

Causal Inference and Research Design

Scott Cunningham (Baylor)

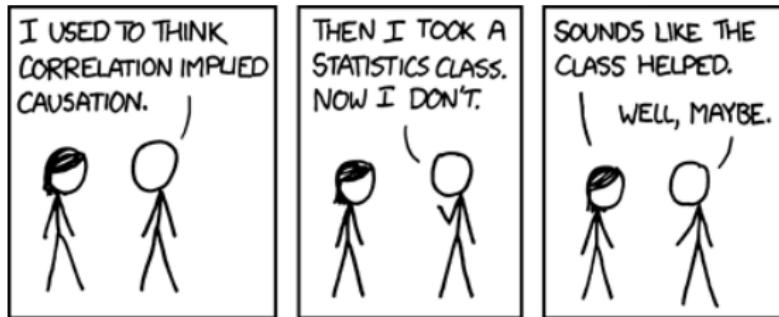


Figure: xkcd

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

- Workflow workflow workflow
- Directories
- Do files and R programs
- Naming conventions
- Versional control
- Soft skills

Welcome

- In this class, we will cover topics in a field in econometrics sometimes called causal inference
- The course will use lectures, simulations, replications, discussion of papers, an midterm, discussion and a research project to help you learn the material
- Goal is to take you to a level of comfort and confidence in applying the tools you've learned in econometrics to your own research as well as to analyze work in your field

Main goals

The main goal of the course is to teach you key concepts and successful implementation of causality relevant research designs

A secondary goal of the workshop is to provide you with programming examples in Stata and R for implementing some but not all of the procedures we'll cover

- Hidden curriculum and programming tips
- R and Stata programming
- R markdown documents at a learnr repo

https://github.com/scunning1975/mixtape_learnr

Textbooks

Main textbooks

- ① Alexander (2021) Telling Stories with Data (Alexander)
- ② Cunningham (2021) mixtape.scunning.com (Mixtape)

Other helpful books but don't feel compelled to buy them

- Angrist and Pischke (2009) Mostly Harmless Econometrics ("Angrist and Pischke")
- Morgan and Winship (2014) Counterfactuals and Causal Inference ("Morgan and Winship")

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

- Workflow workflow workflow
- Directories
- Do files and R programs
- Naming conventions
- Versional control
- Soft skills

Limitations of this material

- Hidden curriculum falls into two categories: empirical workflow and personal factors for lack of a better word
- But there is a glaring omission and that's the unique obstacles that underrepresented minorities and females face in the profession
- Hopefully this is still helpful
- Let me tell you a story

About me

- Professor of economics at Baylor (Waco Texas),
- Graduated in 2007 from University of Georgia with a field in econometrics, IO, public, and labor field courses
- I knew I was going to be an empiricist, so I made econometrics my main field – passed field exam on second attempt
- Since graduating I've focused on topics in crime and risky sex such as sex work, drug policy, abortion, mental healthcare.
- I knew I couldn't achieve my goals without learning causal inference which I could tell I had only a vague understanding of
- This is because causal inference isn't taught historically in traditional econometrics

Sad story (to me!)

- Once upon a time there was a boy who wrote a job market paper using the NLSY97.
- This boy presented the findings a half dozen times, spoke to the media a few times, got 17 interviews at the ASSA, 7 flyouts, and an offer from Baylor
- He submitted the job market paper to the *Journal of Human Resources*, a top field journal in labor, and received a “revise and resubmit” request from the editor (woo hoo!)

The horror!

- But then digging into his one directory, he found countless versions of his do file and hundreds of files with random names
- And once he finally was able to get the code running again, he found a critical coding error that when corrected ("destroyed") his results
- The young boy was devastated and never resubmitted which he does not recommend (but he was sad!)

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

Workflow workflow workflow

Directories

Do files and R programs

Naming conventions

Versional control

Soft skills

All competent empirical work is a mousetrap

“Happy families are all alike; every unhappy family is unhappy in its own way.” - Leo Tolstoy, Anna Karenina

“Good empirical work is all alike; every bad empirical work is bad in its own way.” - Scott Cunningham, This slide

Cunningham Empirical Workflow Conjecture

- The cause of most of your errors is **not** due to insufficient knowledge of syntax in your chosen programming language
- The cause of most of your errors is due to a poorly designed empirical workflow

Workflow

Wikipedia definition:

"A workflow consists of an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources into processes that transform materials, provide services, or process information."

Dictionary definition:

"the sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion."

Empirical workflow

- Workflow is a fixed set of routines you bind yourself to which when followed identifies the most common errors
 - Think of it as your morning routine: alarm goes off, go to wash up, make your coffee, check Twitter, repeat *ad infinitum*
- Finding the outlier errors is a different task; empirical workflows catch typical and common errors created by the modal data generating processes

Why do we use checklists?

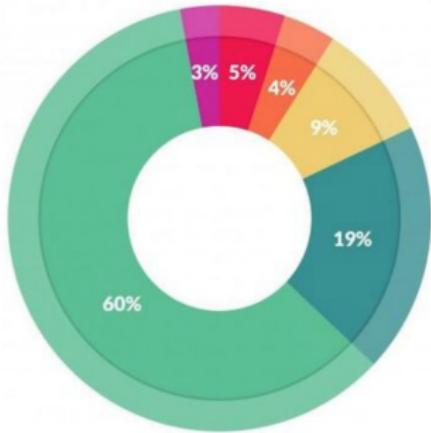
- Before going on a trip, you use a checklist to make sure you have everything you need
 - Charger (check), underwear (check), toothbrush (check), passport (oops), ...
- The empirical checklist is solely referring to the intermediate step between “getting the data” and “analyzing the data”
- It largely focuses on ensuring data quality for the most common, easiest to identify, situations you’ll find yourself in

Simple checks

- Your checklist should be a few simple, yet non-negotiable, programming commands and exercises to check for coding errors
- Let's discuss a few

Time

- People often think empirical research is about “getting the data” and “analyzing the data”
- They have an “off to the races” mindset
- Just like running a marathon involves far far more time training than you ever spend running the marathon, doing empirical research involves far far more time doing tedious, repetitive tasks
- Since you do the tedious tasks repeatedly, they have the *most* potential for error which can be catastrophic
- How can we minimize these errors through a checklist?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Figure: Image from Wenfei Xu at Columbia

Read the codebook

- We stand on the shoulders of giants
- Few like reading the codebook as it is not gripping literature
- But the codebook explains how to interpret the data you have acquired and it is not a step you can skip
- Set aside time to study it, and have it in a place where you can regularly return to it
- This goes for the `readme` that accompanies some datasets, too.

Look at the data

- The eyeball is not nearly appreciated enough for its ability to spot problems
- Use browse or excel to just read the spreadsheet with your eyes.
- Scroll through the variables and accompany yourself with what you've got visually

Data Editor (Browse) — vs.dta

Edit mode Save Find YM

date[1] 1995m1

	date	ers_ym	st_fips	county_fips	month	year	marital_st~D	marital_st~M
1	1995m1	514	6	1	1	1995	0	0
2	1995m2	514	6	1	2	1995	0	0
3	1995m3	514	6	1	3	1995	0	0
4	1995m4	514	6	1	4	1995	0	0
5	1995m5	514	6	1	5	1995	0	0
6	1995m6	514	6	1	6	1995	0	0
7	1995m7	514	6	1	7	1995	0	0
8	1995m8	514	6	1	8	1995	0	0
9	1995m9	514	6	1	9	1995	0	0
10	1995m10	514	6	1	10	1995	0	0
11	1995m11	514	6	1	11	1995	0	0
12	1995m12	514	6	1	12	1995	0	0
13	1996m1	514	6	1	1	1996	0	0
14	1996m2	514	6	1	2	1996	0	0
15	1996m3	514	6	1	3	1996	0	0
16	1996m4	514	6	1	4	1996	0	0
17	1996m5	514	6	1	5	1996	0	0
18	1996m6	514	6	1	6	1996	0	0
19	1996m7	514	6	1	7	1996	0	0
20	1996m8	514	6	1	8	1996	0	0
21	1996m9	514	6	1	9	1996	0	0
22	1996m10	514	6	1	10	1996	0	0
23	1996m11	514	6	1	11	1996	0	0
24	1996m12	514	6	1	12	1996	0	0
25	1997m1	514	6	1	1	1997	0	0
26	1997m2	514	6	1	2	1997	0	0
27	1997m3	514	6	1	3	1997	0	0

Vars: 71 Order: Dataset Obs: 565,260

Variables

Name	date
Label	State of Occurrence
Type	float
Format	%tm
Value label	
Notes	

Properties

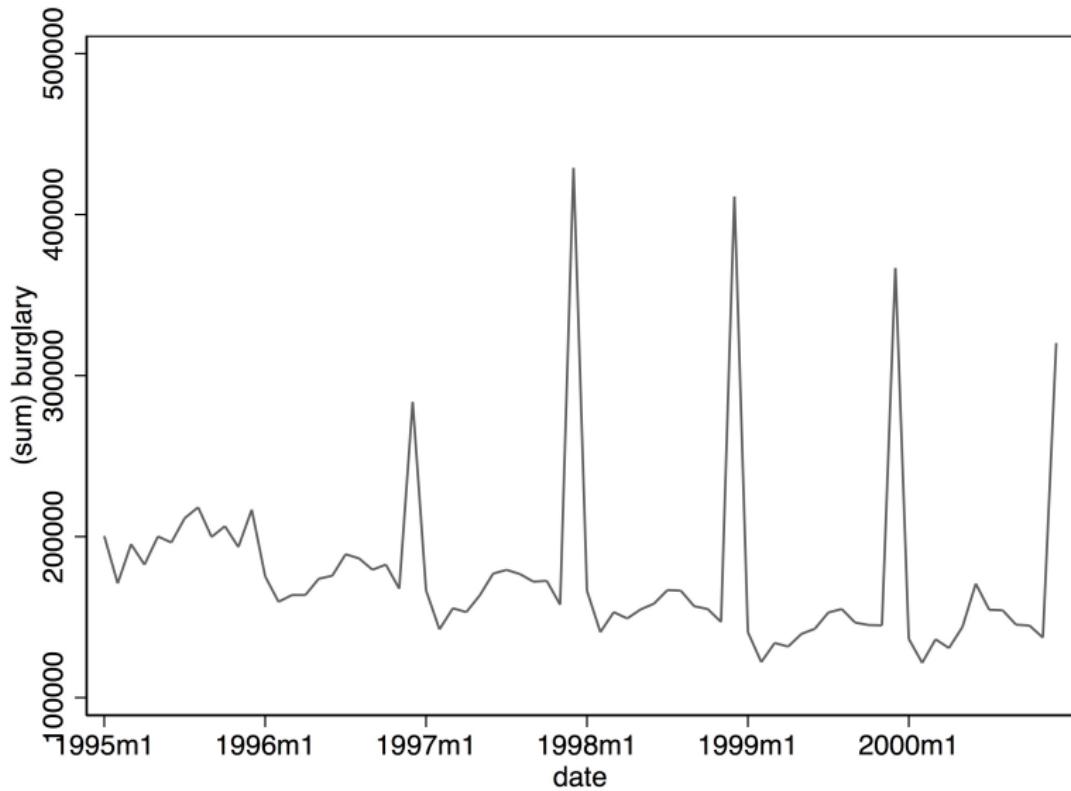
Variables

Name	date
Label	
Type	float
Format	%tm
Value label	
Notes	

Data

Frame	default
Filename	vs.dta
Label	
Notes	

Filter: Off



Missing observations

- Check the size of your dataset in Stata using `count`
- Check the number of observations per variable in Stata using `summarize`
 - String variables will always report zero observations under `summarize` so `count if X==""` will work
- Use `tabulate` also because oftentimes missing observations are recorded with a `-9` or some other illogical negative value

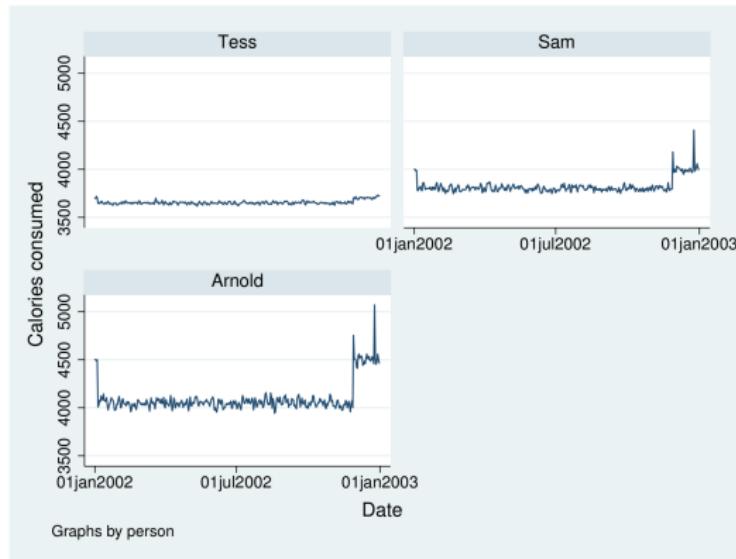
Missing years

- Panel data can be overwhelming bc looking at each state/city/firm/county borders on the impossible
- Start with collapse to the national level by year and simply list to see if anything looks strange
 - What's "strange" look like?
 - Well wouldn't it be strange if national unemployment rates were zero in any year?
- You can use `xtline` to see time series for panel identifiers, with or without the subcommand of `overlay`

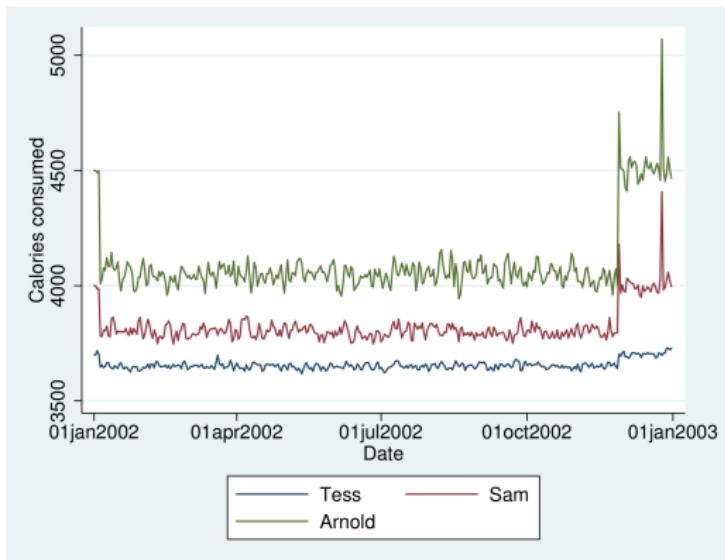
```
. collapse (sum) male_homicide female_homicide, by(year)  
. list
```

	year	male_h~e	female~e
1.	1995	0	0
2.	1996	0	0
3.	1997	0	0
4.	1998	0	0
5.	1999	0	0
6.	2000	0	0
7.	2001	0	0
8.	2002	0	0
9.	2003	4474	910
10.	2004	4270	900
11.	2005	4450	895
12.	2006	4479	889
13.	2007	4480	895
14.	2008	4228	893
15.	2009	3857	866

```
. xtline calories, tlabel(#3)
```



```
. xtline calories, overlay
```



Panel observations are $N \times T$

- Say you have 51 state units (50 states plus DC) and 10 years
- $51 \times 10 = 510$ observations
- If you do not have 510 observations, then you have an unbalanced panel; if you have 510 observations you have a balanced panel
- Check the patterns using `xtdescribe` and simple counting tricks


```
. gen one = 1  
  
. bysort county_group: egen count=sum(one)  
  
. ta count
```

count	Freq.	Percent	Cum.
24	48	0.42	0.42
36	36	0.31	0.73
48	48	0.42	1.15
96	96	0.84	1.99
120	480	4.19	6.18
156	312	2.72	8.90
180	10,440	91.10	100.00
<hr/>			
Total	11,460	100.00	

Merge

- During a stage of arranging datasets, you will likely merge – oftentimes a lot
- Make sure you count before and after you merge so you can figure out what went wrong, if anything
- Also make sure you're using the contemporary m:m syntax as many an excellent empiricists have been hurt by merge syntax errors

```
. count  
48,600  
  
. do "/Users/scott_cunningham/Dropbox/Indy/Do/.tm-stata-55642.do"  
  
. merge 1:1 id date using ../data/seer.dta  
(note: variable month was byte, now float to accommodate using data's values)
```

Result	# of obs.
not matched	517,044
from master	384 (_merge==1)
from using	516,660 (_merge==2)
matched	48,216 (_merge==3)

```
. ta _merge
```

_merge	Freq.	Percent	Cum.
master only (1)	384	0.07	0.07
using only (2)	516,660	91.40	91.47
matched (3)	48,216	8.53	100.00
Total	565,260	100.00	

```
.  
.end of do-file
```

```
. count  
565,260
```

Don't forget the question

- “Exploring the data” is intoxicating to the point of distracting
- “All you can do is write the best paper on the question you’re studying” – Mark Hoekstra
 - Note he didn’t say “Write the best paper you’re capable of writing”
 - He said **the best paper**
 - Important therefore to choose the right questions with real upside
- Slow down, think big picture, force yourself to figure out exactly what your question is, who is in your sample (and importantly who won’t be) and what time periods you’ll pull

Organize your directories

- After the coding error fiasco, I spent a lot of time wondering how this could happen
- I decided it was partly because of four problems related to
 - ① organized subdirectories
 - ② automation
 - ③ naming conventions
 - ④ version control
- I'll discuss each but I highly recommend that you just read Gentzkow and Shapiro's excellent resource "Code and Data for the Social Sciences: A Practitioner's Guide" <https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf>

No correct organization

- There is no correct way to organize your directories,
- But all competent empiricists have adopted an intentional philosophy of how to organize their directories
- Why? Because you're writing for your future self, and your future self is lazy, distracted, disinterested and busy

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

Workflow workflow workflow

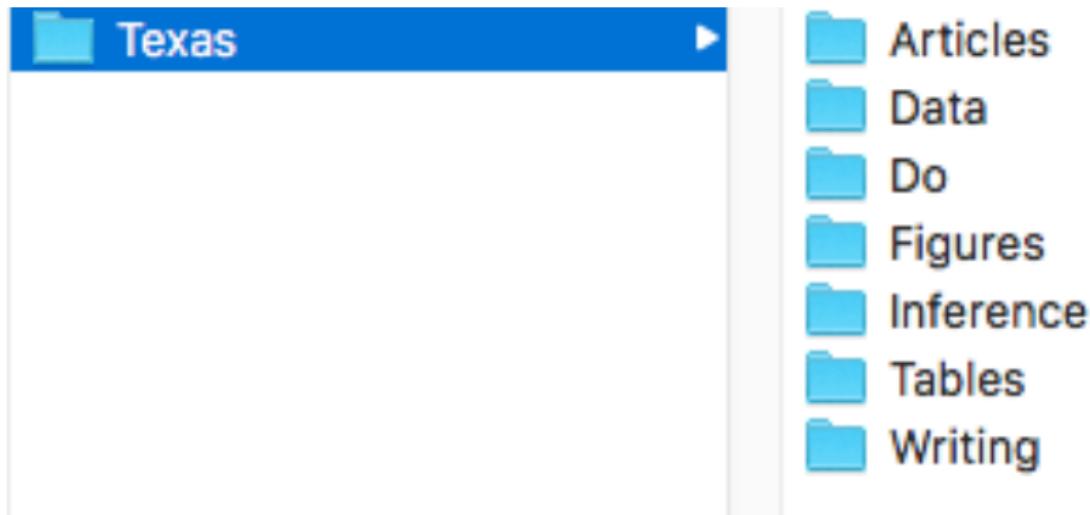
Directories

- Do files and R programs
- Naming conventions
- Versional control
- Soft skills

Directories

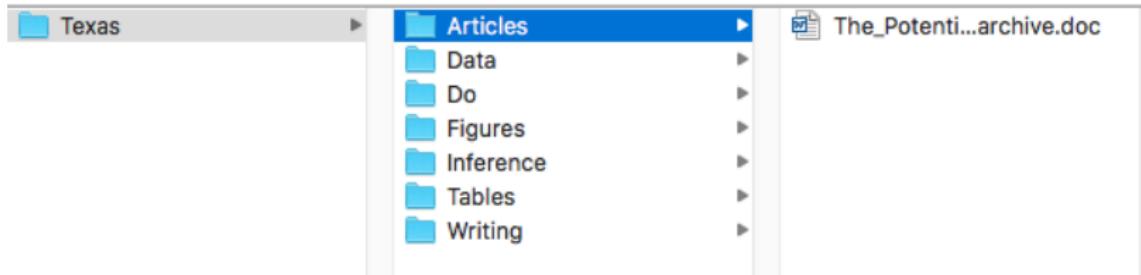
- The typical applied micro project may have hundreds of files of various type and will take *years* just to finish not including time to publication
- So simply finding the files you need becomes more difficult if everything is stored in the same place
- When I start a new project, the first thing I do is create the following directories

Subdirectory organization



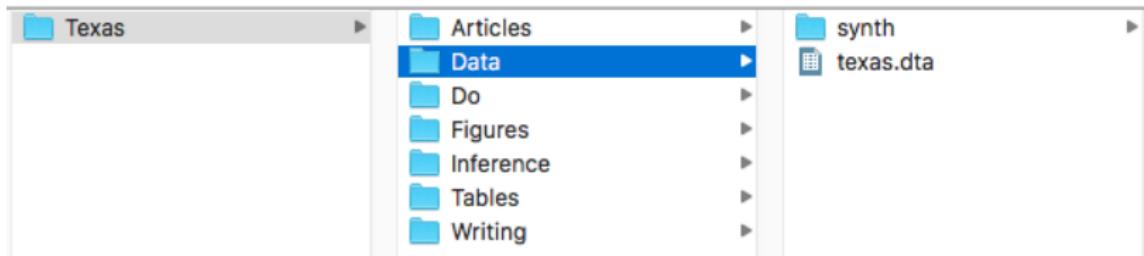
- 1) Name the project ("Texas")

Subdirectory organization



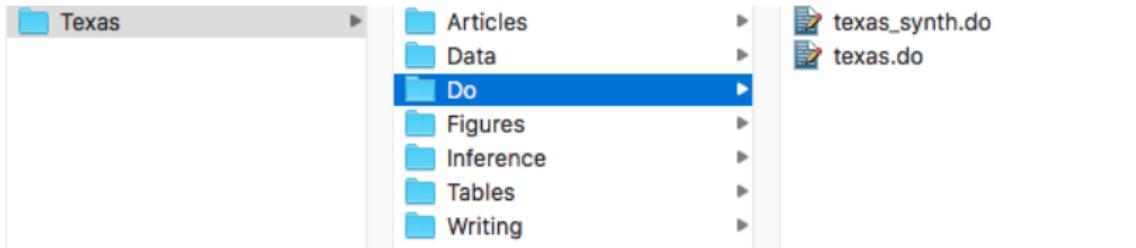
- 2) A subdirectory for all articles you cite in the paper

Subdirectory organization



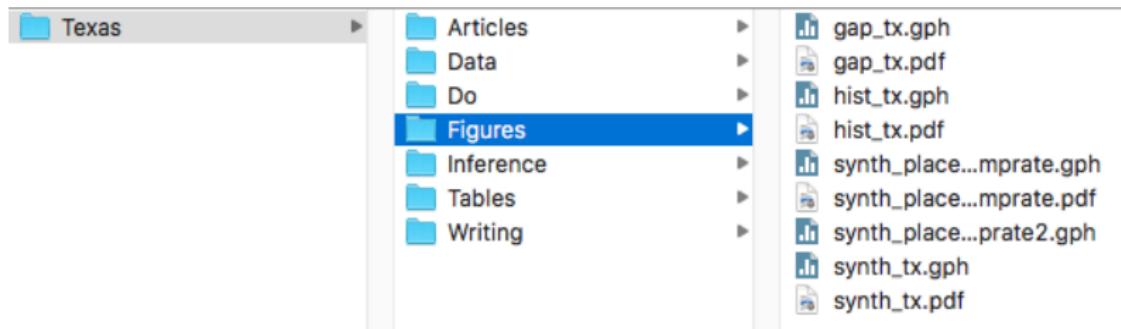
- 3) Data subdirectory containing all datasets

Subdirectory organization



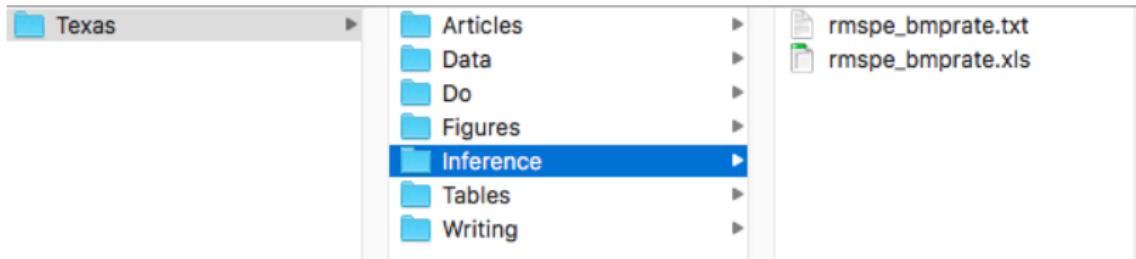
- 4) A subdirectory for all do files and log files

