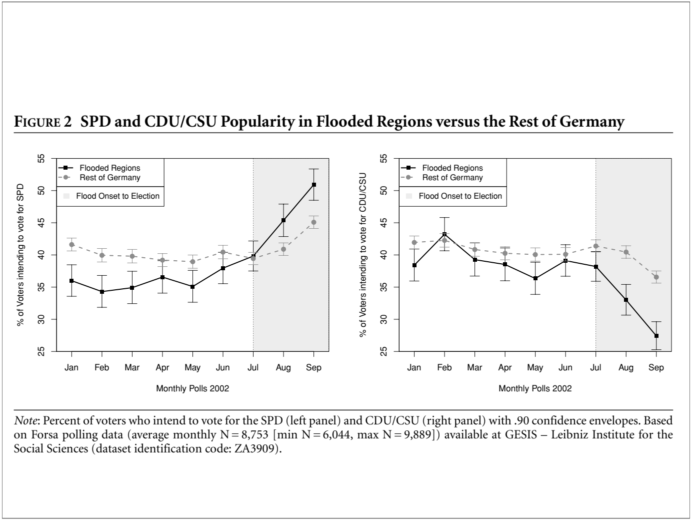


FIGURE 2 SPD and CDU/CSU Popularity in Flooded Regions versus the Rest of Germany

Note: Percent of voters who intend to vote for the SPD (left panel) and CDU/CSU (right panel) with .90 confidence envelopes. Based on Forsa polling data (average monthly N = 8,753 [min N = 6,044, max N = 9,889]) available at GESIS – Leibniz Institute for the Social Sciences (dataset identification code: ZA3909).

Wednesday, 11 June 14

Is this sufficient evidence?