Subdirectory organization



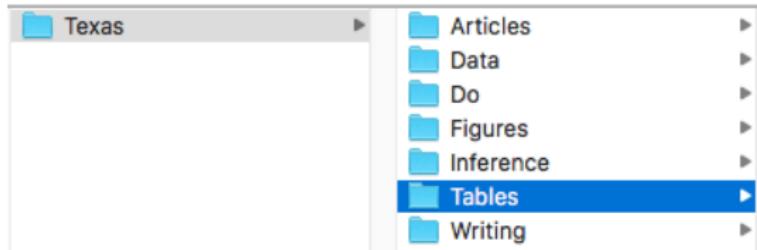
- 5) All figures produced by Stata or image files

Subdirectory organization



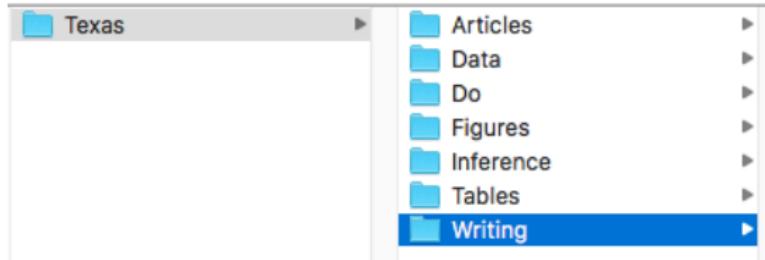
- 6) Project-specific heterogeneity (e.g., “Inference”, “Grants”, “Interview notes”, “Presentations”, “Misc”)

Subdirectory organization



- 7) All tables generated by Stata (e.g., .tex tables produced by -estout-)

Subdirectory organization



- 8) A subdirectory reserved only for writing

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

- Workflow workflow workflow
- Directories
- Do files and R programs
- Naming conventions
- Versional control
- Soft skills

Always use scripting programs NOT GUI

- Guess what - your future self doesn't even remember making do files, tables or figures, let alone typing into GUI command line
- Therefore throw her a bone, hold her hand and walk her exactly through everything
- Which means you've got to have replicable scripting files*
 - * Sure, sometimes use the command line for messing around
 - But then put that messing around in the program

Good text editor

- Remember: the goal is to make beautiful programs
- Invest in a good text editor which has bundling capabilities that will integrate with Stata, R or LaTeX
- I use Textmate 2 because I use a Mac and in addition to a Stata and R bundle, it also allows for *column* editing
- PC users tend to love Sublime for the same reasons
- Stata and Rstudio also come with built-in text editors, which use slick colors for various types of programming commands

Headers

```
*****  
* name: texas.do  
* author: scott cunningham (baylor university)  
* description: estimates the causal effect of prison capacity  
*           expansion on incarceration rates using synth  
* date: march 19, 2018  
*****
```

Speak clearly

“Be conservative in what you do; be liberal in what you accept from others.” - Jon Postel

- Smart sounding quote about both programming and relationships
- Your future self is time constrained, so explain *everything* to her as well as write clear code
- Optimally document your programs
- But speak your future self's love language so she understands

Automating Tables and Figures

- Your goal is to make “beautiful tables” that are never edited post-production as well as readable on their own
- Large fixed costs learning commands like `-estout-` or `-outreg2-`: incur them bc marginal costs are zero
- I use `-estout-` because Jann has written an excellent help file at http://repec.org/bocode/e/estout/hlp_esttab.html but many like `-outreg2-`
- Learn `-twoway-` and/or `-ggplot2-` and make “beautiful pictures” too

Different elements

- When I found my error, and after I regained my exposure, I eventually developed a system of naming
 - ① variables,
 - ② datasets, and
 - ③ do files
- As these are the three things you repeatedly use, you need to have a system, even if not mine

Naming conventions for variables

- Variables should be readable to a stranger
 - Say that you want to create the product of two variables.
Name it the two variables with an underscore
 - `gen price_mpg = price * mpg`
- Otherwise name the variable exactly what it is
 - `gen bmi = weight / (height^2 * 703)`
- Avoid meaningless words (e.g., `lmb2`), dating (e.g., `temp05012020`) and numbering (e.g., `outcome25`) as your future self will be confused

Naming datasets and do files

- The overarching goal is always to name things so that a stranger seeing them can know what they are
- One day you will be the stranger on your own project! Make it easy on your future self!
- Choose some combination of simplicity and clarity but whatever you do, be consistent
- Avoid numbering datasets unless the numbers correspond to some meaningful thing, like randomization inference where each file is a set of coefficients and numbered according to FIPS index

Hidden curriculum

- Foundational causality stuff
- Regression discontinuity designs
- Instrumental variables
- Two-way fixed effects estimator
- Difference-in-differences
- Comparative case studies
- Matching and weighting
- Concluding remarks

Introduction

- Workflow workflow workflow
- Directories
- Do files and R programs
- Naming conventions
- Versional control**
- Soft skills

Version control

- You need a system that allows you to revert to earlier versions without having ten billion files with names like `prison_03102019_sc.do`, etc.
- Most popular these days is git
- Dropbox has version history and is great for coauthoring (as is GitHub), version control is a little sketchier (version history)
- I'm slowly learning git and use github; there's software to make it easier

Selling your work

- If you don't advocate for your work, *no one will*.
- Network, network, network
- You will need to become an expert in 1.5 areas, and you will need experts in those 1.5 areas to agree
- Study the effective of rhetoric of successful economists who expertly communicate their work to others both in their writing of the actual manuscript, as well as the presentation and promotion of their work

Find your mentors and sponsors

- Working with senior people at some point becomes necessary
- Good news: many senior people want to help you
- Bad news: they don't know who you are and can't find you
- It's a two sided matching problem

Finding them

- Introduce yourself in socially appropriate ways!
- This is likely a friction for URM and females given their under-representation in the profession, and prejudices more generally
- Find allies; work networks; I'll be a resource if you need me

Al Roth story

- I wrote Al Roth in 2007 and like Robert Browning to Elizabeth Barrett introduced myself by saying "I love your book on twosided matching with Sotomayor with all my heart."
- We became pen pals and then he won the Nobel Prize
- Scared, I wrote to congratulate him on the day he won and he immediately asked to help me
- "Interpersonal favors are meant to be paid forward not backwards" - Roth to me after a *second* favor!
- Nobody can help you if you don't know them bc help, sponsorship and mentoring is a two sided matching problem

More readings

- I've put several deck of slides and helpful articles for you in the dropbox folder
- Jesse Shapiro's "How to Present an Applied Micro Paper"
- Gentzkow and Shapiro's coding practices manual
- Rachael Meager on presenting as an academic
- Ljubica "LJ" Ristovska's language agnostic guide to programming for economists
- Grant McDermott on Version Control using Github
<https://raw.githubusercontent.com/uo-ec607/lectures/master/02-git/02-Git.html#1>

Data Visualization

Every project should present compelling graphics summarizing the main results and main takeaway

- Study other people's pictures and get help from experts
 - ➊ Kieran Healy's 2018 Visualization: A Practical Introduction (Princeton University Press); free version is <http://socviz.co/index.html#preface>.
 - ➋ Ed Tufte's book Visual display of quantitative information is classic, but more a coffee table book plus no programming assistance.
- Learn Stata's -twoway- capabilities and/or R's -ggplot2-

Hidden curriculum
Foundational causality stuff
Regression discontinuity designs
Instrumental variables
Twoway fixed effects estimator
Difference-in-differences
Comparative case studies
Matching and weighting
Concluding remarks

Regression review
Potential outcomes
Randomization and selection bias
Randomization inference
Causal models and Directed Acyclical Graphs

Introduction: OLS Review

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

Foundations of scientific knowledge

Scientific methodologies are the epistemological foundation of scientific knowledge

- Science does not collect evidence in order to “prove” what people already believe or want others to believe
- Science accepts unexpected and even undesirable answers
- Science is process oriented, not outcome oriented

Terminology

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

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

We said we must confront three issues:

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

We will argue that the simple regression model

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

addresses each of them.

Simple linear regression model

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

The error term

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

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

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

The intercept

The presence of β_0 in

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

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

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

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

Mean independence of the error term

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

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

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

Distribution of ability across education

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

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

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

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

Zero conditional mean assumption

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

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

Population regression function

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

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

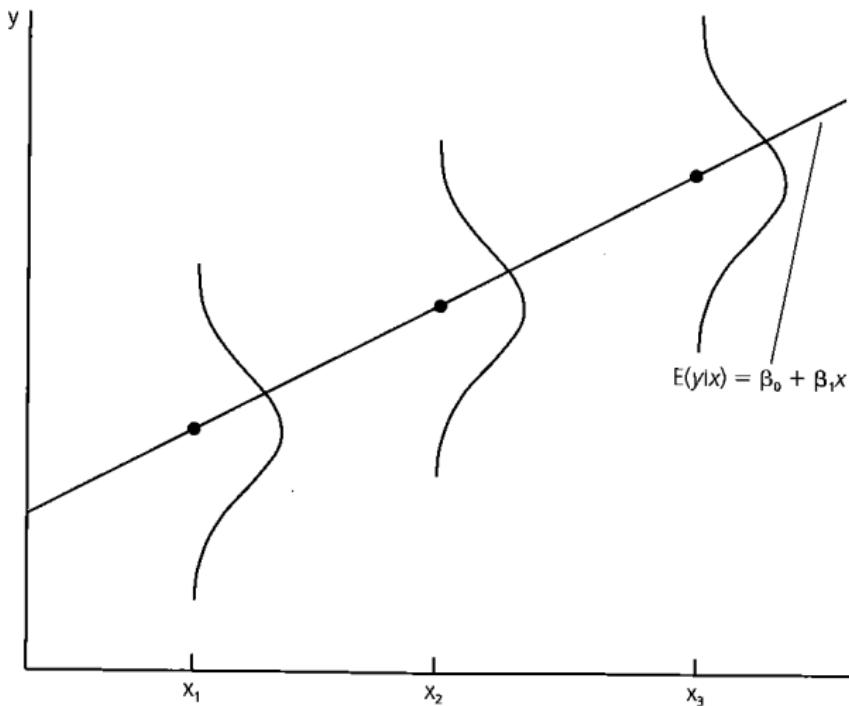
which shows the **population regression function** is a linear function of x .

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

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

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

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



Deriving the Ordinary Least Squares Estimates

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

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

where the i subscript indicates a particular observation.

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

We use the two population restrictions:

$$E(u) = 0$$

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

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

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

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

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

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

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

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

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

Pass the summation operator through the first equation:

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

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

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

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

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

We have shown that the first equation,

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

implies

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

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

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

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

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

so

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

Simple algebra gives

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

So, the equation to solve is

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

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

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

OLS

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

Solving for $\hat{\beta}$

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

Predicting y

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

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

We have n of these.

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

The residual

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

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

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

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

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

Algebraic Properties of OLS Statistics

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

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

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

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

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

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

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

Second moment

Similarly the way we obtained our estimates,

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

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

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

Bringing things together

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

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

Averages

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

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (28)$$

Again, we chose the estimates to make this true.

Expected Value of OLS

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

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

Don't forget why we're here

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

Uncertainty and sampling variance

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

Assumptions

Assumption SLR.1 (Linear in Parameters)

- The population model can be written as

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

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

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

Assumption SLR.2 (Random Sampling)

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

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

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

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

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

Assumption SLR.4 (Zero Conditional Mean)

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

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

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

Showing OLS is unbiased

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

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

where the expected value means averaging across random samples.

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

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

which is one of several equivalent forms.

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

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

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

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

The numerator becomes

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

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

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

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

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

We have shown

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

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

Now define

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

so we have

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM (Unbiasedness of OLS)

Under Assumptions SLR.1 through SLR.4

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

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

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

Reminder

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

SE and the data

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

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

Variance of the OLS Estimators

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

Assumption SLR.5 (Homoskedasticity, or Constant Variance)

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

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

where σ^2 is (virtually always) unknown.

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

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

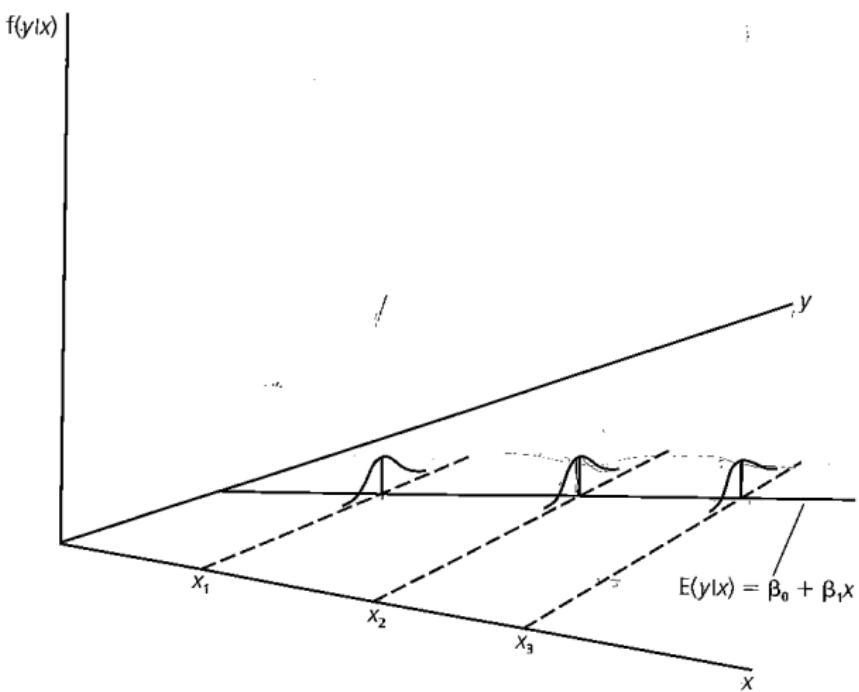
Under the population Assumptions SLR.1 ($y = \beta_0 + \beta_1x + u$),
SRL.4 ($E(u|x) = 0$) and SLR.5 ($Var(u|x) = \sigma^2$),

$$E(y|x) = \beta_0 + \beta_1x$$
$$Var(y|x) = \sigma^2$$

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

Figure 2.8

The simple regression model under homoskedasticity.



THEOREM (Sampling Variances of OLS)

Under Assumptions SLR.1 to SLR.2,

$$\begin{aligned}Var(\hat{\beta}_1|x) &= \frac{\sigma^2}{\sum_{i=1}^n(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x} \\Var(\hat{\beta}_0|x) &= \frac{\sigma^2(n^{-1}\sum_{i=1}^nx_i^2)}{SST_x}\end{aligned}$$

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

To show this, write, as before,

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

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

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

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

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

Now we have

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

We have shown

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

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

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

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

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

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

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

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

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

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

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

Estimating the Error Variance

In the formula

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

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

Recall that

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

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

$$n^{-1} \sum_{i=1}^n u_i^2 \tag{57}$$

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

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

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

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

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

for any i .

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

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

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

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

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

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

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

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

There is no such restriction on the unobserved errors.

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

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

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

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

In regression output, it is

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

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

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

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

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

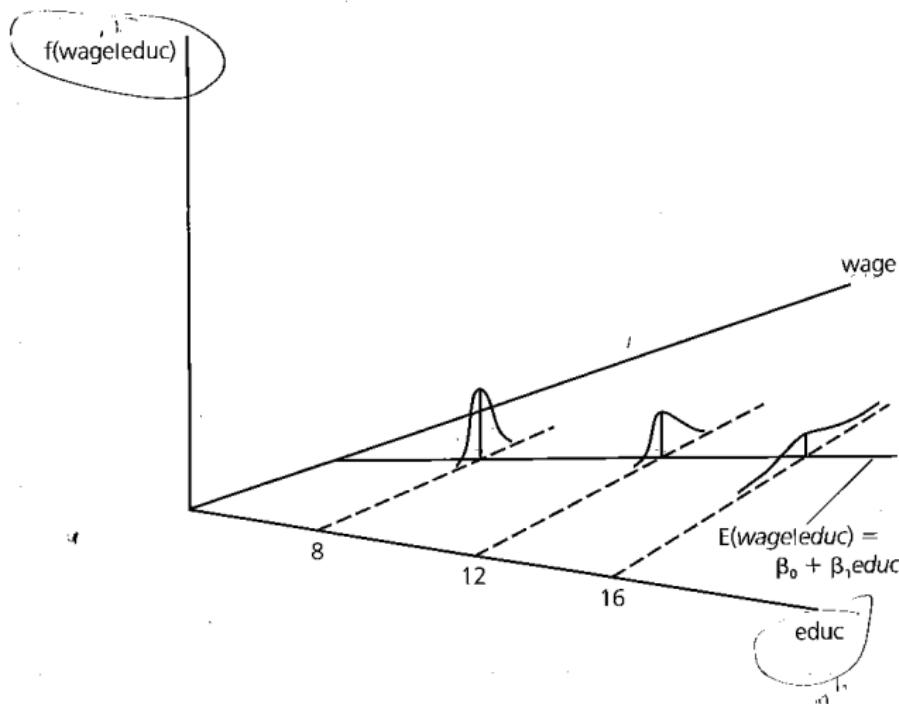
where both the numerator and denominator are computed from the data.

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

- OLS inference is generally faulty in the presence of heteroskedasticity

Figure 2.9

$\text{Var } (\underline{\text{wage}}|\text{educ})$ increasing with educ .



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

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

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

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

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

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

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

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

which is easily computed from the data after the OLS regression

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

Clustered data

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

Simulations

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

Least squares estimates of non-clustered data

Monte Carlo simulation of the slope

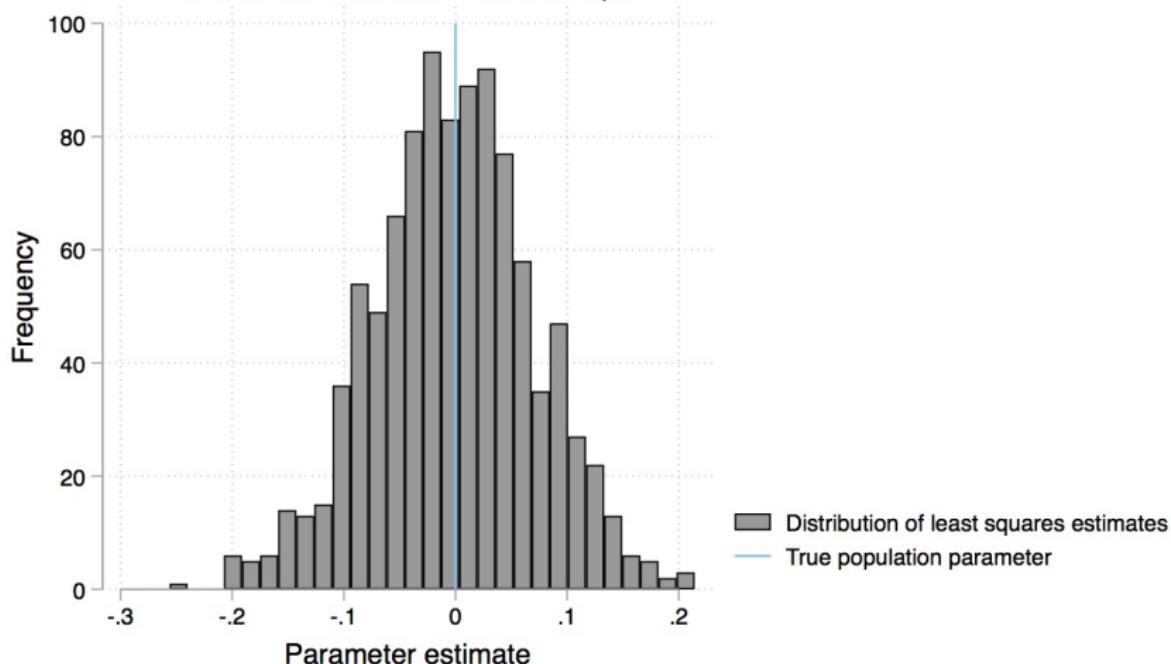


Figure: Distribution of the least squares estimator over 1,000 random draws.

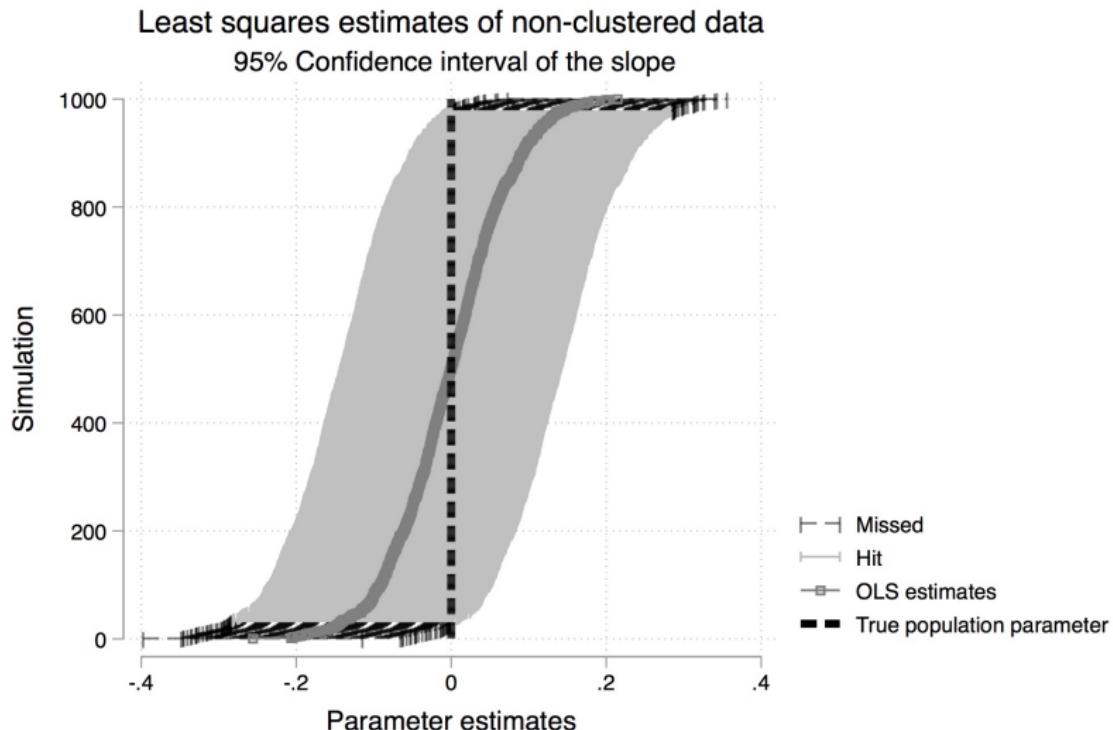


Figure: Distribution of the 95% confidence intervals with coloring showing those which are incorrectly rejecting the null.

Clustered data and heteroskedastic robust

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

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

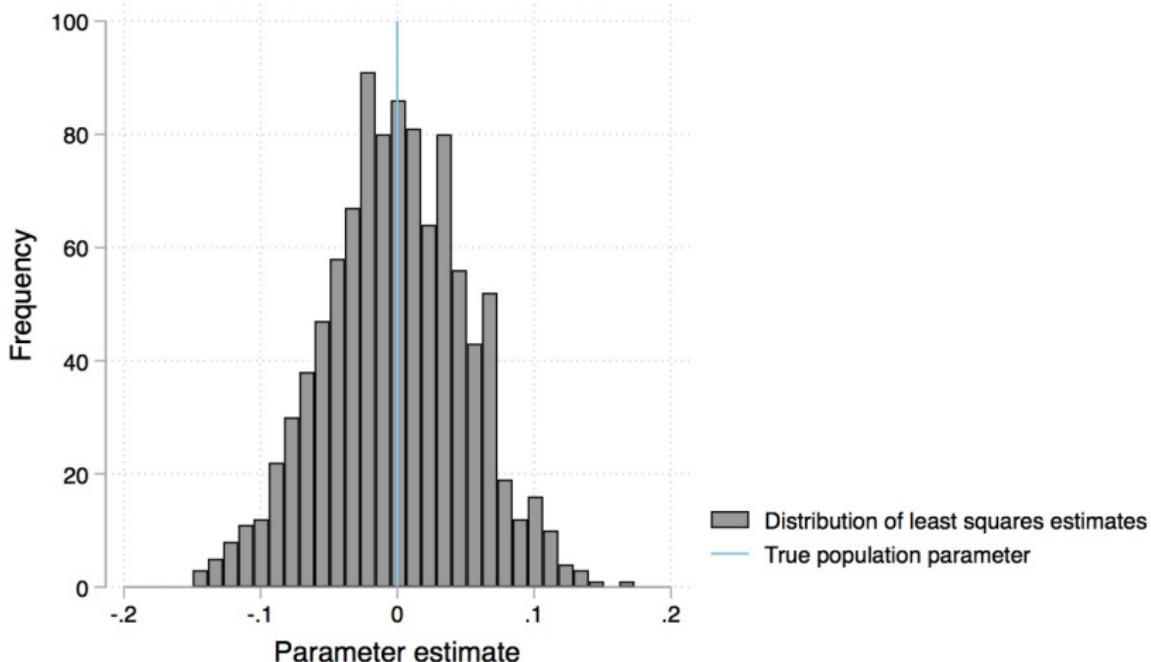


Figure: Distribution of the least squares estimator over 1,000 random draws. Clustered data without correcting for clustering

Least squares estimates of clustered data

95% Confidence interval of the slope

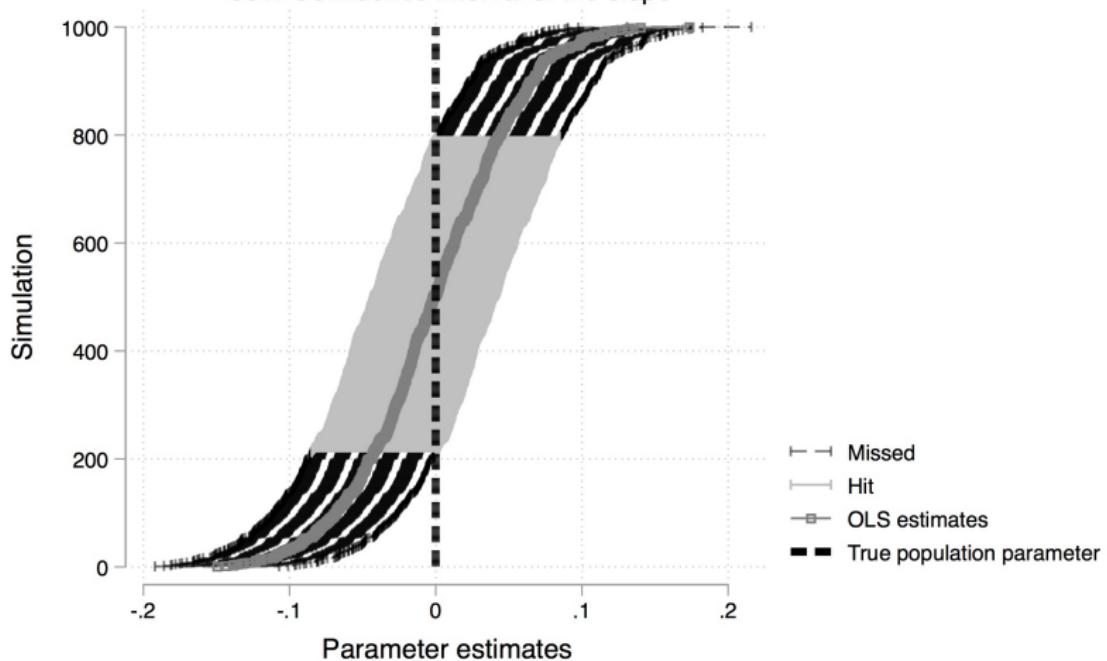


Figure: Distribution of 1,000 95% confidence intervals with dashed region representing those estimates that incorrectly reject the null.

Over-rejecting the null

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

Robust least squares estimates of clustered data

95% Confidence interval of the slope

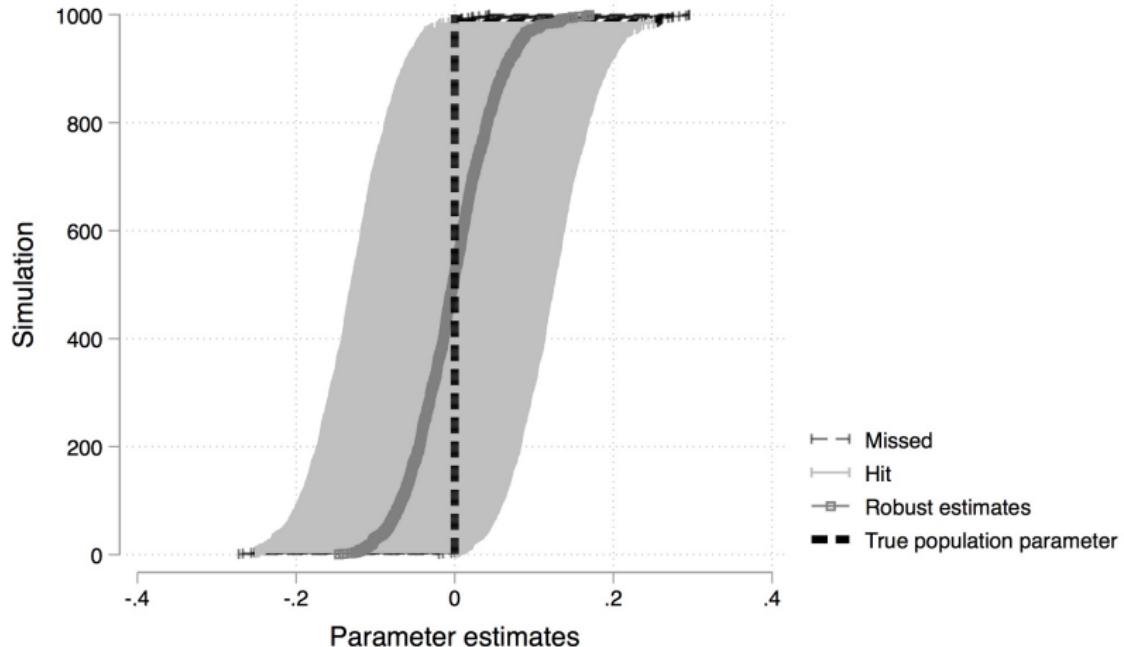


Figure: Distribution of 1,000 95% confidence intervals from a cluster robust least squares regression with dashed region representing those estimates that incorrectly reject the null.

Cluster robust standard errors

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

Clustered data

- Let's stack the observations by cluster

$$y_g = x_g \beta + u_g$$

- The OLS estimator of β is:

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

- The variance is given by:

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

Clustered data

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

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

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

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

The importance of knowing your data

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

Foundations of scientific knowledge

- Scientific methodologies are the epistemological foundation of **scientific knowledge**, which is a particular kind of knowledge
- Science **does not** collect evidence in order to “prove” what people already believe or want others to believe.
- Science is **process oriented**, not **outcome oriented**.
- Therefore science allows us to accept unexpected and sometimes even undesirable answers.

My strong pragmatic claim

- “Credible” causal inference is essential to scientific discovery, publishing and **your career**
- Non-credibly identified empirical micro papers, even ones with ingenious theory, will have trouble getting published and won’t be taken seriously
- Causal inference in 2019 is a necessary, not a sufficient, condition

Outline

- Properties of the conditional expectation function (CEF)
- Reasons for using linear regression
- Regression anatomy theorem
- Omitted variable bias

Properties of the conditional expectation function

- Assume we are interested in the returns to schooling in a wage regression.
- We can summarize the predictive power of schooling's effect on wages with the **conditional expectation function**

$$E(y_i|x_i) \tag{64}$$

- The CEF for a dependent variable, y_i , given covariates X_i , is the expectation, or population average, of y_i with x_i held constant.

- $E(y_i|x_i)$ gives the expected value of y for given values of x
- It provides a reasonable representation of how y changes with x
- If x is random, then $E(y_i|x_i)$ is a random function

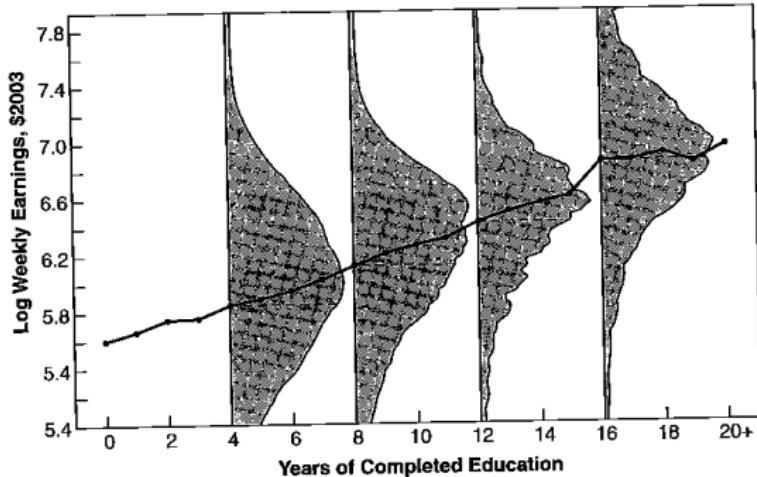


Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

- When there are only two values that x_i can take on, then there are only two values the CEF can take on – but the dummy variable is a special case
- We're often interested in CEFs that are functions of many variables, conveniently subsumed in the vector x_i , and for a specific value of x_i , we will write

$$E(y_i|x_i = x)$$

Helpful result: Law of Iterated Expectations

Definition of Law of Iterated Expectations (LIE)

The unconditional expectation of a random variable is equal to the expectation of the conditional expectation of the random variable conditional on some other random variable

$$E(Y) = E(E[Y|X])$$

We use LIE for a lot of stuff, and it's actually quite intuitive. You may even know it and not know you know it!

Simple example of LIE

- Say you want to know average IQ but only know average IQ by gender.
- LIE says we get the former by taking conditional expectations by gender and combining them (properly weighted)

$$\begin{aligned} E[IQ] &= E(E[IQ|Sex]) \\ &= \sum_{Sex_i} Pr(Sex_i) \cdot E[IQ|Sex_i] \\ &= Pr(Male) \cdot E[IQ|Male] \\ &\quad + Pr(Female) \cdot E[IQ|Female] \end{aligned}$$

- In words: the weighted average of the conditional averages is the unconditional average.

Person	Gender	IQ
1	M	120
2	M	115
3	M	110
4	F	130
5	F	125
6	F	120

- $E[IQ] = 120$
- $E[IQ | \text{Male}] = 115; E[IQ | \text{Female}] = 125$
- LIE: $E(E[IQ | \text{Sex}]) = (0.5) \times 115 + (0.5) \times 125 = 120$

Proof.

For the continuous case:

$$\begin{aligned} E[E(Y|X)] &= \int E(Y|X = u)g_x(u)du \\ &= \int \left[\int tf_{y|x}(t|X = u)dt \right] g_x(u)du \\ &= \int \int tf_{y|x}(t|X = u)g_x(u)dudt \\ &= \int t \left[\int f_{y|x}(t|X = u)g_x(u)du \right] dt \\ &= \int t [f_{x,y} du] dt \\ &= \int tg_y(t)dt \\ &= E(y) \end{aligned}$$



Proof.

For the discrete case,

$$\begin{aligned} E(E[Y|X]) &= \sum_x E[Y|X=x]p(x) \\ &= \sum_x \left(\sum_y yp(y|x) \right) p(x) \\ &= \sum_x \sum_y yp(x,y) \\ &= \sum_y y \sum_x p(x,y) \\ &= \sum_y yp(y) \\ &= E(Y) \end{aligned}$$



Property 1: CEF Decomposition Property

The CEF Decomposition Property

$$y_i = E(y_i|x_i) + u_i$$

where

- ① u_i is mean independent of x_i ; that is

$$E(u_i|x_i) = 0$$

- ② u_i is uncorrelated with any function of x_i

In words: Any random variable, y_i , can be decomposed into two parts: the part that can be explained by x_i and the part left over that can't be explained by x_i . Proof is in Angrist and Pischke (ch. 3)

Property 2: CEF Prediction Property

The CEF Prediction Property

Let $m(x_i)$ be any function of x_i . The CEF solves

$$E(y_i|x_i) = \arg \min_{m(x_i)} E[(y_i - m(x_i))^2].$$

In words: The CEF is the minimum mean squared error predictor of y_i given x_i . Proof is in Angrist and Pischke (ch. 3)

3 reasons why linear regression may be of interest

Linear regression may be interesting even if the underlying CEF is not linear. We review some of the linear theorems now. These are merely to justify the use of linear models to approximate the CEF.

The Linear CEF Theorem

Suppose the CEF is linear. Then the population regression is it.

Comment: Trivial theorem imho because if the population CEF is linear, then it makes the most sense to use linear regression to estimate it. Proof in Angrist and Pischke (ch. 3). Proof uses the CEF Decomposition Property from earlier.

The Best Linear Predictor Theorem

- ① The CEF, $E(y_i|x_i)$, is the minimum mean squared error (MMSE) predictor of y_i given x_i in the class of all functions x_i by the CEF prediction property
- ② The population regression function, $E(x_iy_i)E(x_ix_i')^{-1}$, is the best we can do in the class of all linear functions

Proof is in Angrist and Pischke (ch. 3).

The Regression CEF Theorem

The function $x_i\beta$ provides the minimum mean squared error (MMSE) linear approximation to $E(y_i|x_i)$, that is

$$\beta = \arg \min_b E\{(E(y_i|x_i) - x_i' b)^2\}$$

Again, proof in Angrist and Pischke (ch. 3).

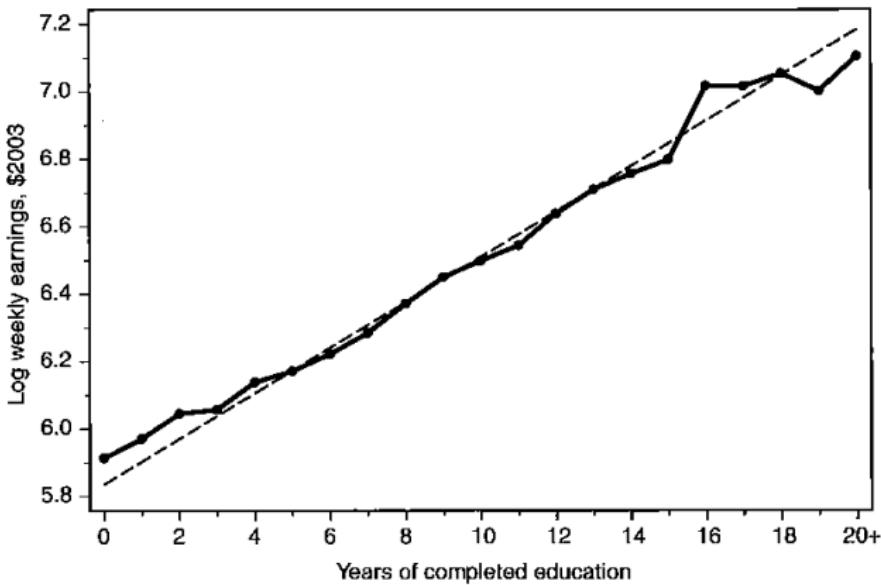


Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Random families

- We are interested in the causal effect of family size on labor supply so we regress labor supply onto family size

$$labor_supply_i = \beta_0 + \beta_1 numkids_i + \varepsilon_i$$

- If couples had kids by flipping coins, then $numkids_i$ independent of ε_i , then estimation is simple - just compare families with different sizes to get the causal effect of $numkids$ on $labor_supply$
- But how do we interpret $\hat{\beta}_1$ if families don't flip coins?

Non-random families

- If family size is random, you could visualize the causal effect with a scatter plot and the regression line
- If family size is non-random, then we can't do this because we need to control for multiple variables just to remove the factors causing family size to be correlated with ε

Non-random families

- Assume that family size is random once we condition on race, age, marital status and employment.

$$\text{labor_supply}_i = \beta_0 + \beta_1 \text{Numkids}_i + \gamma_1 \text{White}_i + \gamma_2 \text{Married}_i + \gamma_3 \text{Age}_i + \gamma_4 \text{Employed}_i + \varepsilon_i$$

- To estimate this model, we need:
 - a data set with all 6 variables;
 - Numkids must be randomly assigned conditional on the other 4 variables
- Now how do we interpret $\hat{\beta}_1$? And can we visualize $\hat{\beta}_1$ when there's multiple dimensions to the data? Yes, using the regression anatomy theorem, we can.

Regression Anatomy Theorem

Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

and an auxiliary regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i$$

and $\tilde{x}_{1i} = x_{1i} - \hat{x}_{1i}$ being the residual from the auxiliary regression.
The parameter β_1 can be rewritten as:

$$\beta_1 = \frac{Cov(y_i, \tilde{x}_{1i})}{Var(\tilde{x}_{1i})}$$

In words: The regression anatomy theorem says that $\hat{\beta}_1$ is a scaled covariance with the \tilde{x}_1 residual used instead of the actual data x .

Regression Anatomy Proof

To prove the theorem, note $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\hat{x}_{ki}] = E[f_i]$, and plug y_i and residual \tilde{x}_{ki} from x_{ki} auxiliary regression into the covariance $\text{cov}(y_i, \tilde{x}_{ki})$

$$\begin{aligned}\beta_k &= \frac{\text{cov}(y_i, \tilde{x}_{ki})}{\text{var}(\tilde{x}_{ki})} \\ &= \frac{\text{cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{\text{var}(\tilde{x}_{ki})} \\ &= \frac{\text{cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{\text{var}(f_i)}\end{aligned}$$

- ① Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$.
- ② Since f_i is a linear combination of all the independent variables with the exception of x_{ki} , it must be that

$$\beta_1 E[f_i x_{1i}] = \cdots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \cdots = \beta_K E[f_i x_{Ki}] = 0$$

Regression Anatomy Proof (cont.)

- ③ Consider now the term $E[e_i f_i]$. This can be written as:

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i(x_{ki} - \hat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}] \end{aligned}$$

Since e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki} : accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the x_{ki} auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\hat{\gamma}_0 + \hat{\gamma}_1 x_{1i} + \cdots + \hat{\gamma}_{k-1} x_{(k-1)i} + \hat{\gamma}_{k+1} x_{(k+1)i} + \cdots + \hat{\gamma}_K x_{Ki})]$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows $E[e_i f_i] = 0$.

Regression Anatomy Proof (cont.)

- ④ The only remaining term is $E[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki} \tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term x_{ki} can be substituted using a rewriting of the auxiliary regression model, x_{ki} , such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= E[\beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})]] \\ &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}] \tilde{x}_{ki})]\} \\ &= \beta_k \text{var}(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between $E[x_{ki}|X_{-k}]$ and \tilde{x}_{ki} . From previous derivations we finally get

$$\text{cov}(y_i, \tilde{x}_{ki}) = \beta_k \text{var}(\tilde{x}_{ki})$$

which completes the proof. □

Stata command: reganat (i.e., regression anatomy)

```
. ssc install reganat, replace
. sysuse auto
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(weight length) biline
```

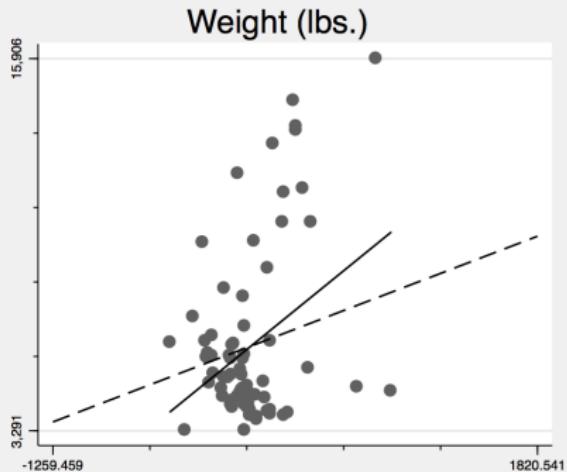
```
. regress price length weight headroom mpg
```

Source	SS	df	MS	Number of obs	=	74
Model	236190226	4	59047556.6	F(4, 69)	=	10.21
Residual	398875170	69	5780799.56	Prob > F	=	0.0000
Total	635065396	73	8699525.97	R-squared	=	0.3719
				Adj R-squared	=	0.3355
				Root MSE	=	2404.3

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
length	-94.49651	40.39563	-2.34	0.022	-175.0836	-13.90944
weight	4.335045	1.162745	3.73	0.000	2.015432	6.654657
headroom	-490.9667	388.4892	-1.26	0.211	-1265.981	284.048
mpg	-87.95838	83.5927	-1.05	0.296	-254.7213	78.80449
_cons	14177.58	5872.766	2.41	0.018	2461.735	25893.43

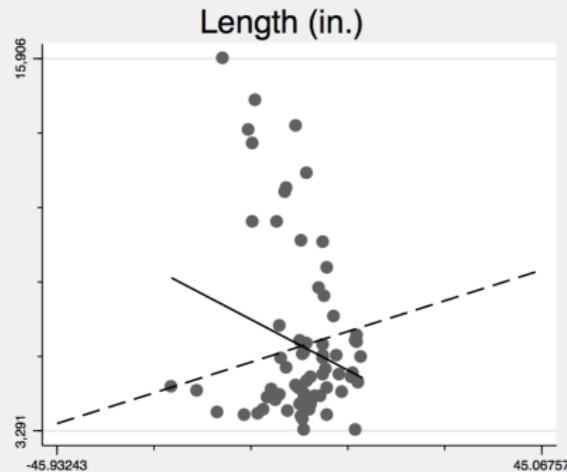
Regression Anatomy

Dependent variable: Price



Multivariate slope: 4.335 (1.163)

Bivariate slope: 2.044 (0.377)



Multivariate slope: -94.497 (40.396)

Bivariate slope: 57.202 (14.080)

Covariates: Length (in.), Weight (lbs.), Headroom (in.), Mileage (mpg).

Regression lines: Solid = Multivariate, Dashed = Bivariate.

Big picture

- ① Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
- ② If we prefer to think of approximating $E(y_i|x_i)$ as opposed to predicting y_i , the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
- ③ Regression anatomy theorem helps us interpret a single slope coefficient in a multiple regression model by the aforementioned decomposition.

Omitted Variable Bias

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

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

Y_i = log of earnings

S_i = schooling measured in years

A_i = individual ability

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

Shorter regression

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

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

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

Derivation of Ability Bias

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

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

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

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

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

Summary

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

Introduction to Counterfactuals and Causality

- Aliens come and orbit earth, see sick people in hospitals and conclude “these ‘hospitals’ are hurting people”
- Motivated by anger and compassion, they kill the doctors to save the patients
- Sounds stupid, but earthlings do this too - all the time
- Let’s look at the challenges of making causality synonymous with correlations

#1: Correlation and causality are very different concepts

- Causal question: “If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?”
- Correlation question:

$$\frac{1}{n} \frac{\text{Cov}(D, Y)}{\sqrt{\text{Var}_D} \sqrt{\text{Var}_Y}}$$

- These are not the same thing

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- *Post hoc ergo propter hoc*: “after this, therefore, because of this”



#3: No correlation does not mean no causality!

- A sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- The same aliens observe from space and say “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken.” So they send her a new rudder
- They’re wrong but why are they wrong? There is, after all, no correlation

Potential outcomes notation

- Let the treatment be a binary variable:

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

where i indexes an individual observation, such as a person

- Potential outcomes:

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

where j indexes a counterfactual state of the world

Moving between worlds

- I'll drop t subscript, but note – these are potential outcomes for the same person at the exact same moment in time
- A potential outcome Y^1 is not the historical outcome Y either conceptually or notationally
- Potential outcomes are hypothetical states of the world but historical outcomes are ex post realizations
- Major philosophical move here: go from the potential worlds to the actual (historical) world based on your treatment assignment

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i , equals $Y_i^1 - Y_i^0$

Definition 3: Switching equation

An individual's observed health outcomes, Y , is determined by treatment assignment, D_i , and corresponding potential outcomes:

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

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

So what's the problem?

Definition 4: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i , is *unknowable*.

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

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

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

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

Causality and comparisons

- Does the ventilator make someone have severe COVID symptoms? Or are they sick with COVID symptoms, and that's why they are on a ventilator?
- Why can't I just compare symptoms for those on vents versus those who aren't? After all, there's a control group
- What are we actually measuring if we compare average health outcomes for those on vents versus those who aren't?
- Let's look at our first estimator and see if we can't better understand what comparisons are and are not saying.

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by the sample averages:

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

in large samples. I'll usually use expectation operators but we use samples for estimation.

SDO vs. ATE

Notice the subtle difference between the SDO and ATE notation:

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

- The SDO is an *estimate*, whereas ATE is a *parameter*
- SDO is a crank that turns data into numbers
- ATE is a parameter that is unknowable because of the fundamental problem of causal inference
- SDO might line up with the ATE but also might not
- Under what situations is SDO a biased estimate of the ATE?

Potentially biased comparisons

Decomposition of the SDO

The SDO can be decomposed into the sum of three parts:

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

Seeing is believing so let's work through this identity!

Decomposition of SDO

Use LIE to decompose ATE into the sum of four conditional average expectations

$$\begin{aligned}\text{ATE} &= E[Y^1] - E[Y^0] \\ &= \{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &\quad - \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\}\end{aligned}$$

Substitute letters for expectations

$$\begin{aligned}E[Y^1|D = 1] &= a \\ E[Y^1|D = 0] &= b \\ E[Y^0|D = 1] &= c \\ E[Y^0|D = 0] &= d \\ \text{ATE} &= e\end{aligned}$$

Rewrite ATE

$$e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$$

Move SDO terms to LHS

$$e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$$

$$e = \pi a + b - \pi b - \pi c - d + \pi d$$

$$e = \pi a + b - \pi b - \pi c - d + \pi d + (a - a) + (c - c) + (d - d)$$

$$0 = e - \pi a - b + \pi b + \pi c + d - \pi d - a + a - c + c - d + d$$

$$a - d = e - \pi a - b + \pi b + \pi c + d - \pi d + a - c + c - d$$

$$a - d = e + (c - d) + a - \pi a - b + \pi b - c + \pi c + d - \pi d$$

$$a - d = e + (c - d) + (1 - \pi)a - (1 - \pi)b + (1 - \pi)d - (1 - \pi)c$$

$$a - d = e + (c - d) + (1 - \pi)(a - c) - (1 - \pi)(b - d)$$

Substitute conditional means

$$\begin{aligned} E[Y^1|D=1] - E[Y^0|D=0] &= \text{ATE} \\ &\quad + (E[Y^0|D=1] - E[Y^0|D=0]) \\ &\quad + (1 - \pi)(\{E[Y^1|D=1] - E[Y^0|D=1]\}) \\ &\quad - (1 - \pi)\{E[Y^1|D=0] - E[Y^0|D=0]\}) \end{aligned}$$

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

Decomposition of difference in means

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

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

Independence assumption

Independence assumption

Treatment is independent of potential outcomes

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

In plain language: Random assignment means the treatment has been assigned to units without regard to their potential outcomes. This ensures that mean potential outcomes for the treatment group and control group are the same. Also ensures other variables are distributed the same for a large sample.

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

Random Assignment Solves the Selection Problem

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

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

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

Random Assignment Solves the Heterogenous Treatment Effects

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

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

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

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

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

- If treatment is independent of potential outcomes, then:

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

SUTVA

- Potential outcomes model places a limit on what we can measure: the “stable unit-treatment value assumption”. Horrible acronym.
 - ① **S**: *stable*
 - ② **U**: across all *units*, or the population
 - ③ **TV**: *treatment-value* (“treatment effect”, “causal effect”)
 - ④ **A**: *assumption*
- SUTVA means that average treatment effects are parameters that assume (1) homogenous dosage, (2) potential outcomes are invariant to who else (and how many) is treated (e.g., externalities), and (3) partial equilibrium

SUTVA: (1) Homogenous dose

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable”
- If we are estimating the effect of vents on covid symptoms, we assume everyone is getting the same kinds of vents more or less.
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*

SUTVA: (2) No spillovers to other units

- What if putting someone on a ventilator causes someone else to be more or less likely to develop severe covid symptoms?
- Have to think hard about externalities, particularly with transmissible diseases
- SUTVA means that you don't have a problem like this.
- If there are no externalities from treatment, then δ_i is stable for each i unit regardless of whether someone else receives the treatment too, but herd immunity must be considered when it comes to cures

SUTVA: (3) Partial equilibrium only

Easier to imagine this with a different example.

- Let's say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?
- Scaling up a policy can be challenging to predict if there are rising costs of production
- What if expansion requires hiring lower quality teachers just to make classes?
- That's a general equilibrium effect; we only estimated a partial equilibrium effect (external versus internal validity)

Demand for Learning HIV Status

- Rebecca Thornton implemented an RCT in rural Malawi for her job market paper at Harvard in mid-2000s
- At the time, it was an article of faith that you could fight the HIV epidemic in Africa by encouraging people to get tested; but Thornton wanted to see if this was true
- She randomly assigned cash incentives to people to incentivize learning their HIV status
- Also examined whether learning changed sexual behavior.

Experimental design

- Respondents were offered a free door-to-door HIV test
- Treatment is randomized vouchers worth between zero and three dollars
- These vouchers were redeemable once they visited a nearby voluntary counseling and testing center (VCT)
- Estimates her models using OLS with controls

Why Include Control Variables?

- To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the prior equation, we might estimate:

$$Y_i = \alpha + \delta D_i + \gamma X_i + \eta_i$$

- There are 2 main reasons for including additional controls in the regression models:
 - ① Conditional random assignment. Sometimes randomization is done *conditional* on some observable (e.g., gender, school, districts)
 - ② Exogenous controls increase precision. Although control variables X_i are uncorrelated with D_i , they may have substantial explanatory power for Y_i . Including controls thus reduces variance in the residuals which lowers the standard errors of the regression estimates.

Table: Impact of Monetary Incentives and Distance on Learning HIV Results

	1	2	3	4	5
Any incentive	0.431*** (0.023)	0.309*** (0.026)	0.219*** (0.029)	0.220*** (0.029)	0.219 *** (0.029)
Amount of incentive		0.091*** (0.012)	0.274*** (0.036)	0.274*** (0.035)	0.273*** (0.036)
Amount of incentive ²			-0.063*** (0.011)	-0.063*** (0.011)	-0.063*** (0.011)
HIV	-0.055* (0.031)	-0.052 (0.032)	-0.05 (0.032)	-0.058* (0.031)	-0.055* (0.031)
Distance (km)				-0.076*** (0.027)	
Distance ²				0.010** (0.005)	
Controls	Yes	Yes	Yes	Yes	Yes
Sample size	2,812	2,812	2,812	2,812	2,812
Average attendance	0.69	0.69	0.69	0.69	0.69

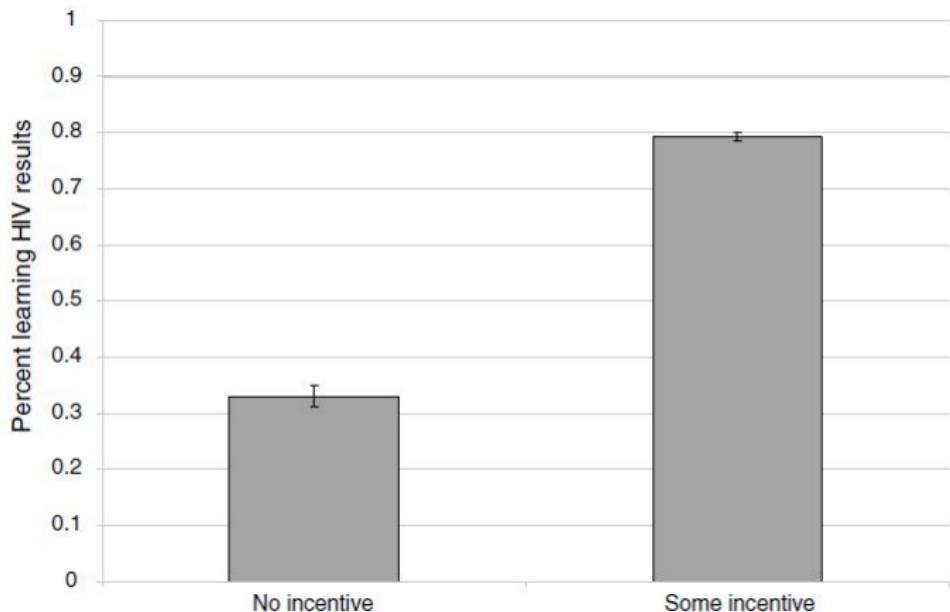


Figure: Visual representation of cash transfers on learning HIV test results.

Results

- Even small incentives were effective
- Any incentive increases learning HIV status by 43% compared to the control (mean 34%)
- Next she looks at the effect that learning HIV status has on risky sexual behavior

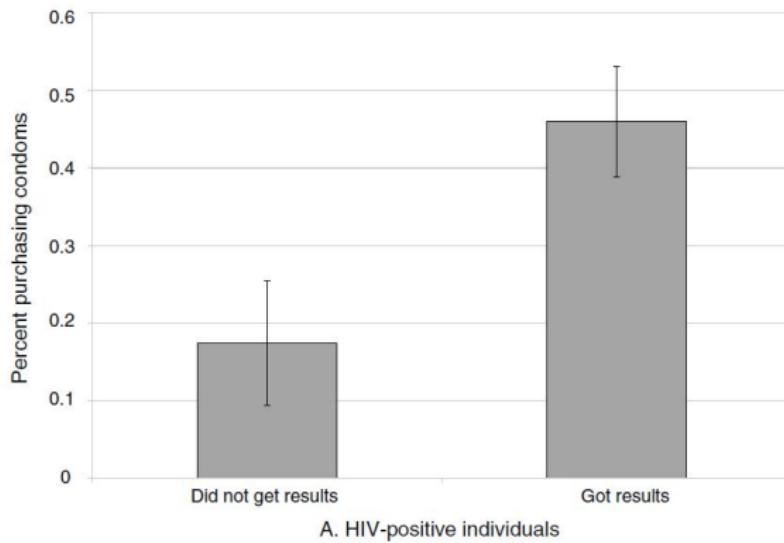


Figure: Visual representation of cash transfers on condom purchases for HIV positive individuals.

Table: Reactions to Learning HIV Results among Sexually Active at Baseline

Dependent variables:	Bought condoms		Number of condoms bought	
	OLS	IV	OLS	IV
Got results	-0.022 (0.025)	-0.069 (0.062)	-0.193 (0.148)	-0.303 (0.285)
Got results × HIV	0.418*** (0.143)	0.248 (0.169)	1.778*** (0.564)	1.689** (0.784)
HIV	-0.175** (0.085)	-0.073 (0.123)	-0.873 (0.275)	-0.831 (0.375)
Controls	Yes	Yes	Yes	Yes
Sample size	1,008	1,008	1,008	1,008
Mean	0.26	0.26	0.95	0.95

Results

- For those who were HIV+ and got their test results, 42% more likely to buy condoms (but shrinks and becomes insignificant at conventional levels with IV).
- Number of condoms bought – very small. HIV+ respondents who learned their status bought 2 more condoms

Randomization inference and causal inference

- “In randomization-based inference, uncertainty in estimates arises naturally from the random assignment of the treatments, rather than from hypothesized sampling from a large population.” (Athey and Imbens 2017)
- Athey and Imbens is part of growing trend of economists using randomization-based methods for doing causal inference
- Unclear (to me) why we are hearing more and more about randomization inference, but we are.
- Could be due to improved computational power and/or the availability of large data instead of samples?

Lady tasting tea experiment

- Ronald Aylmer Fisher (1890-1962)
 - Two classic books on statistics: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), as well as a famous work in genetics, *The Genetical Theory of Natural Science*
 - Developed many fundamental notions of modern statistics including the theory of randomized experimental design.

Lady tasting tea

- Muriel Bristol (1888-1950)
 - A PhD scientist back in the days when women weren't PhD scientists
 - Worked with Fisher at the Rothamsted Experiment Station (which she established) in 1919
 - During afternoon tea, Muriel claimed she could tell from taste whether the milk was added to the cup before or after the tea
 - Scientists were incredulous, but Fisher was inspired by her strong claim
 - He devised a way to test her claim which she passed using randomization inference

Description of the tea-tasting experiment

- Original claim: Given a cup of tea with milk, Bristol claims she can discriminate the order in which the milk and tea were added to the cup
- Experiment: To test her claim, Fisher prepares 8 cups of tea – 4 **milk then tea** and 4 **tea then milk** – and presents each cup to Bristol for a taste test
- Question: How many cups must Bristol correctly identify to convince us of her unusual ability to identify the order in which the milk was poured?
- Fisher's sharp null: Assume she can't discriminate. Then what's the likelihood that random chance was responsible for her answers?

Choosing subsets

- The lady performs the experiment by selecting 4 cups, say, the ones she claims to have had the tea poured first.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- "8 choose 4" – $\binom{8}{4}$ – ways to choose 4 cups out of 8
 - Numerator is $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order.
 - Denominator is $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups.

Choosing subsets

- There are 70 ways to choose 4 cups out of 8, and therefore a 1.4% probability of producing the correct answer by chance

$$\frac{24}{1680} = 1/70 = 0.014.$$

- For example, the probability that she would correctly identify all 4 cups is $\frac{1}{70}$

Statistical significance

- Suppose the lady correctly identifies all 4 cups. Then . . .
 - ① Either she has no ability, and has chosen the correct 4 cups purely by chance, or
 - ② She has the discriminatory ability she claims.
- Since choosing correctly is highly unlikely in the first case (one chance in 70), the second seems plausible.
- Bristol actually got all four correct
- I wonder if seeing this, any of the scientists present changed their mind

Null hypothesis

- In this example, the null hypothesis is the hypothesis that the lady has no special ability to discriminate between the cups of tea.
- We can never prove the null hypothesis, but the data may provide evidence to reject it.
- In most situations, rejecting the null hypothesis is what we hope to do.

Null hypothesis of no effect

- Randomization inference allows us to make probability calculations revealing whether the treatment assignment was “unusual”
- Fisher’s sharp null is when entertain the possibility that no unit has a treatment effect
- This allows us to make “exact” p-values which do not depend on large sample approximations
- It also means the inference is not dependent on any particular distribution (e.g., Gaussian); sometimes called nonparametric

Sidebar: bootstrapping is different

- Sometimes people confuse randomization inference with bootstrapping
- Bootstrapping randomly draws a percent of the total observations for estimation; “uncertainty over the sample”
- Randomization inference randomly reassigns the treatment; “uncertainty over treatment assignment”

(Thanks to Jason Kerwin for helping frame the two against each other)

6-step guide to randomization inference

The following is from Imbens and Rubin's textbook on causal inference, as well as Matthew Blackwell's helpful lectures

- ① Choose a sharp null hypothesis (e.g., no treatment effects)
- ② Calculate a test statistic (T is a scalar based on D and Y)
- ③ Then pick a randomized treatment vector \tilde{D}_1
- ④ Calculate the test statistic associated with (\tilde{D}_1, Y)
- ⑤ Repeat steps 3 and 4 for all possible combinations to get
 $\tilde{T} = \{\tilde{T}_1, \dots, \tilde{T}_K\}$
- ⑥ Calculate exact p-value as $p = \frac{1}{K} \sum_{k=1}^K I(\tilde{T}_k \geq T)$

Pretend experiment

Table: Pretend DBT intervention for some homeless population

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

For concreteness, assume a program where we pay homeless people \$15 to take dialectical behavioral therapy (DBT). Outcomes are some measure of mental health 0-20 with higher scores being improvements in mental health symptoms.

Step 1: Sharp null of no effect

Fisher's Sharp Null Hypothesis

$$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \quad \forall i$$

- Assuming no effect means any test statistic is due to chance
- Neyman and Fisher test statistics were different – Fisher was exact, Neyman was not
- Neyman's null was no average treatment effect ($ATE=0$). If you have a treatment effect of 5 and I have a treatment effect of -5, our ATE is zero. This is not the sharp null even though it also implies a zero ATE

More sharp null

- Since under the Fisher sharp null $\delta_i = 0$, it means each unit's potential outcomes under both states of the world are the same
- We therefore know each unit's missing counterfactual
- The randomization we will perform will cycle through all treatment assignments under a null well treatment assignment doesn't matter because all treatment assignments are associated with a null or zero unit treatment effects
- We are looking for evidence *against* the null

Step 1: Fisher's sharp null and missing potential outcomes

Table: Missing potential outcomes are no longer missing

Name	D	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	1	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Fisher sharp null allows us to **fill in** the missing counterfactuals bc under the null there's zero treatment effect at the unit level. This guarantees zero ATE, but is different in formulation than Neyman's null effect of no ATE.

Step 2: Choosing a test statistic

Test Statistic

A test statistic $T(D, Y)$ is a scalar quantity calculated from the treatment assignments D and the observed outcomes Y

- By scalar, I just mean it's a number (vs. a function) measuring some relationship between D and Y
- Ultimately there are many tests to choose from; I'll review a few later
- If you want a test statistic with high statistical power, you need large values when the null is false, and small values when the null is true (i.e., *extreme*)

Simple difference in means

- Consider the absolute SDO from earlier

$$\delta_{SDO} = \left| \frac{1}{N_T} \sum_{i=1}^N D_i Y_i - \frac{1}{N_C} \sum_{i=1}^N (1 - D_i) Y_i \right|$$

- Larger values of δ_{SDO} are evidence *against* the sharp null
- Good estimator for constant, additive treatment effects and relatively few outliers in the potential outcomes

Step 2: Calculate test statistic, $T(D, Y)$

Table: Calculate T using D and Y

Name	D	Y	Y^0	Y^1	δ_i
Andy	1	10	10	10	0
Ben	1	5	5	5	0
Chad	1	16	16	16	0
Daniel	1	3	3	3	0
Edith	0	5	5	5	0
Frank	0	7	7	7	0
George	0	8	8	8	0
Hank	0	10	10	10	0

We'll start with this simple the simple difference in means test statistic, $T(D, Y)$: $\delta_{SDO} = 34/4 - 30/4 = 1$

Steps 3-5: Null randomization distribution

- Randomization steps reassign treatment assignment for every combination, calculating test statistics each time, to obtain the entire distribution of counterfactual test statistics
- The key insight of randomization inference is that under Fisher's sharp null, the treatment assignment shouldn't matter
- Ask yourself:
 - if there is no unit level treatment effect, can you picture a distribution of counterfactual test statistics?
 - and if there is no unit level treatment effect, what must average counterfactual test statistics equal?

Step 6: Calculate “exact” p-values

- Question: how often would we get a test statistic as big or bigger as our “real” one if Fisher’s sharp null was true?
- This can be calculated “easily” (sometimes) once we have the randomization distribution from steps 3-5
 - The number of test statistics ($t(D, Y)$) bigger than the observed divided by total number of randomizations

$$Pr(T(D, Y) \geq T(\tilde{D}, Y | \delta = 0)) = \frac{\sum_{D \in \Omega} I(T(D, Y) \leq T(\tilde{D}, Y))}{K}$$

Approximate p-values

These have been “exact” tests when they use every possible combination of D

- When you can't use every combination, then you can get *approximate* p-values from a simulation (TBD)
- With a rejection threshold of α (e.g., 0.05), randomization inference test will falsely reject less than $100 \times \alpha\%$ of the time

First permutation (holding N_T fixed)

Name	\tilde{D}_2	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

$$\tilde{T}_1 = |36/4 - 28/4| = 9 - 7 = 2$$

Second permutation (again holding N_T fixed)

Name	\tilde{D}_3	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	1	8	8	8
Hank	0	10	10	10

$$T_{rank} = |36/4 - 27/4| = 9 - 6.75 = 2.25$$

Sidebar: Should it be 4 treatment groups each time?

- In this experiment, I've been using the same N_T under the assumption that N_T had been fixed when the experiment was drawn.
- But if the original treatment assignment had been generated by something like a Bernoulli distribution (e.g., coin flips over every unit), then you should be doing a complete permutation that is also random in this way
- This means that for 8 units, sometimes you'd have 1 treated, or even 8
- Correct inference requires you know the original data generating process

Randomization distribution

Step 2: Other test statistics

- The simple difference in means is fine when effects are additive, and there are few outliers in the data
- But outliers create more variation in the randomization distribution
- What are some alternative test statistics?

Transformations

- What if there was a constant multiplicative effect:
$$Y_i^1 / Y_i^0 = C?$$
- Difference in means will have low power to detect this alternative hypothesis
- So we transform the observed outcome using the natural log:

$$T_{log} = \left| \frac{1}{N_T} \sum_{i=1}^N D_i \ln(Y_i) - \frac{1}{N_C} \sum_{i=1}^N (1 - D_i) \ln(Y_i) \right|$$

- This is useful for skewed distributions of outcomes

Difference in medians/quantiles

- We can protect against outliers using other test statistics such as the difference in quantiles
- Difference in medians:

$$T_{median} = |\text{median}(Y_T) - \text{median}(Y_C)|$$

- We could also estimate the difference in quantiles at any point in the distribution (e.g., 25th or 75th quantile)

Rank test statistics

- Basic idea is rank the outcomes (higher values of Y_i are assigned higher ranks)
- Then calculate a test statistic based on the transformed ranked outcome (e.g., mean rank)
- Useful with continuous outcomes, small datasets and/or many outliers

Rank statistics formally

- Rank is the domination of others (including oneself):

$$\tilde{R} = \tilde{R}_i(Y_1, \dots, Y_N) = \sum_{j=1}^N I(Y_j \leq Y_i)$$

- Normalize the ranks to have mean 0

$$\tilde{R}_i = \tilde{R}_i(Y_1, \dots, Y_N) = \sum_{j=1}^N I(Y_j \leq Y_i) - \frac{N+1}{2}$$

- Calculate the absolute difference in average ranks:

$$T_{rank} = |\bar{R}_T - \bar{R}_C| = \left| \frac{\sum_{i:D_i=1} R_i}{N_T} - \frac{\sum_{i:D_i=0} R_i}{N_C} \right|$$

- Minor adjustment (averages) for ties

Randomization distribution

Name	D	Y	Y^0	Y^1	Rank	R_i
Andy	1	10	10	10	6.5	2
Ben	1	5	5	5	2.5	-2
Chad	1	16	16	16	8	3.5
Daniel	1	3	3	3	1	-3.5
Edith	0	5	5	5	2.5	-2
Frank	0	7	7	7	4	-0.5
George	0	8	8	8	5	0.5
Hank	0	10	10	10	6.5	2

$$T_{rank} = |0 - 0| = 0$$

Effects on outcome distributions

- Focused so far on “average” differences between groups.
- Kolmogorov-Smirnov test statistics is based on the difference in the distribution of outcomes
- Empirical cumulative distribution function (eCDF):

$$\hat{F}_C(Y) = \frac{1}{N_C} \sum_{i:D_i=0} 1(Y_i \leq Y)$$

$$\hat{F}_T(Y) = \frac{1}{N_T} \sum_{i:D_i=1} 1(Y_i \leq Y)$$

- Proportion of observed outcomes below a chosen value for treated and control separately
- If two distributions are the same, then $\hat{F}_C(Y) = \hat{F}_T(Y)$

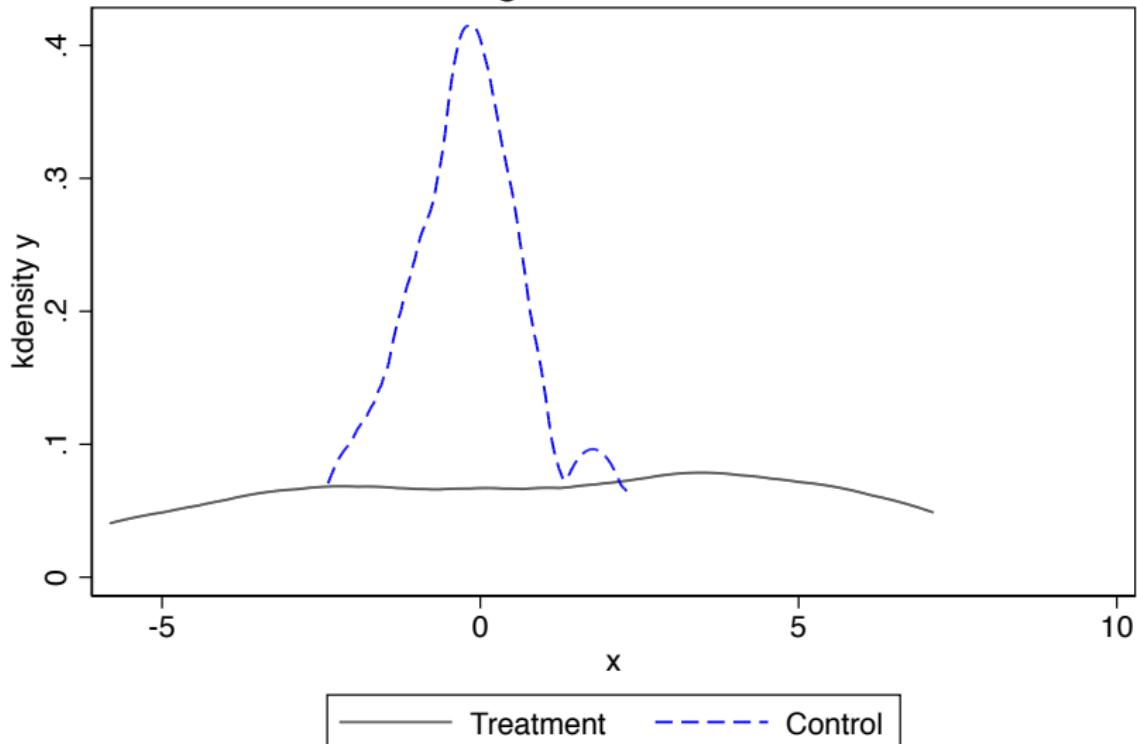
Kolmogorov-Smirnov statistic

- Test statistics are scalars not functions
- eCDFs are functions, not scalars
- Solution: use the maximum discrepancy between the two eCDFs:

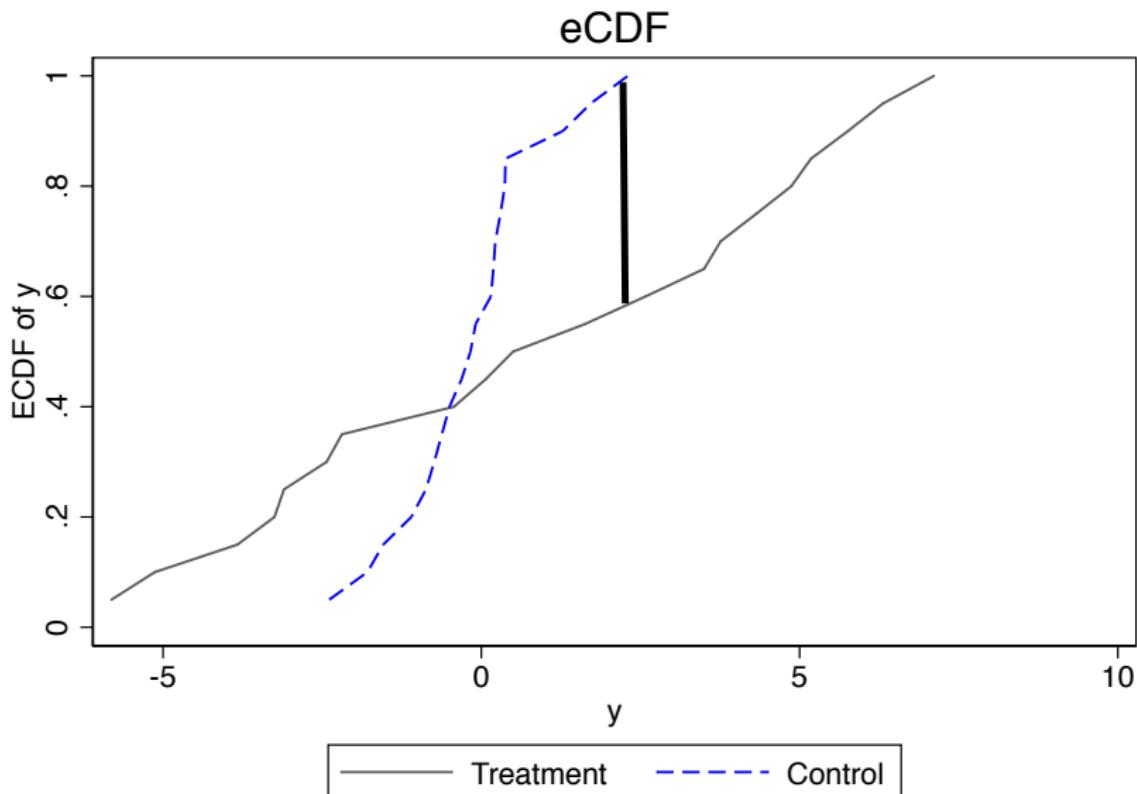
$$T_{KS} = \max | \hat{F}_T(Y_i) - \hat{F}_C(Y_i) |$$

Kernel density by group status

Kolmogorov-Smirnov test



eCDFs by treatment status and test statistic



KS Test Statistic

Treatment	D	Exact P-value
K-S	0.4500	0.034

Max distance is 0.45. Exact p is 0.034.

“Which bear is best?” – Jim Halpert

A good test statistic is the one that best fits your data. Some test statistics will have weird properties in the randomization as we'll see in synthetic control.

One-sided or two-sided?

- So far, we have defined all test statistics as absolute values
- We are testing against a two-sided alternative hypothesis

$$H_0 : \delta_i = 0 \quad \forall i$$

$$H_1 : \delta_i \neq 0 \text{ for some } i$$

- What about a one-sided alternative

$$H_0 : \delta_i = 0 \quad \forall i$$

$$H_1 : \delta_i > 0 \text{ for some } i$$

- For these, use a test statistic that is bigger under the alternative:

$$T_{diff*} = \bar{Y}_T - \bar{Y}_C$$

Small vs. Modest Sample Sizes are non-trivial

Computing the exact randomization distribution is not always feasible (Wolfram Alpha)

- $N = 6$ and $N_T = 3$ gives us 20 assignment vectors
- $N = 8$ and $N_T = 4$ gives us 70 assignment vectors
- $N = 10$ and $N_T = 5$ gives us 252 assignment vectors
- $N = 20$ and $N_T = 10$ gives us 184,756 assignment vectors
- $N = 50$ and $N_T = 25$ gives us 1.2641061×10^{14} assignment vectors

Exact p calculations are not realistic bc the number of assignments explodes at even modest size

Approximate p values

- Use simulation to get approximate p -values
 - Take K samples from the treatment assignment space
 - Calculate the randomization distribution in the K samples
 - Tests no longer exact, but bias is under your control (increase K)
- Imbens and Rubin show that p values converge to stable p values pretty quickly (in their example after 1000 replications)

Sample dataset

Let's do this now with Thornton's data. You can replicate that using thorton_ri.do or thornton_ri.R

Thornton's experiment

ATE	Iteration	Rank	<i>p</i>	no. trials
0.45	1	1	0.01	100
0.45	1	1	0.002	500
0.45	1	1	0.001	1000

Table: Estimated *p*-value using different number of trials.

Including covariate information

- Let X_i be a pretreatment measure of the outcome
- One way is to use this as a gain score: $Y^{d'} = Y_i^d - X_i$
- Causal effects are the same $Y^{1i} - Y^{0i} = Y_i^1 - Y_i^0$
- But the test statistic is different:

$$T_{gain} = \left| (\bar{Y}_T - \bar{Y}_C) - (\bar{X}_T - \bar{X}_C) \right|$$

- If X_i is strongly predictive of Y_i^0 , then this could have higher power
 - Y_{gain} will have lower variance under the null
 - This makes it easier to detect smaller effects

Regression in RI

- We can extend this to use covariates in more complicated ways
- For instance, we can use an OLS regression:

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon$$

- Then our test statistic could be $T_{OLS} = \widehat{\delta}$
- RI is justified even if the model is wrong
 - OLS is just another way to generate a test statistic
 - The more the model is “right” (read: predictive of Y_i^0), the higher the power T_{OLS} will have
- See if you can do this in Thornton’s dataset using the loops and saving the OLS coefficient (or just use `ritest`)

Hidden curriculum
Foundational causality stuff
Regression discontinuity designs
Instrumental variables
Twoway fixed effects estimator
Difference-in-differences
Comparative case studies
Matching and weighting
Concluding remarks

Regression review
Potential outcomes
Randomization and selection bias
Randomization inference
Causal models and Directed Acyclical Graphs

Judea Pearl and DAGs

- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology
- Their causality concepts are extremely clear, they provide a map to the estimation strategy, and maybe best of all, they communicate to others what must be true about the data generating process to recover the causal effect

Judea Pearl, 2011 Turing Award winner, drinking his first IPA



Further reading

- ① Pearl (2018) The Book of Why: The New Science of Cause and Effect, Basic Books (*popular*)
- ② Morgan and Winship (2014)
Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press, 2nd edition
(*excellent*)
- ③ Pearl, Glymour and Jewell (2016)
Causal Inference In Statistics: A Primer, Wiley Books
(*accessible*)
- ④ Pearl (2009) Causality: Models, Reasoning and Inference, Cambridge, 2nd edition (*difficult*)
- ⑤ Cunningham (2021) Causal Inference: The Mixtape, Yale, 1st edition (*best choice, no question*)

Causal model

- The causal model is sometimes called the structural model, but for us, I prefer the former as it's less alienating
- It's the system of equations describing the relevant aspects of the world
- It necessarily is filled with causal effects associated with some particular comparative statics
- To illustrate, I will assume a Beckerian human capital model

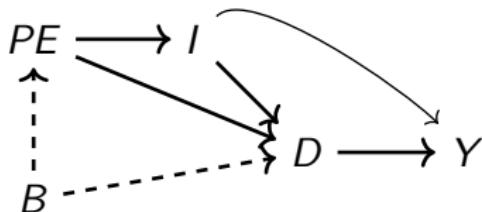
Human capital model: statements *and* graphs

Let's describe my simplified Beckerian human capital model.

- Individuals maximize utility by choosing consumption and schooling (D) subject to multi-period budget constraint
- Education has current costs but longterm returns
- But people choose different levels of schooling based on a number of things we will call “background” (B) which won’t be in the dataset (“unobserved”)
- And own-schooling will also be because of parental schooling (PE)
- Finally, wages (Y) are a function of parental schooling

Becker's human capital causal model

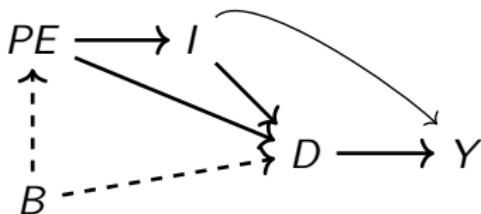
We can represent that causal model visually



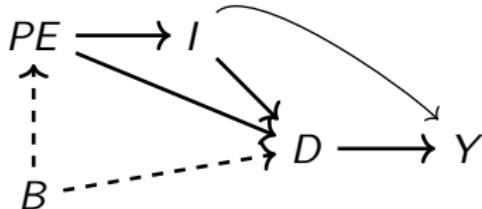
PE is parental education, B is “unobserved background factors (i.e., “ability”)\”, I is family income, D is college education and Y is log wages. The DAG is an approximation of Becker’s underlying (causal) human capital model.

Arrows, but also *missing arrows*

Before we dive into all this notation, couple of things

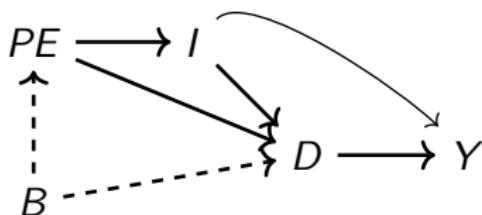


PE and D are caused by B . But why doesn't B cause Y ? Do you believe this? Why/why not? We can dispute this, but notice – we can see the assumption, which is transparent and communicates the author's beliefs, as well as the needed assumptions in their forthcoming *empirical* model. Every empirical strategy makes assumptions, but oftentimes they are not as transparent to us as this is.



- B is a **parent** of PE and D
- PE and D are **descendants** of B
- There is a **direct (causal) path** from D to Y
- There is a **mediated (causal) path** from B to Y through D
- There are four **paths** from PE to Y but none are direct, and one is unlike the others

Colliders



Notice anything different with this DAG? Look closely.

- D is a **collider** along the path $B \rightarrow D \leftarrow I$ (i.e., “colliding” at D)
- D is a **noncollider** along the path $B \rightarrow D \rightarrow Y$

Summarizing Value of DAGs imo

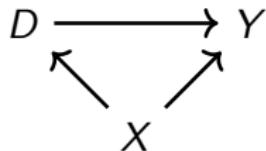
- ① Facilitates the task of designing identification strategy for estimating average causal effects
- ② Facilitates the task of testing compatibility of the model with your data
- ③ Visualizes the identifying assumptions which opens up the model to critical scrutiny

Creating DAGs

- The DAG is a *relevant* causal relationships describing the relationship between D and Y
- It will include:
 - All direct causal effects among the *relevant* variables in the graph
 - All common causes of any pair of *relevant* variables in the graph
- No need to model a dinosaur stepping on a bug causing in a million years some evolved created that impacted your decision to go to college
- We get ideas for DAGs from theory, models, observation, experience, prior studies, intuition
- Sometimes called the data generating process.

Confounding

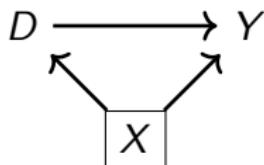
- Omitted variable bias has a name in DAGs: “confounding”
- Confounding occurs when the treatment and the outcomes have a common cause or parent which creates spurious correlation between D and Y



The *correlation* between D and Y no longer reflects the causal effect of D on Y

Backdoor Paths

- Confounding creates **backdoor paths** between treatment and outcome ($D \leftarrow X \rightarrow Y$) – i.e., spurious correlations
- Not the same as mediation ($D \rightarrow X \rightarrow Y$)
- We can “block” backdoor paths by conditioning on the common cause X
- Once we condition on X , the correlation between D and Y estimates the causal effect of D on Y
- Conditioning means calculating $E[Y|D = 1, X] - E[Y|D = 0, X]$ for each value of X then combining (e.g., integrating)



Blocked backdoor paths

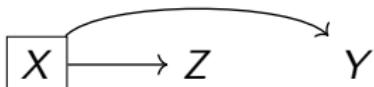
A backdoor path is blocked if and only if:

- It contains a noncollider that has been conditioned on
- Or it contains a collider that has not been conditioned on

Examples of blocked paths

Examples:

- ① Conditioning on a noncollider blocks a path:



- ② Conditioning on a collider opens a path (i.e., creates spurious correlations):



- ③ Not conditioning on a collider blocks a path:



Backdoor criterion

Backdoor criterion

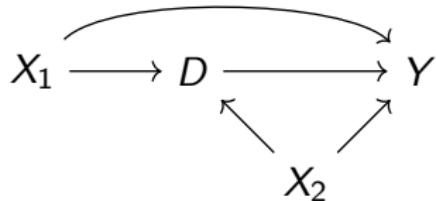
Conditioning on X satisfies the backdoor criterion with respect to (D, Y) directed path if:

- ① All backdoor paths are blocked by X
- ② No element of X is a collider

In words: If X satisfies the backdoor criterion with respect to (D, Y) , then controlling for or matching on X identifies the causal effect of D on Y

What control strategy meets the backdoor criterion?

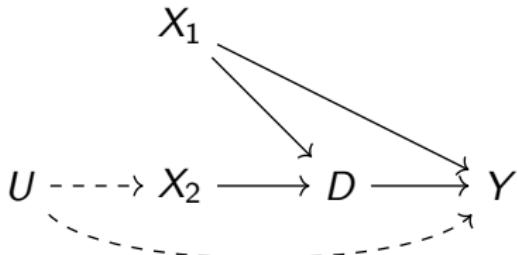
- List all backdoor paths from D to Y . I'll wait.



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?

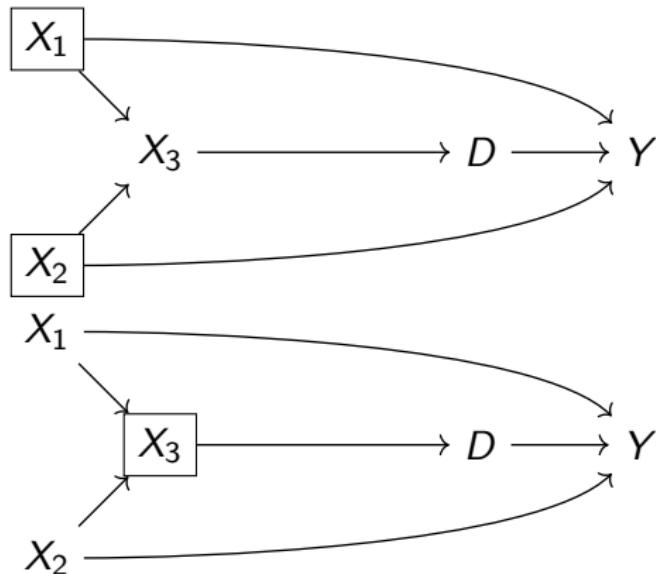
What if you have an unobservable?

- List all the backdoor paths from D to Y .



- What are the necessary and sufficient set of controls which will satisfy the backdoor criterion?
- What about the unobserved variable, U ?

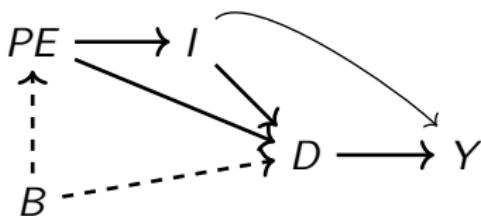
Multiple strategies

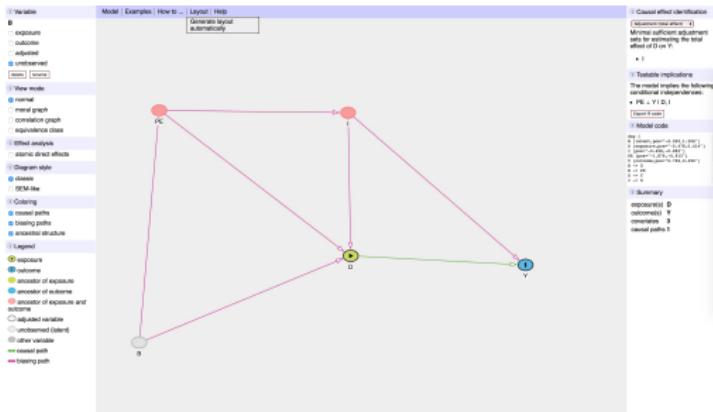


- Conditioning on the common causes, X_1 and X_2 , is sufficient
- ... but so is conditioning on X_3

Testing the Validity of the DAG

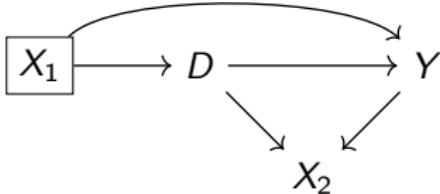
- The DAG makes testable predictions
- Conditional on D and I , parental education (PE) should no longer be correlated with Y
- Can be hard to figure this out by hand, but software can help (e.g., Daggity.net is browser based)



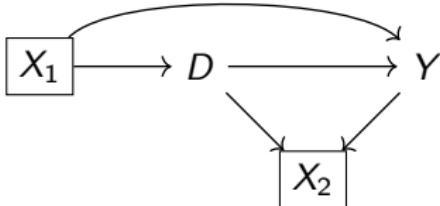


Collider bias

- Conditioning on a collider introduces spurious correlations; can even mask causal directions
 - There is only one backdoor path from D to Y

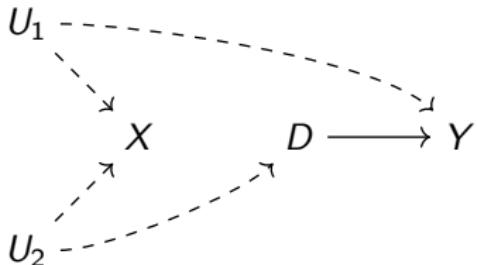


- Conditioning on X_1 blocks the backdoor path
- But what if we also condition on X_2 ?

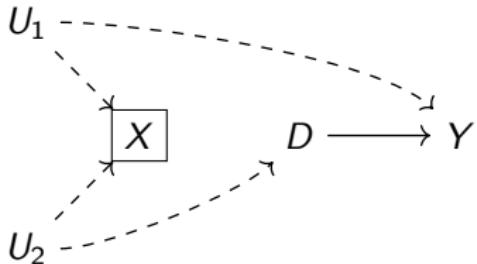


- Conditioning on X_2 opens up a new path, creating new spurious correlations between D and Y

- Even controlling for pretreatment covariates can create bias
 - Name the backdoor paths. Is it open or closed?



- But what if we condition on X ?



Living in reality - he doesn't love you

- **Fact #1:** We can't know if we have a collider bias (confounder) problem without making assumptions about the causal model (i.e. not in the codebook)
- **Fact # 2:** You can't just haphazardly throw in a bunch of controls on the RHS (i.e., "the kitchen sink") bc you may inadvertently be conditioning on a collider which can lead to massive biases
- **Fact # 3:** You have no choice but to leverage economic theory, intuition, intimate familiarity with institutional details and background knowledge for research designs.
- **Fact #4:** You can only estimate causal effects with **data and assumptions**.

Examples of collider bias

Bad controls

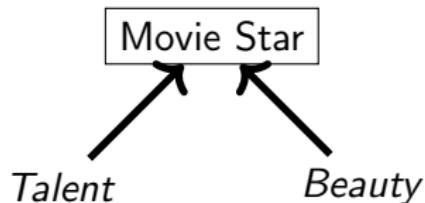
- Angrist and Pischke in MHE talk about a specific type of danger associated with controlling for an outcome – “bad controls”
- The problem is not controlling for an outcome;
- The problem is controlling for a collider and don’t correct for *that*
- This has implications for when you work with non-random administrative data, too

Sample selection example of collider bias

Important: Since unconditioned colliders block back-door paths, what exactly does conditioning on a collider do? Let's illustrate with a fun example and some made-up data

- CNN.com headline: Megan Fox voted worst – but sexiest – actress of 2009 ([link](#))
- Are these two things actually negatively correlated in the world?
- Assume talent and beauty are independent, but each causes someone to become a movie star. What's the correlation between talent and beauty for a sample of movie stars compared to the population as a whole (stars and non-stars)?

- What if the sample consists *only* of movie stars?



Stata code

```
clear all
set seed 3444

* 2500 independent draws from standard normal distribution
set obs 2500
generate beauty=rnormal()
generate talent=rnormal()

* Creating the collider variable (star)
gen score=(beauty+talent)
egen c85=pctile(score), p(85)
gen star=(score>=c85)
label variable star "Movie star"

* Conditioning on the top 15%
twoway (scatter beauty talent, mcolor(black) msymbol(smx)),
ytitle(Beauty) xtitle(Talent) subtitle(Aspiring actors and actresses)
by(star, total)
```

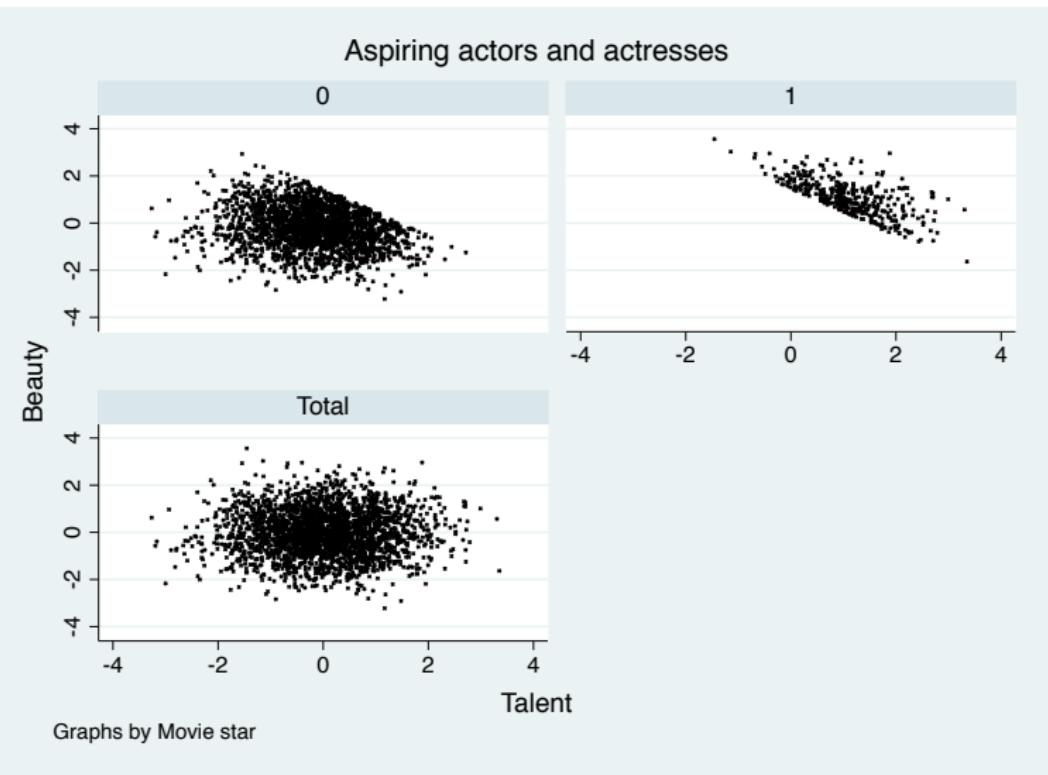


Figure: Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent.

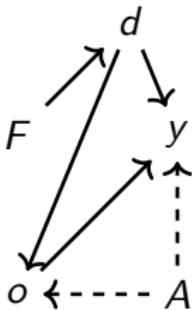
Stata

- Run Stata file star.do

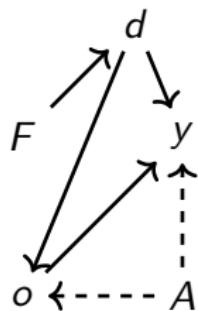
Occupational sorting and discrimination example of collider bias

- Let's look at another example: very common for think tanks and journalists to say that the gender gap in earnings disappears once you control for occupation.
- But what if occupation is a collider, which it could be in a model with occupational sorting
- Then controlling for occupation in a wage regression searching for discrimination can lead to all kinds of crazy results even *in a simulation where we explicitly design there to be discrimination*

DAG



F is female, d is discrimination, o is occupation, y is earnings and A is ability. Dashed lines mean the variable cannot be observed. Note, by design, being a female has no effect on earnings or occupation, and has no relationship with ability. So earnings is coming through discrimination, occupation, and ability.



Mediation and Backdoor paths

- ① $d \rightarrow o \rightarrow y$
- ② $d \rightarrow o \leftarrow A \rightarrow y$

Stata model (Erin Hengel)

- Erin Hengel (www.erinhengel.com) and I worked out this code and she gave me permission to put in my Mixtape
- Let's look at `collider_discrimination.do` or `collider_discrimination.R` together

Table: Regressions illustrating collider bias with simulated gender disparity

Covariates:	Unbiased combined effect	Biased	Unbiased wage effect only
Female	-3.074*** (0.000)	0.601*** (0.000)	-0.994*** (0.000)
Occupation		1.793*** (0.000)	0.991*** (0.000)
Ability			2.017*** (0.000)
N	10,000	10,000	10,000
Mean of dependent variable	0.45	0.45	0.45

- Recall we designed there to be a discrimination coefficient of -1
- If we do not control for occupation, then we get the combined effect of $d \rightarrow o \rightarrow y$ and $d \rightarrow y$
- Because it seems intuitive to control for occupation, notice column 2 - the sign flips!
- We are only able to isolate the direct causal effect by conditioning on ability and occupation, but ability is unobserved

Administrative data

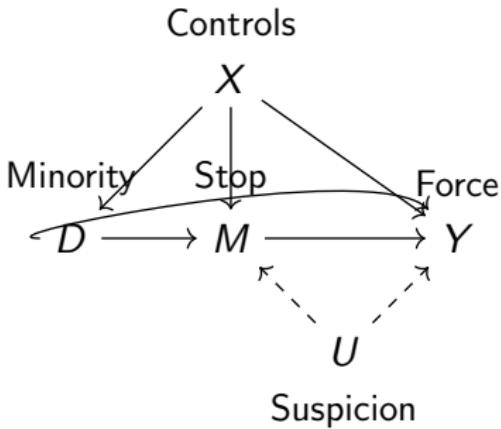
- Admin data has become extremely common, if not absolutely necessary
- But naive use of admin data can be dangerous if the drawing of the sample is itself a collider problem (Heckman 1979; Elwert and Winship 2014)
- Let's look at a new paper by Fryer (2019) and a critique by Knox, et al. (2019)

Collider bias and police use of force

- Claims of excessive and discriminatory use of police force against minorities (e.g., Black Lives Matter, Trayvon Martin, Michael Brown, Eric Garner)
- Challenging to identify
 - Police-citizen interactions are conditional on interactions having already been triggered
 - That initial interaction is unobserved
- Fryer (2019) is a monumental study for its data collection and analysis: Stop and Frisk, Police-Public Contact Survey, and admin data from two jurisdictions
- Codes up almost 300 variables from arrest narratives which range from 2-100 pages in length – shoe leather!

Initial interaction

- Fryer finds that blacks and Hispanics were more than 50% more likely to have an interaction with the policy in NYC Stop and Frisk as well as Police-Public Contact survey
- It survives extensive controls – magnitudes fall, but still very large (21%)
- Moves to admin data
- Conditional on police interaction, *no* racial differences in officer-related shootings
- Fryer calls it one of the most surprising findings in his career
- Lots of eyes on this study as a result of the counter intuitive results; published in JPE
- Knox, et al (202) claim his data is itself a collider. What?



Fryer told us $D \rightarrow M$ exists from both Stop and Frisk and Police-Public. But note: admin data is instances of M stops, which is itself a collider. If this DAG is true, then spurious correlations enter between M and Y which may dilute our ability to estimate causal effects.

Knox, et al (2020)

- Move from DAG to more contemporary potential outcomes notation to design relevant parameters
- Use potential outcomes and bounds
- Even with lower bound estimates of the incidence of police violence against civilians is more than 5x higher than what Fryer (2019) finds
- Heckman (1979) – we *cannot* afford to ignore sample selection

Summarizing all of this

- Your dataset will not come with a codebook flagging some variables as “confounders” and other variables as “colliders” because those terms are always context specific
- Except for some unique situations that aren’t generally applicable, you also don’t always know statistically you have an omitted variable bias problem; but both of these are fatal for any application
- You only know to do what you’re doing based on *knowledge about data generating process.*
- All identification must be guided by theory, experience, observation, common sense and knowledge of institutions
- DAGs absorb that information and can be then used to write out the explicit identifying model

DAGs are not panacea

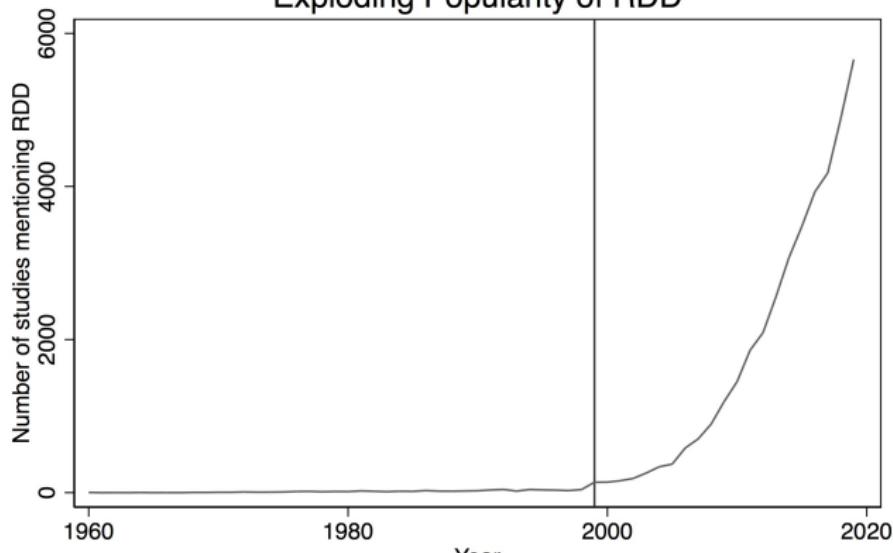
- DAGs cannot handle, though, reverse causality or simultaneity
- So there are limitations. “All models are wrong but some are useful”
- They are also not popular (see Twitter ongoing debates which have descended into light hearted jokes as well as aggressive debates)
- But I think they are helpful and while not *necessary*, showcase what is necessary – assumptions
- Heckman (1979) can maybe provide some justification at times

What is regression discontinuity design?

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

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

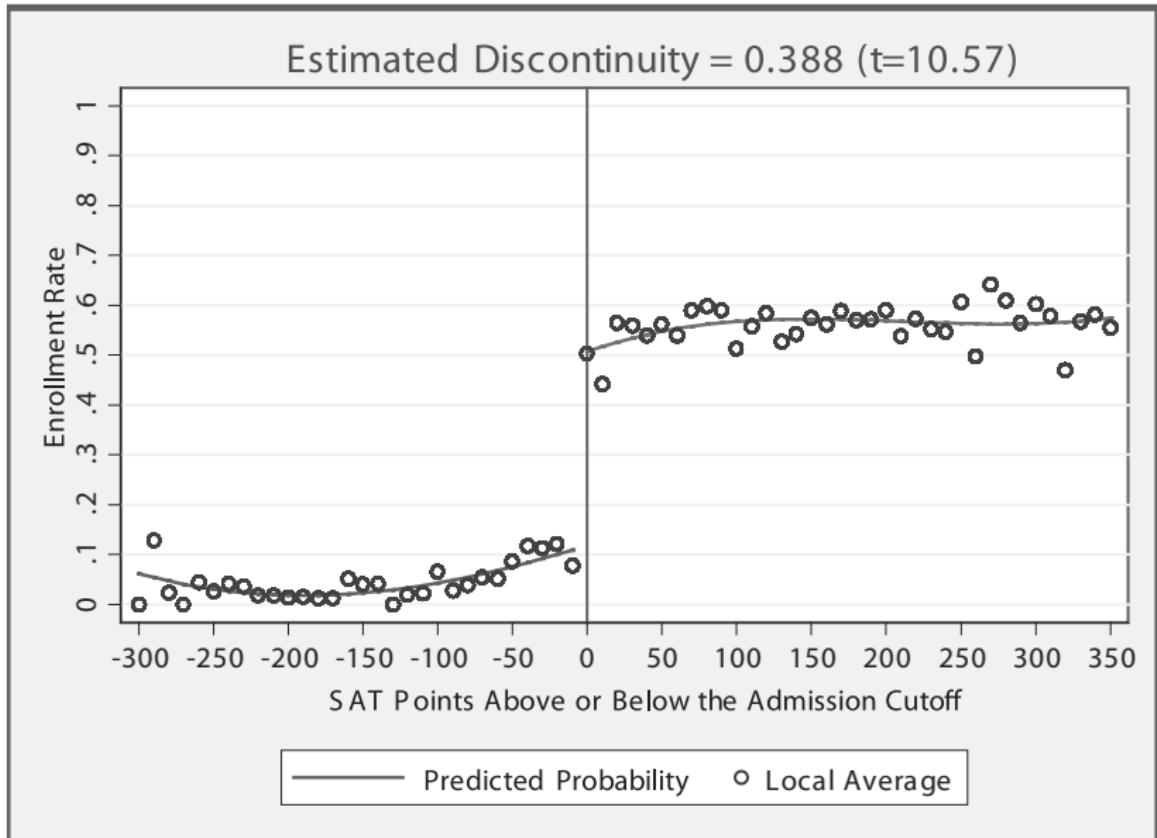
Exploding Popularity of RDD



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

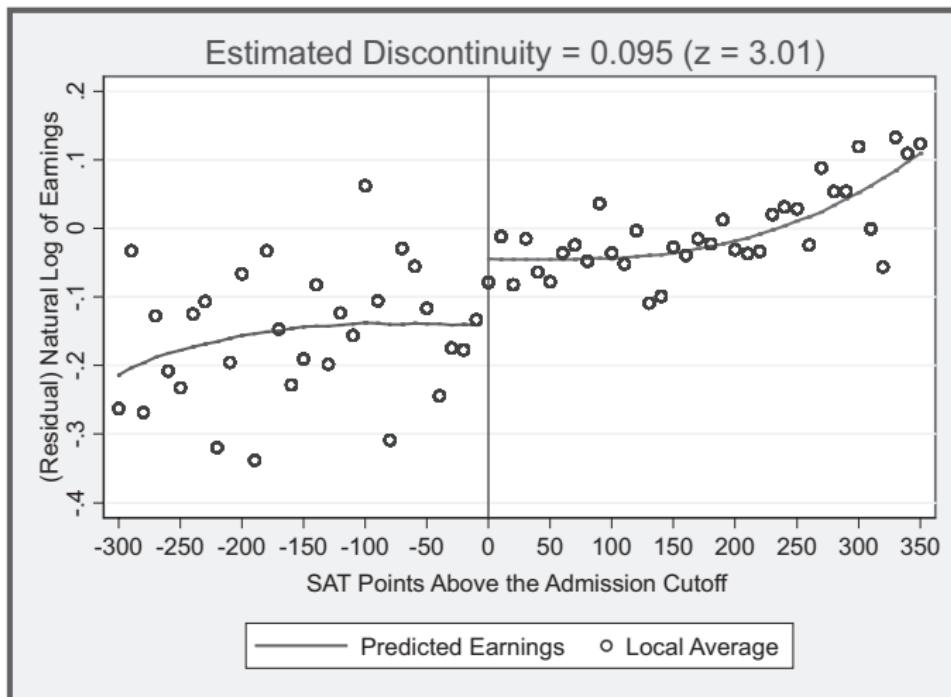
Tell me what you think is happening

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Tell me what you think is happening

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



What is a regression discontinuity design?

- We want to estimate some causal effect of a treatment on some outcome, but we're worried about selection bias due to self-selection into treatment
- But what if treatment assignment occurs abruptly when some underlying variable X called the “running variable” passes a cutoff c_0 ?
- RDD formalizes the effort to estimate causal effects using just such an event.

Running and jumping

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

Examples from the literature

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

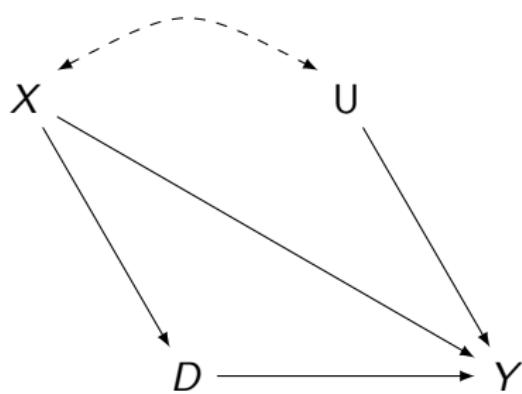
Data requirements

Large sample sizes are characteristic features of the RDD

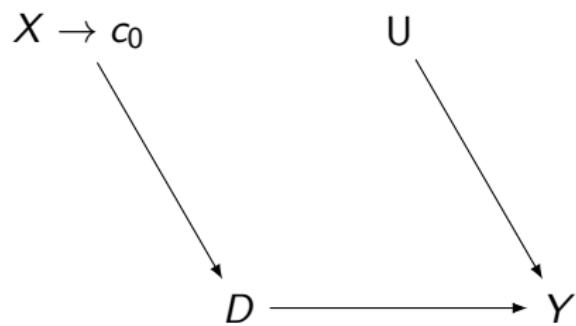
- Usually we think of “trends” as time trends, but in an RDD, trends refer to that “running variable” – but meaning is the same
- If there are strong trends in the running variable, one typically needs a lot more data than if there weren’t
- We need a lot of data bc we need to fill out the running variable so there is large mass at the cutoff
- Researchers are typically using administrative data or settings such as birth records where there are **many** observations

Might explain why the method never caught on until the 00's

(A) Data generating graph



(B) Limiting graph



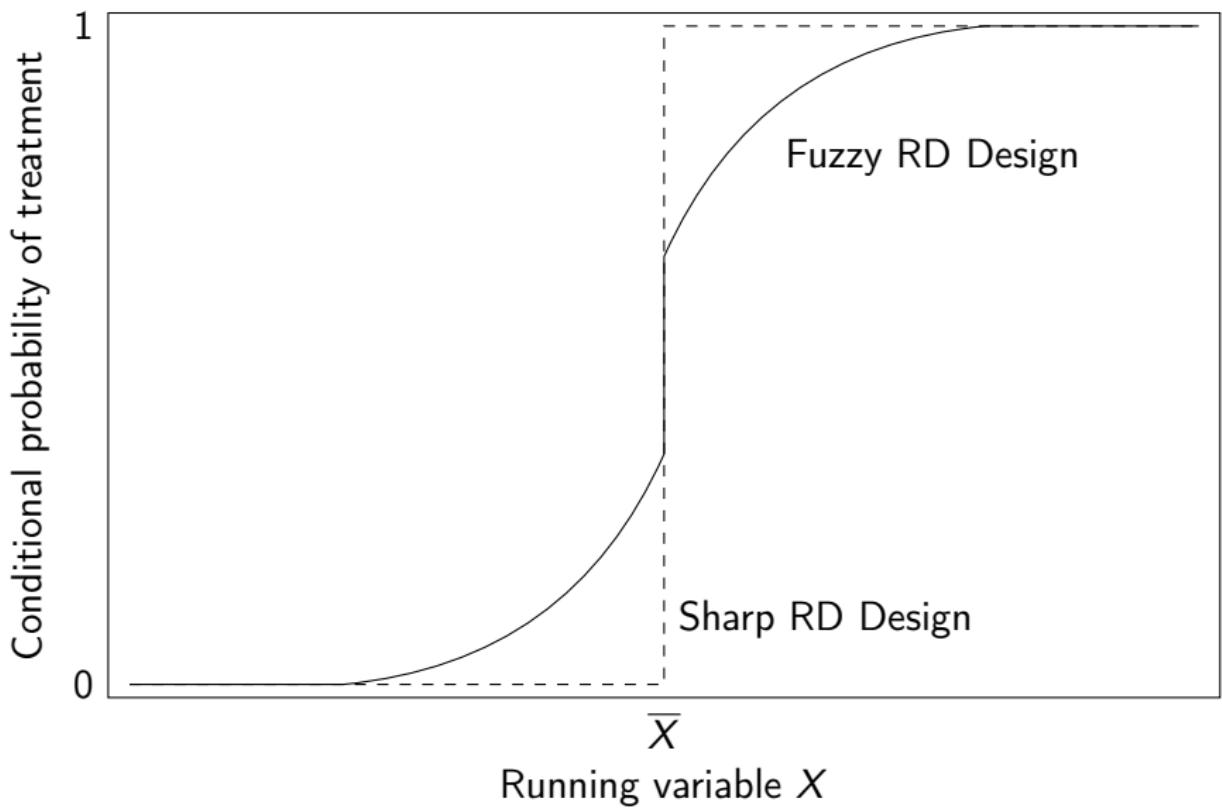


Figure: Sharp vs. Fuzzy RDD

Sharp vs. Fuzzy RDD

- There's traditionally thought to be two kinds of RD designs:
 - ① Sharp RDD: Treatment is a deterministic function of running variable, X . Example: Medicare benefits.
 - ② Fuzzy RDD: Discontinuous “jump” in the *probability* of treatment when $X > c_0$. Cutoff is used as an instrumental variable for treatment. Example: attending state flagship
- Fuzzy is a type of IV strategy and requires explicit IV estimators like 2SLS; sharp is reduced form IV and doesn't require IV-like estimators

Overlap

- Notice that in the sharp design, we have people in the treated or untreated along the running variable but not both – no “overlap”
- But independence will give us an *equal* likelihood of being in either group
- This implies that we have units in treatment and control at every point along the running variable in other words

Overlap (cont.)

- This doesn't happen with RDD sharp designs because the cutoff randomizes individuals to left and right of cutoff not up and down as independence ordinarily creates
- *This is why we use extrapolation beyond the support of the data* (left and right of the cutoff) – because we can't compare units with the same running variable score
- So which estimator we use is necessarily important because not all estimators yield the same extrapolated points of interest

Treatment assignment in the sharp RDD

Deterministic treatment assignment (“sharp RDD”)

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

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

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

Example: Medicare: Americans aged 64 are *not* eligible for Medicare, but Americans aged 65 are eligible for Medicare (ignoring disability exemptions). Notice no 64 year olds are in Medicare, and no 65 year olds are in the control group (no overlap)

Treatment effect definition and estimation

Definition of treatment effect

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

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

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

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

Extrapolation

- In RDD, the counterfactuals are conditional on X .
- We use *extrapolation* in estimating treatment effects with the sharp RDD bc we do not have overlap
 - Left of cutoff, only non-treated observations, $D_i = 0$ for $X < c_0$
 - Right of cutoff, only treated observations, $D_i = 1$ for $X \geq c_0$
- The extrapolation is to a counterfactual

Extrapolation

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

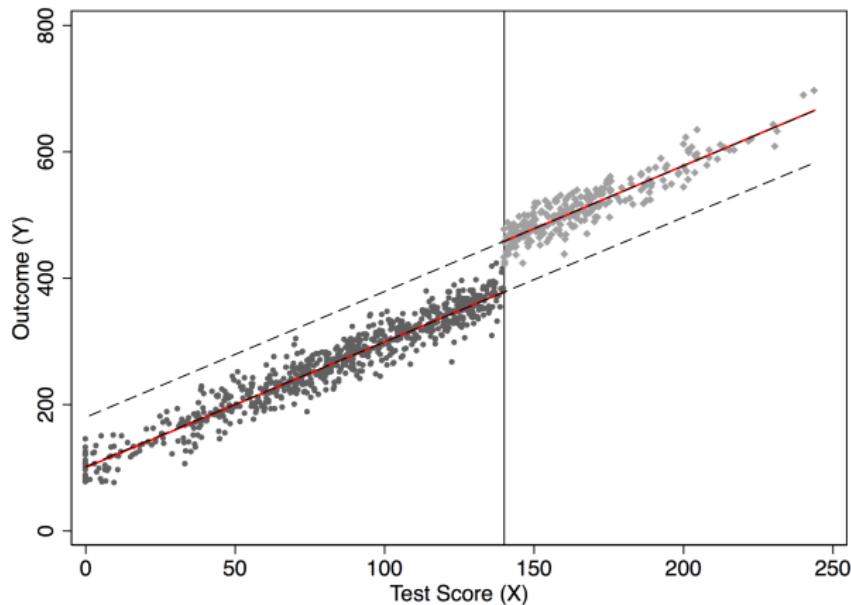


Figure: Dashed lines are extrapolations (Marcelo Perraillon simulated random variables)

Key identifying assumption

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

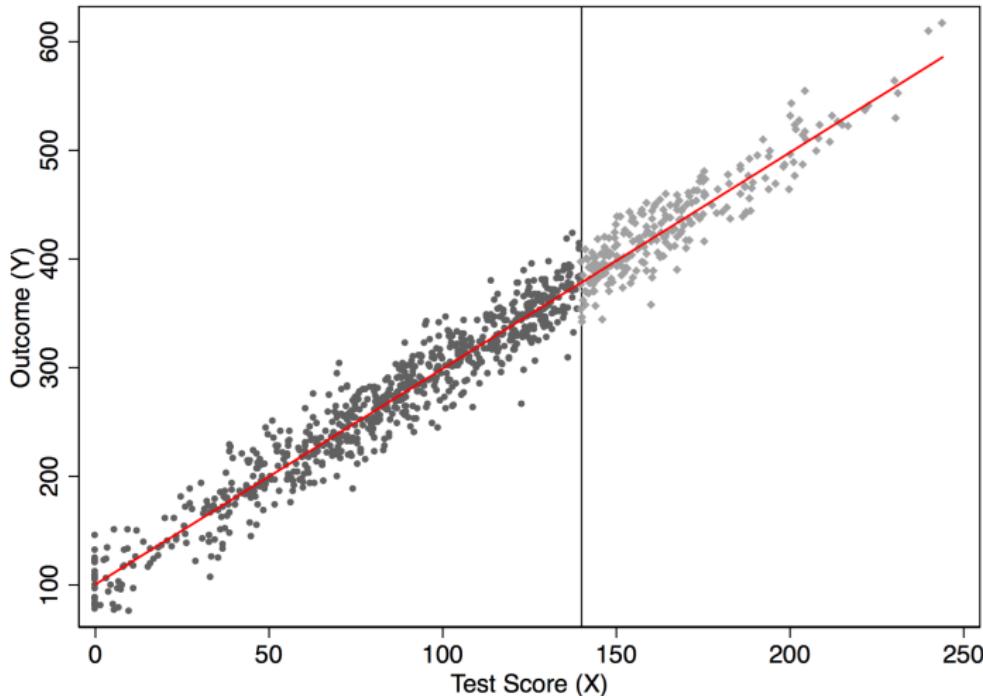
$E[Y_i^0|X = c_0]$ and $E[Y_i^1|X = c_0]$ are continuous (smooth) in X at c_0 .

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

Smoothness is the identifying assumption and untestable

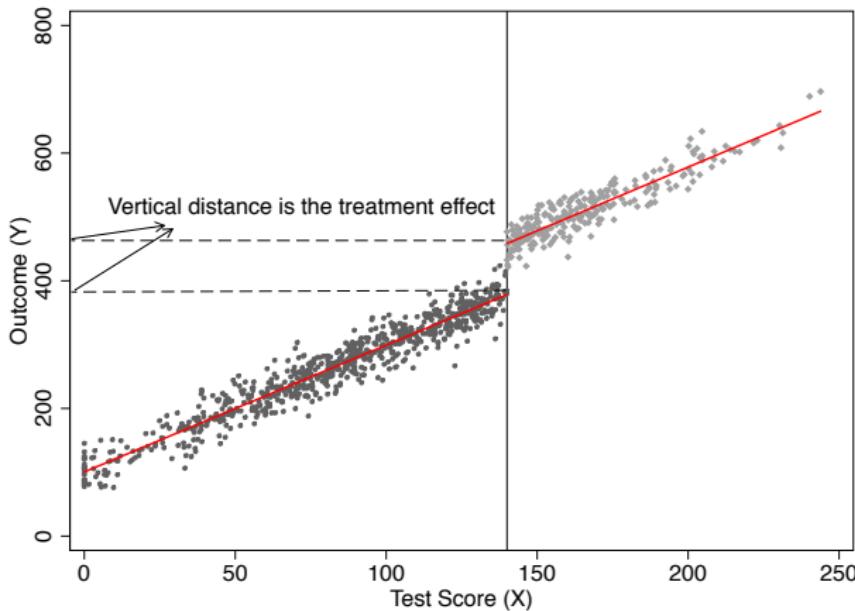
- The smoothness assumption allows us to use average outcome of units right below the cutoff as a valid counterfactual for units right above the cutoff.
- In other words, extrapolation is allowed if smoothness is credible, and extrapolation is nonsensical if smoothing isn't credible
- The causal effect of the treatment will be based on **extrapolation** from the trend, $E[Y_i^0|X < c_0]$, to those values of $X > c_0$ for the $E[Y_i^0|X > c_0]$.
- Means you have to think long and hard about smoothness and what violations mean in your context
- Why then is it not directly testable? Because potential outcomes are counterfactual

Graphical example of the smoothness assumption



Note these are *potential* not *actual* outcomes (Marcelo Perraillon simulated random variables)

Graphical example of the treatment effect, not the smoothness assumption



Note that these are *actual*, not *potential* outcomes. (Marcelo Perraillon simulated random variables)

Re-centering the data

- It is common for authors to transform X by “centering” at c_0 :

$$Y_i = \alpha + \beta(X_i - c_0) + \delta D_i + \varepsilon_i$$

- This doesn't change the interpretation of the treatment effect
– only the interpretation of the intercept.

Re-centering the data

- Example: Medicare and age 65. Center the running variable (age) by subtracting 65:

$$\begin{aligned}Y &= \beta_0 + \beta_1(Age - 65) + \beta_2Edu \\&= \beta_0 + \beta_1Age - \beta_165 + \beta_2Edu \\&= \alpha + \beta_1Age + \beta_2Edu\end{aligned}$$

where $\alpha = \beta_0 - \beta_165$.

- All other coefficients, notice, have the same interpretation, except for the intercept.

Regression without re-centering

```
reg y D x
```

Source	SS	df	MS	Number of obs	=	999
Model	15842893.9	2	7921446.97	F(2, 996)	=	19988.47
Residual	394715.557	996	396.30076	Prob > F	=	0.0000
Total	16237609.5	998	16270.1498	R-squared	=	0.9757

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	80.01418	2.144779	37.31	0.000	75.80537 84.22298
x	1.986975	0.186779	106.38	0.000	1.950322 2.023627
_cons	100.3885	1.70944	58.73	0.000	97.03397 103.743

Regression with centering

```
gen x_c = x - 140
```

```
reg y D x_c
```

Source	SS	df	MS	Number of obs =	999
Model	15842893.9	2	7921446.97	F(2, 996) =	19988.47
Residual	394715.554	996	396.300757	Prob > F =	0.0000
Total	16237609.5	998	16270.1498	R-squared =	0.9757
				Adj R-squared =	0.9756
				Root MSE =	19.907

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	80.01418	2.144779	37.31	0.000	75.80537 84.22298
x_c	1.986975	.0186779	106.38	0.000	1.950322 2.023627
cons	378.565	1.290755	293.29	0.000	376.032 381.0979

Nonlinearity bias

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

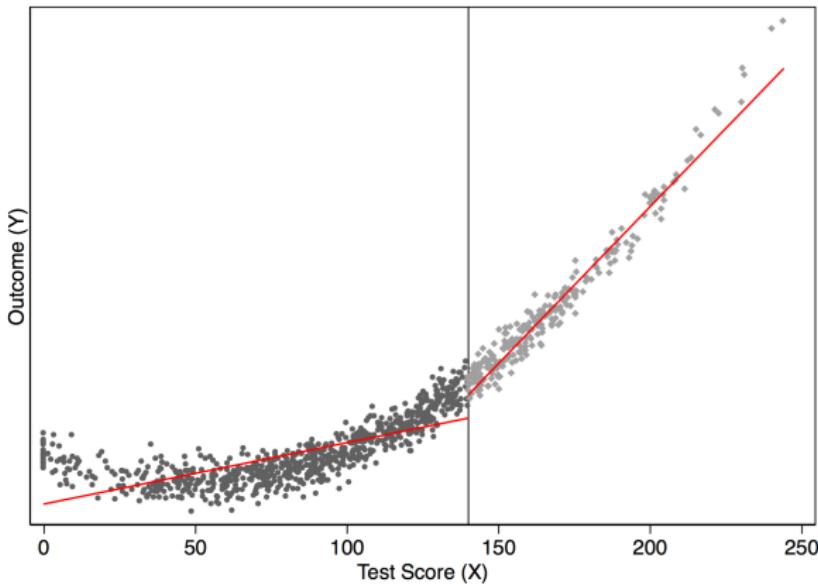
Marcelo Perraillon simulated random variables

```
gen x2 = x*x
```

```
gen x3 = x*x*x
```

```
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)
```

```
scatter y x if D==0, msize(vsmall) || scatter y x ///
if D==1, msize(vsmall) legend(off) xline(140, ///
lstyle(foreground)) ylabel(none) || lfit y x ///
if D ==0, color(red) || lfit y x if D ==1, ///
color(red) xtitle("Test Score (X)") ///
ytitle("Outcome (Y)")
```



See how the two lines don't touch at c_0 but empirically should?
That's bc the linear fit is the wrong functional form – we know this
from the simulation that it's the wrong functional form. (Marcelo
Perraillon simulated random variables)

Sharp RDD: Nonlinear Case

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

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

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

Nonlinearities

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

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

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

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

Different polynomials on the 2 sides of the discontinuity

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

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

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

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

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

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

Different polynomials on the 2 sides of the discontinuity

- To derive a regression model, first note that the observed values must be used in place of the potential outcomes:

$$E[Y|X] = E[Y^0|X] + (E[Y^1|X] - E[Y^0|X]) D$$

which is the switching equation from earlier expressed in terms of conditional expectation functions

- Regression model you estimate is:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \delta D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \varepsilon_i \end{aligned}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{21} - \beta_{11}$ and $\beta_p^* = \beta_{1p} - \beta_{0p}$

- The treatment effect at c_0 is δ

Polynomial simulation example (Marcelo Perraillon simulated random variables)

```
capture drop y x2 x3

gen x2 = x*x
gen x3 = x*x*x
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)

reg y D x x2 x3
predict yhat

scatter y x if D==0, msize(vsmall) || scatter y x
if D==1, msize(vsmall) legend(off) xline(140,
lstyle(foreground)) ylabel(none) || line yhat x
if D ==0, color(red) sort || line yhat x if D==1,
sort color(red) xtitle("Test Score (X)")
ytitle("Outcome (Y)")
```

Polynomial simulation example

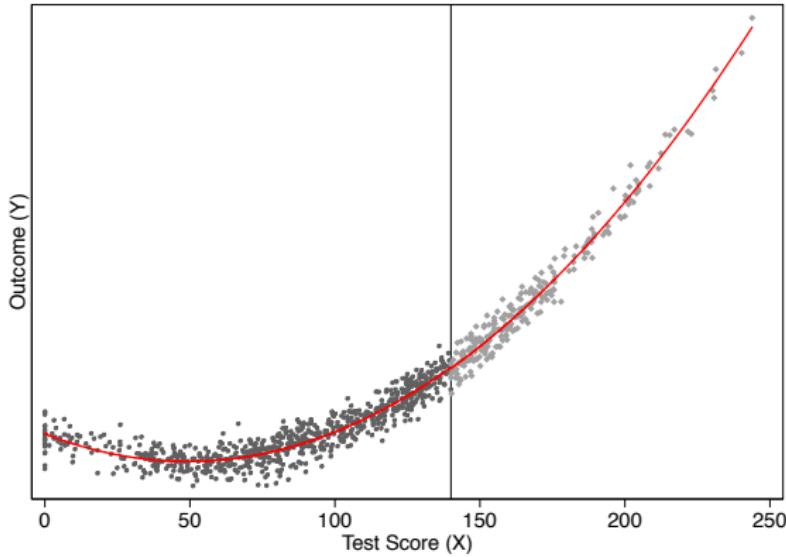


Figure: Third degree polynomial. Actual model second degree polynomial. (Marcelo Perraillon simulated random variables).

Notice: no more gap at c_0 once we model the function $f(x)$

Stata simulation

```
gen x2_c = x2 - 140
```

```
gen x3_c = x3 - 140
```

```
reg y D x x2
```

```
reg y D x_c x2_c
```

Polynomial simulation example

Source	SS	df	MS	Number of obs = 999
Model	3.7863e+10	3	1.2621e+10	F(3, 995) = 13115.22
Residual	957507024	995	962318.617	Prob > F = 0.0000
Total	3.8821e+10	998	38898361.8	R-squared = 0.9753 Adj R-squared = 0.9753 Root MSE = 980.98

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	-115.5381	127.4967	-0.91	0.365	-365.7314 134.6552
x	-98.57582	2.285769	-43.13	0.000	-103.0613 -94.09034
x2	1.000001	.0122767	81.45	0.000	.9759098 1.024092
_cons	9864.218	111.1206	88.77	0.000	9646.16 10082.28

Notice: no more gap at c_0 once we model the function $f(x)$ (e.g., D is insignificant once we include polynomials)

Polynomial simulation example

Source	SS	df	MS	Number of obs	=	999
Model	3.7863e+10	3	1.2621e+10	F(3, 995)	=	13115.22
Residual	957507020	995	962318.613	Prob > F	=	0.0000
Total	3.8821e+10	998	38898361.8	R-squared	=	0.9753

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
D	-115.5381	127.4967	-0.91	0.365	-365.7315	134.6552
x_c	-98.57582	2.285769	-43.13	0.000	-103.0613	-94.09034
x2_c	1.000001	.0122767	81.45	0.000	.9759098	1.024092
_cons	-3796.397	227.7894	-16.67	0.000	-4243.4	-3349.394

And centering did nothing to the interpretation of the main results (D), only to the intercept.

Hidden curriculum	
Foundational causality stuff	
Regression discontinuity designs	
Instrumental variables	
Twoway fixed effects estimator	
Difference-in-differences	
Comparative case studies	
Matching and weighting	
Concluding remarks	
	Introduction
	Sharp Design
	Smoothness, Extrapolation and Estimators
	Testing for violations
	Visualization
	Inference, kernels, bandwidths
	Sub-RDD: Close election designs

Can we evaluate credibility of smoothness assumption?

- Your main results are only causal insofar as smoothness is a credible belief, and since smoothness isn't guaranteed by "the science" like an RCT, you have to build your case
- You must now scrutinize alternative hypotheses that are consistent with your main results through sensitivity checks, placebos and alternative approaches

Main Challenges

Classify your concern regarding smoothness violations into two categories:

- Manipulation on the running variable
- Endogeneity of the cutoff

Most robustness is aimed at building credibility around these,

Manipulation of your running variable score

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

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

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

Test 1: Manipulation of the running variable

Manipulation of the running variable

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

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

A badly designed RCT

- Imagine a treatment (statin) that people widely believe will save lives
- Now imagine a team randomly assigns patients to treatment or control to study the effect of the statin on heart attacks within 10 years
- 200 patients are placed in two different waiting rooms – 100 in *A* and 100 in *B*
- If people know which room has the statin, people want to move between rooms, and people have time to move into the rooms, how many people will be in *A*? *B*?

Choosing your running variable value

- 200 people in A ; 0 people in B
- While we cannot test for smoothness because it involves potential outcomes, we can test for a break in the number of units around the cutoff
- This can be tested using a density test (McCrary 2008)

McCrary Density Test

- Assume a null where the *density* is continuous at the cutoff point (why not? Jumps demand explanations if continuity is typical)
- Under the alternative hypothesis, the density increases at the cutoff as people sort onto the desirable side of the cutoff
- This is oftentimes visualized with confidence intervals illustrating the effect of the discontinuity on density - you need no jump to pass this test

Steps for a density test in RDD

- ① Count observations for a chosen bin (but again – which bin is the right bin??)
- ② Estimate your nonlinear OLS model with quadratics in the running variable on the *counts*
- ③ Check whether you can reject the null at the cutoff?

McCrary Density Ttest

- Density tests are **mandatory** for every analysis using RDD
- How do you know if you can do a density test? If you were able to calculate conditional expectations in the first place
- You can download the (no longer supported) Stata ado package, DCdensity, to implement McCrary's density test (<http://eml.berkeley.edu/~jmccrary/DCdensity/>) or use rddensity (R and Stata)

Caveats

- A discontinuity in the density is “suspicious” because it *suggests* manipulation.
- Smoothness *might* not hold because if the units sorting have different potential outcomes such that smoothness no longer holds, then smoothness is violated.
- But technically one doesn't need a smooth density for smoothness in potential outcomes to be true
- Density tests are demanding on data as we will see – you can fail to reject with nonrandom heaping for instance
- Density tests are not helpful for RD in time (Hausman and Rapson 2018), which is an entirely different subject

Simulations of density tests

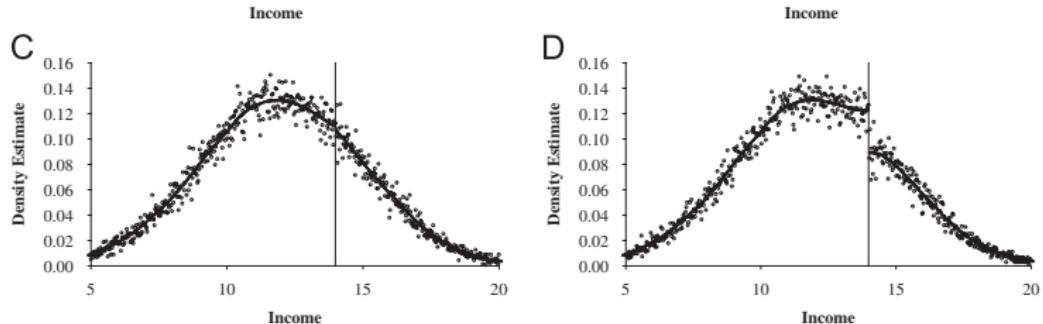


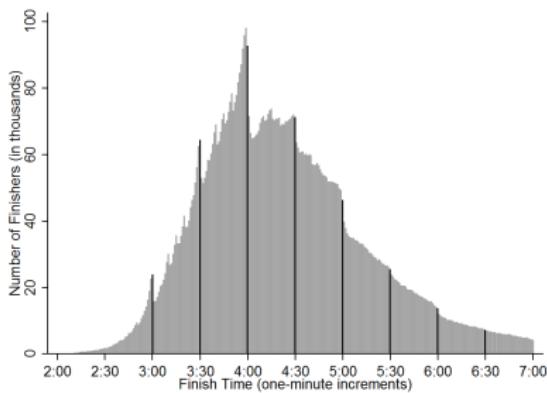
Figure: From McCrary (2008). Left shows failing to reject. Right shows rejection of the null.

I have consistently found manipulation on the running variable when evaluating SNAP (food stamps) using income cutoffs fwiw

Density tests in marathons

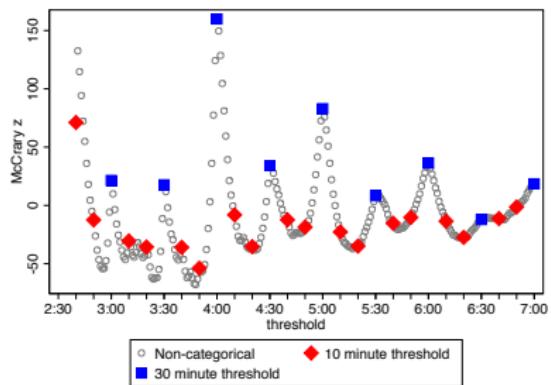
Do people maximize speed in a marathon? Or do they have “reference dependent” times (e.g., making Boston)? Let’s look at raw data by running variable and tests for rejection across the running variable

Figure 2: Distribution of marathon finishing times ($n = 9,378,546$)



NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

Figure 3: Running McCrary z -statistic



NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

Figure: From Allen, Dechow, Pope and Wu (2013) “Reference-Dependent Preferences: Evidence from Marathon Runners”

Nonrandom heaping

- Premature babies both receive expensive medical spending *and* are more likely to perish.
- What is the causal effect of medical spending on mortality?
- Severe selection bias – only babies with high risk of mortality ex ante are more likely to be in the intensive care units for children
- But it's an important question
- A group uses RDD to answer it, but how? We need a cutoff and a running variable. Ask yourself – what might that be?

Almond et al. (201) RDD strategy

- Almond, et al. (2010) attempted to estimate the causal effect of medical expenditures on health outcomes using RDD
- In the US, newborns whose birthweight falls below 1500 grams are placed in intensive care bc 1500 is the “very low birth weight” range
- Compare those just above with those below 1500 using a variety of estimators and visualizations
- They used hospital administrative records and found 1-year infant mortality decreased by 1pp just below 1500 grams compared to just above
- Concluded these medical expenditures are cost-effective (compounded value of life over typical lifespan)
- But now let's look at density tests

Heaping problem

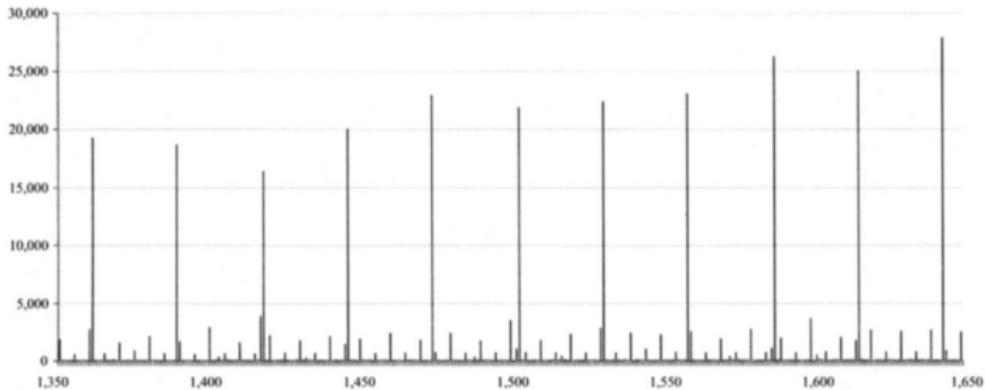


FIGURE I
Frequency of Births by Gram: Population of U.S. Births
between 1,350 and 1,650 g

NCHS birth cohort linked birth/infant death files, 1983–1991 and 1995–2003,
as described in the text.

Figure: Distribution of births by gram from Almond, et al. 2010

Heaping along the Running variable

- This picture shows “heaping” which is excess mass at certain points along the running variable
- Unlikely births actually heap at certain intervals; more likely someone is rounding – but who?
- Some scales may be less sophisticated, some practices may be more common in some types of hospitals than others, some may push for rounding to get favorable treatment

Failure to reject

- Almond, et al. 2010 used the McCrary density test but found no evidence of manipulation
- Ironically, the McCrary density test may fail to reject in a heaping scenario
- In this scenario, the heaping is associated with high mortality children who are outliers compared to newborns both to the left and to the right

Heaping may make it hard to reject the null in the density test

- Density tests may fail to reject in heaping scenarios
- WHY? Because we don't have enough observations for the bins, much use data therefore covering the heaps, making it hard to reject
- Eyeballs, eyeballs, eyeballs

Barreca et al. 2011

"This [heaping at 1500 grams] may be a signal that poor-quality hospitals have relatively high propensities to round birth weights but is also consistent with manipulation of recorded birth weights by doctors, nurses, or parents to obtain favorable treatment for their children. Barreca, et al. 2011 show that this nonrandom heaping leads one to conclude that it is "good" to be strictly less than any 100-g cutoff between 1,000 and 3,000 grams."

Logic of the donut hole RDD specification

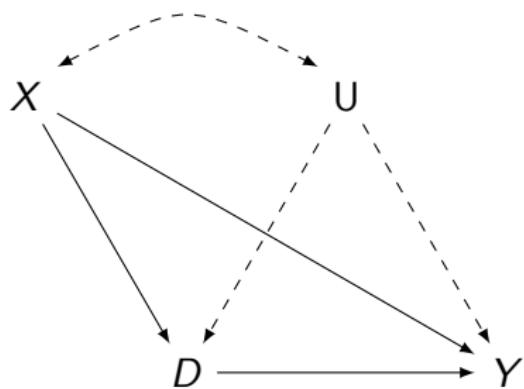
- RDD compares means as we approach c_0 from either direction along X
- Estimates should not logically be sensitive to the observations at the cutoff – if it is, then smoothness may be violated
- Barreca, et al. (2016) suggest dropping units in the vicinity of 1500 grams, and re-estimate the model – if it changes, heaping may be creating a problem
- They call this a “donut” RDD bc you drop the units at the cutoff (the “donut hole”) and estimate your model on the units in the neighborhood instead

Newborn mortality and medical expenditure

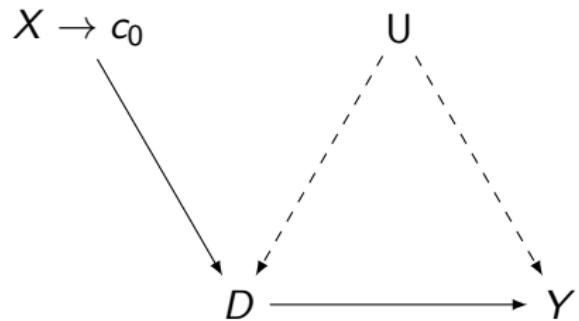
- Dropping units (e.g., trimming) always changes the parameter we're estimating
- In this case, dropping at the threshold reduced sample size by 2%
- But the strength of this practice is that it allows for the possibility that units at the heap differ markedly due to selection bias than those in the surrounding area
- Donut RDD analysis found effect sizes that were 50% smaller than Almond, et al (2010) found
- Be careful with heaping

Endogenous cutoffs

(A) Data generating graph



(B) Limiting graph



Endogeneous cutoffs

- RDD blocks the backdoor path from $D \leftarrow X \leftarrow ? \rightarrow U \rightarrow Y$; but *assumes* that the backdoor path $D \leftarrow U \rightarrow Y$
- But if cutoffs are endogenous, then it is there, which means absent the treatment, smoothness would've been violated *anyway*
- Smoothness isn't guaranteed by an RDD unless $D \leftarrow U \rightarrow Y$ isn't present – which is why it is *the critical identifying assumption*

Endogenous cutoffs

- Examples of endogenous cutoffs
 - Age thresholds used for policy (i.e., person turns 18, and faces more severe penalties for crime) is correlated with other variables that affect the outcome (i.e., graduation, voting rights, etc.)
 - Age 65 is correlated with factors that directly affect healthcare expenditure and mortality such as retirement
- But some of these can be weakly defended with balance tests (observables), or may be directly testable through placebos assuming you have the data

Evaluating smoothness through balance

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

Balance implementation

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

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

Visualizing Balance

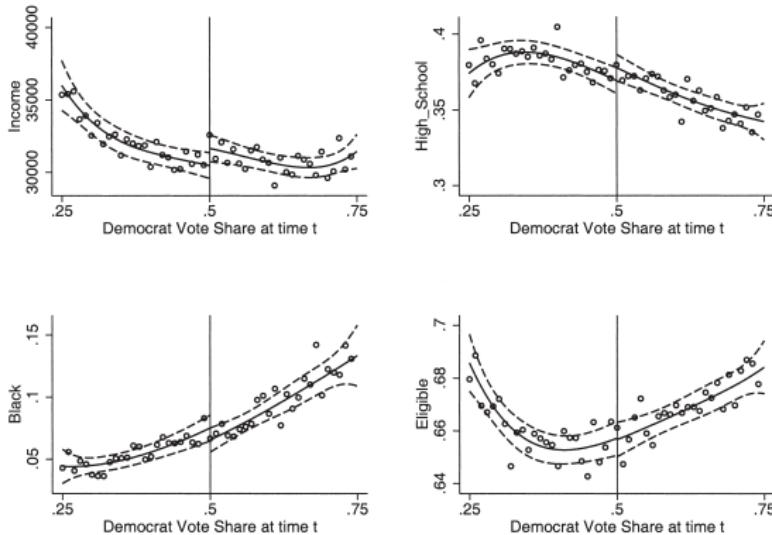


Figure: Figure 3 from Lee, Moretti and Butler (2004), "Do Voters Affect or Elect Policies?" *Quarterly Journal of Economics*. Panels refer to (top left to bottom right) the following district characteristics: real income, percentage with high-school degree, percentage black, percentage eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

Placebos at non-discontinuous points

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

Hidden curriculum	
Foundational causality stuff	
Regression discontinuity designs	
Instrumental variables	
Twoway fixed effects estimator	
Difference-in-differences	
Comparative case studies	
Matching and weighting	
Concluding remarks	
	Introduction
	Sharp Design
	Smoothness, Extrapolation and Estimators
	Testing for violations
	Visualization
	Inference, kernels, bandwidths
	Sub-RDD: Close election designs

Pictures, pictures and more pictures

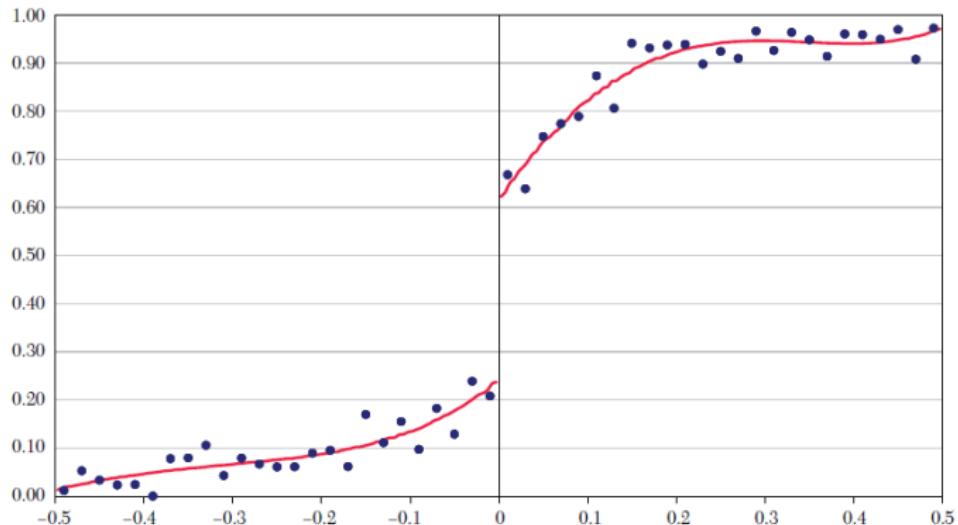
- Synthetic control and RDD are visually intense
- Eyeball tests are rampant (and deservedly) in RDD studies
- Even if your main results are all parametric, you'll still want to present at least some nonparametric style pictures according to Imbens and Lemieux (2010)
- Let's review some of the graphs you have to include

Outcomes

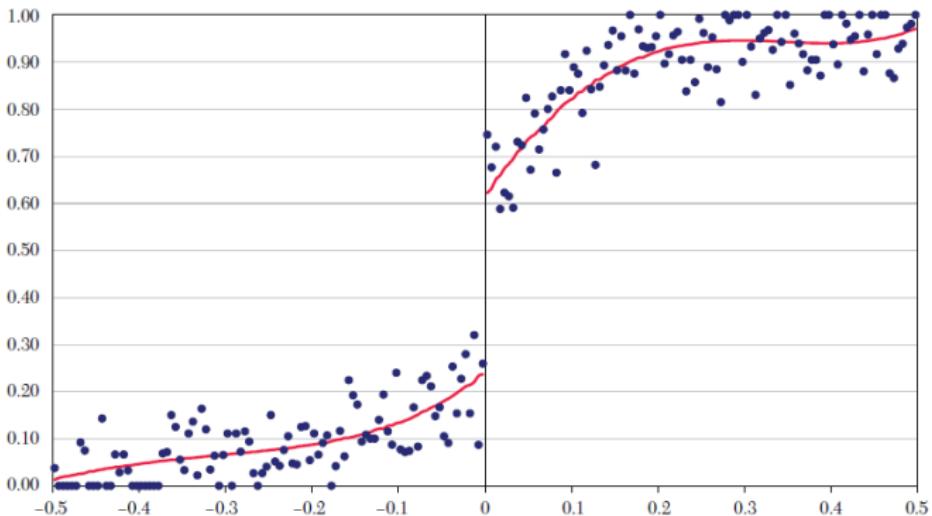
① Outcome by running variable, (X_i):

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

Example: Outcomes by Running Variables



Example: Outcomes by Running Variables with smaller bins



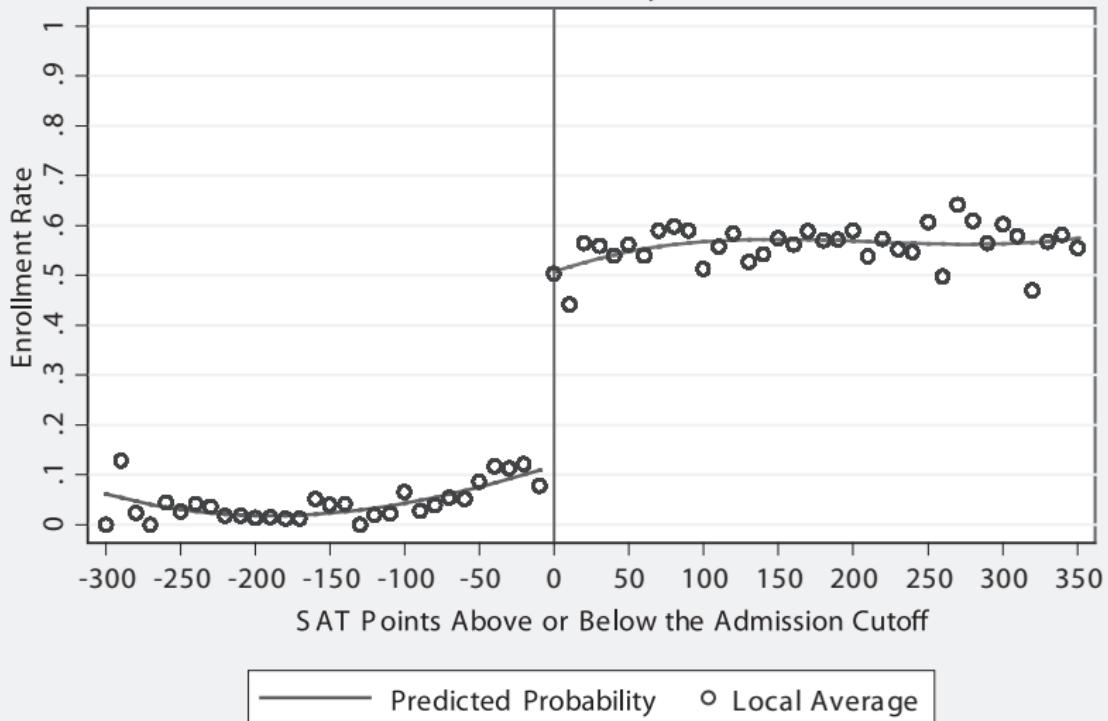
Probability of treatment

② Probability of treatment by running variable if fuzzy RDD

- In a fuzzy RDD, you also want to see that the treatment variable jumps at c_0
- This tells you whether you have a first stage ("bite")
- Let's look at that again from earlier Hoekstra (2008) and enrollment at the flagship

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY

Estimated Discontinuity = 0.388 ($t=10.57$)



McCrary Density

③ Density of the running variable

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

Density of the running variable

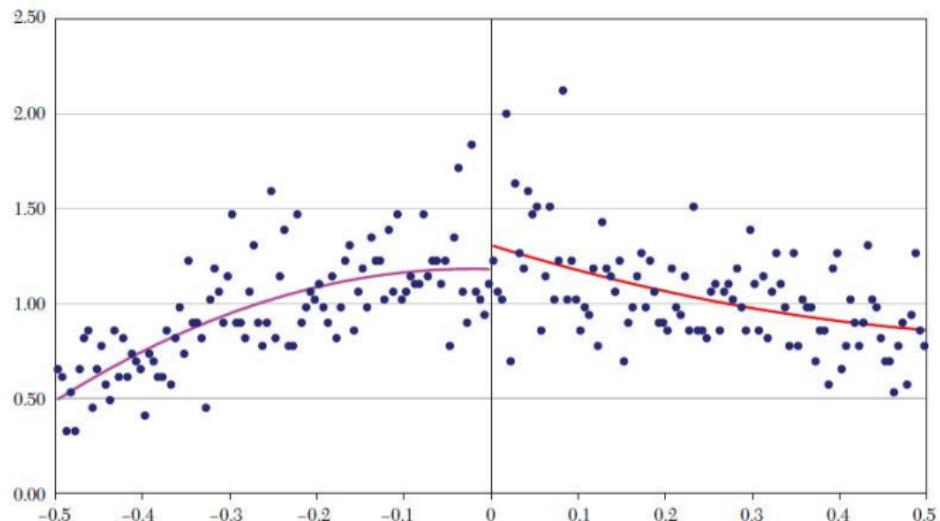


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Balance pictures

④ Covariates by a running variable

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

Example: Covariates by Running Variable

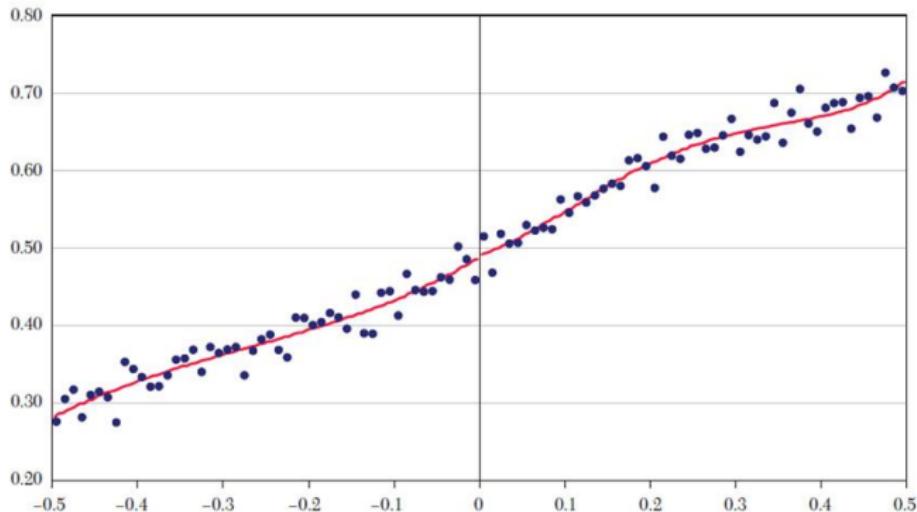


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

Hidden curriculum	
Foundational causality stuff	
Regression discontinuity designs	
Instrumental variables	
Twoway fixed effects estimator	
Difference-in-differences	
Comparative case studies	
Matching and weighting	
Concluding remarks	
	Introduction
	Sharp Design
	Smoothness, Extrapolation and Estimators
	Testing for violations
	Visualization
	Inference, kernels, bandwidths
	Sub-RDD: Close election designs

Inference – honesty

- Lee and Card (2008) and Lee and Lemieux (2010) recommend clustering standard errors on the running variable
- Kolesár and Rothe (2018) provide extensive theoretical and simulation-based evidence that this is not good; you'd be better off just with heteroskedastic robust
- They propose two alternative confidence intervals that achieve correct coverage in large samples – called “honest” (great intro! Still studying this procedure)
- Unavailable in Stata, but is available in R – RDHonest – at <https://github.com/kolesarm/RDHonest>

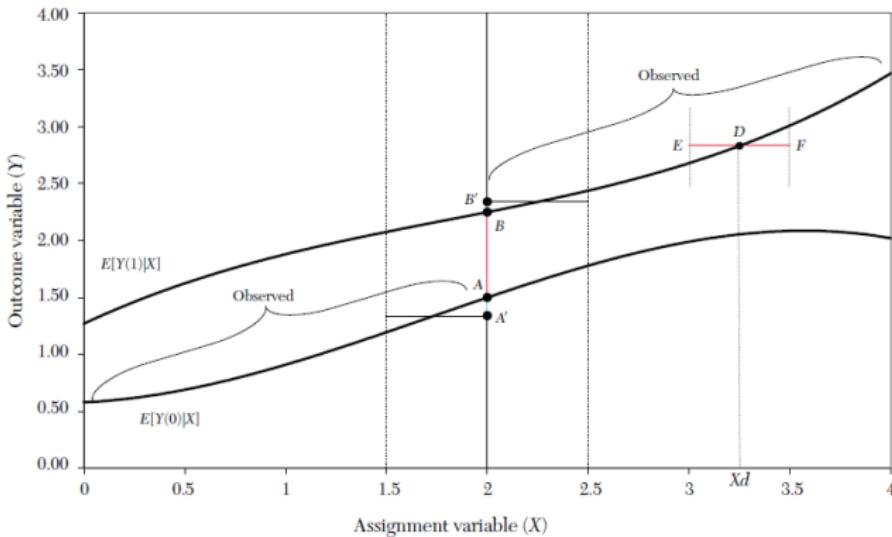
Inference – randomization inference

- Cattaneo, et al. (2015) say to consider that the cutoff is a randomized experiment
- Use randomization inference which is a test of the null of no individual unit level treatment effect at the cutoff

Parametric vs. nonparametric approaches

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

Kernel regression



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

Kernel weighted local polynomial regression

- The nonparametric one-sided kernel estimation problems are called “boundary problems” at the cutoff (Hahn, Todd and Van der Klaauw 2001)
- Kernel estimation (such as lowess) may have poor properties because the point of interest is at a boundary
- They proposed to use “local linear nonparametric regressions” instead

Local linear regression with weights

- Local linear nonparametric regression substantially reduces the bias
- Think of it as a weighted regression restricted to a window – kernel provides the weights to that regression.

$$(\hat{a}, \hat{b}) \equiv_{a,b} \sum_{i=1}^n (y_i - a - b(x_i - c_0))^2 K\left(\frac{x_i - c_0}{h}\right) 1(x_i > c_0)$$

where x_i is the value of the running variable, c_0 is the cutoff, K is a kernel function and $h > 0$ is a suitable bandwidth

Animation of a local linear regression

https://twitter.com/page_eco/status/958687180104245248

Estimation

- Stata's `poly` estimates kernel-weighted local polynomial regressions.
- A rectangular kernel would give the same result as $E[Y]$ at a given bin on X . The triangular kernel gives more importance to observations close to the center.
- This method will be sensitive to how large the bandwidth (window) you choose

Optimal bandwidths

- A rectangular kernel would give the same result as taking $E[Y]$ at a given bin on X whereas the triangular kernel gives more importance to the observations closer to the center.
- While estimating this in a given window of width h around the cutoff is straightforward, it's more difficult to choose this bandwidth (or window), and the method is sensitive to the choice of bandwidth.

Bandwidths

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

Bandwidths

- Imbens and Kalyanaraman (2012), and more recently Calonico, et al. (2017), have proposed methods for estimating “optimal” bandwidths which may differ on either side of the cutoff.
- Calonico, et al (2017) propose local-polynomial regression discontinuity estimators with robust confidence intervals
- Stata ado package and R package are both called `rdrobust`

Hidden curriculum	
Foundational causality stuff	
Regression discontinuity designs	
Instrumental variables	
Twoway fixed effects estimator	
Difference-in-differences	
Comparative case studies	
Matching and weighting	
Concluding remarks	
	Introduction
	Sharp Design
	Smoothness, Extrapolation and Estimators
	Testing for violations
	Visualization
	Inference, kernels, bandwidths
	Sub-RDD: Close election designs

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

DAVID S. LEE
ENRICO MORETTI
MATTHEW J. BUTLER

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

Implementation

- The following paper is a seminal paper in public choice both scientifically and methodologically – the close election RDD
- I call the close election RDD a type of sub-RDD in that it's widely used in political science and economics to the point that it's taken on a life of its own
- Let's take everything we've done and apply it by replicating this paper using programs I've provided

Public choice

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

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

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

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

Falsification of either hypothesis had been hard

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

Outcome is Congress person's liberal voting score

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

Democratic “voteshare” is the running variable

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

Remember these results

TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

Variable	Total effect			Elect component	Affect component
	γ	$\pi_1 (P_{t+1}^D - P_{t+1}^R)$	$\pi_1 [(P_{t+1}^D - P_{t+1}^R)]$	$\pi_0 [P_{t+1}^{sD} - P_{t+1}^{sR}]$	(col. (2) $\hat{\wedge}$ col. (3)) (col. (1)) – (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2 (1.9)	47.6 (1.3)	0.48 (0.02)	22.84 (2.2)	-1.64 (2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

Figure: Lee, Moretti, and Butler 2004, Table 1.

Nonparametric estimation

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

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

Nonparametric estimation (cont.)

- We'll do this estimation just rolling $E[ADA]$ across the running variable *voteshare* visually
- Stata has an option to do this called `cmogram` and it has a lot of useful options, though many people prefer to graph it themselves bc it gives more flexibility.
- We can recreate Figures I, IIA and IIB using it

Future liberal voting score

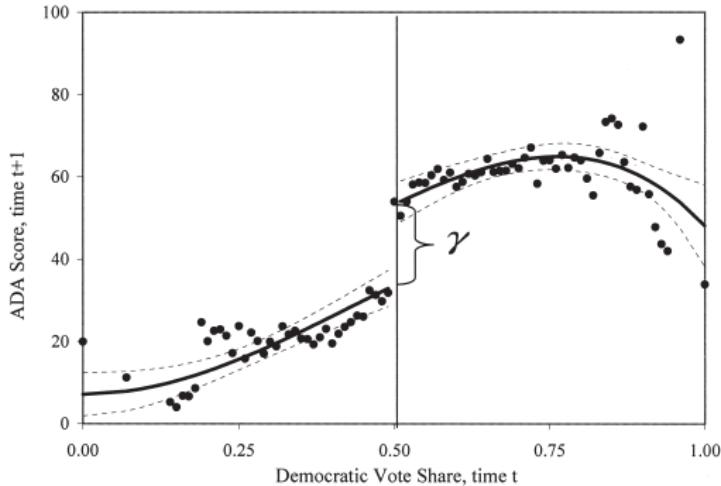


FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

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

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

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

Contemporaneous liberal voting score

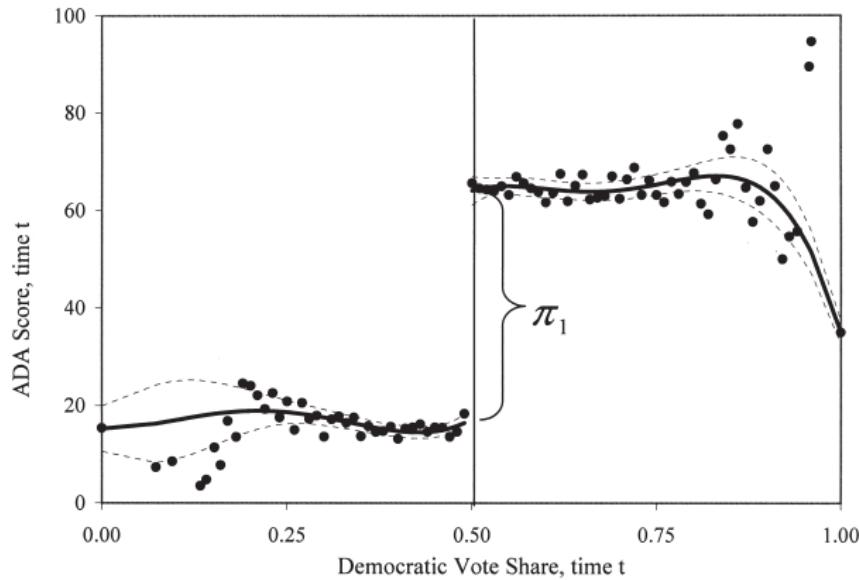


FIGURE IIa
Effect of Party Affiliation: π_1

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

Incumbency advantage

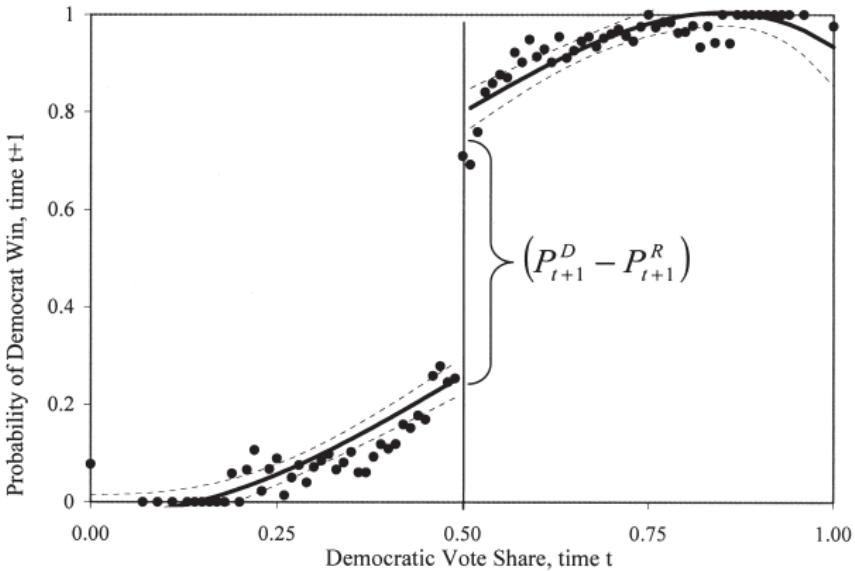


FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P_{t+1}^D - P_{t+1}^R)$

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

Concluding remarks

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

Instrumental variables

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

When is IV used?

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

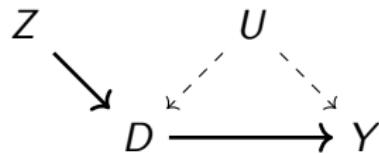
- ① Omitted variable bias
- ② Classical measurement error
- ③ Simultaneity (eg supply and demand)
- ④ Reverse causality
- ⑤ Randomized control trials with noncompliance
- ⑥ Fuzzy RDD

Selection on unobservables



Then D is endogenous due to backdoor path $D \leftarrow U \rightarrow Y$ and causal effect $D \rightarrow Y$ is not identified using the backdoor criterion.

Instruments



Notice how the path from $Z \rightarrow D \leftarrow U \rightarrow Y$ is blocked by a collider.

Phillip Wright

- Philip Wright was a renaissance man - published in JASA, QJE, AER, you name it, while on a very intense teaching load.
- Also published poetry, and even personally published Carl Sandburg's first book of poetry!
- Spent a long time at Tufts
- He was very concerned about the negative effects of tariffs and wrote a book about commodity markets

Elasticity of demand is unidentified

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

Exhibit 1

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

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

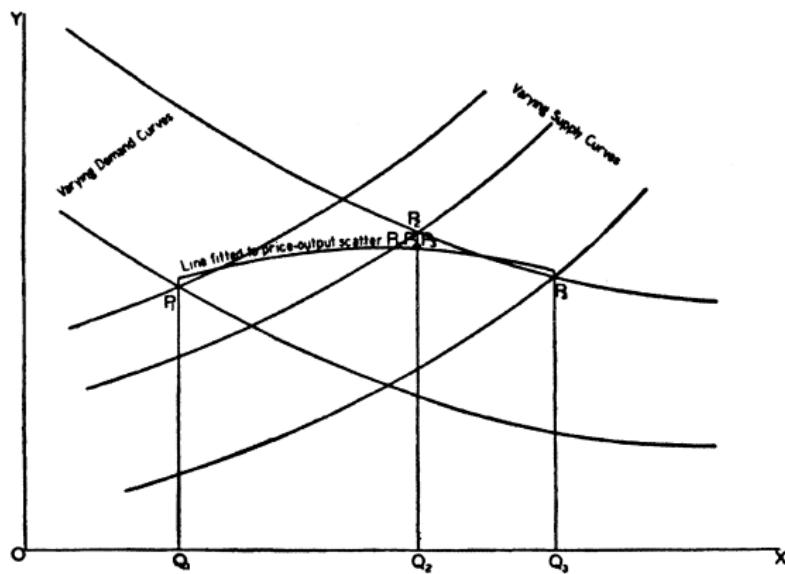


Figure: Wright's graphical demonstration of the identification problem

Sewell Wright

- Sewell was his son, who did *not* go into the family business
- Rather, he decided to become a genius and invent genetics
- Developed path diagrams (which Pearl revived 50 years later for causal inference)
- Father and son engage in letter correspondence as Philip tried to solve the “identification problem”

March 4, 1926.

Dear Sewell:

It may interest you to see a very simple geometric demonstration which I have worked out for you without estimating supply and demand curves without reference to the theory of path coefficients.

Figure: Wright's letter to Sewell, his son

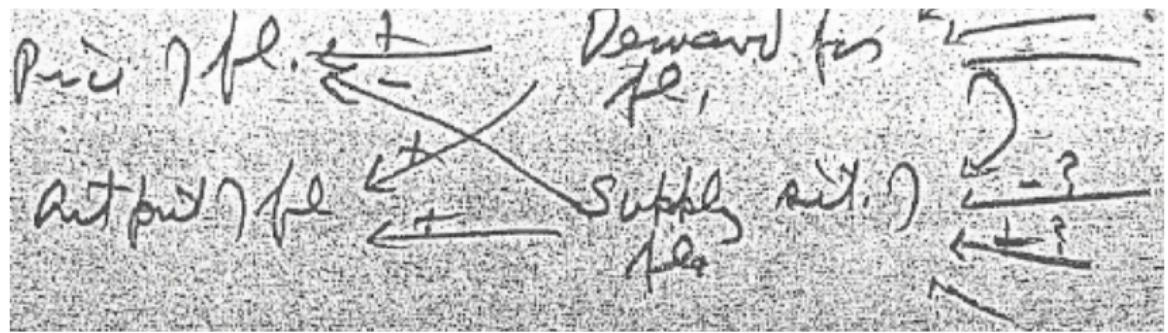


Figure: Recognize these?

QJE Rejects

- QJE misses a chance to make history and rejects his paper proving an IV estimator
- Sticks his proof in Appendix B of 1928 book,
The Tariff on Animal and Vegetable Oils
- His work on IV is ignored, and is then rediscovered 15 years later (e.g., Olav Reiersøl).
- James Stock and others have helped correct the record

Sidebar: stylometric analysis

- Long standing question was who *wrote* Appendix B? Answer according to Stock and Trebbi (2003) using stylometric methods is that Philip *wrote* it.
- But who invented it? It was collaborative, but Sewell acknowledged he didn't know how to handle endogeneity and simultaneity (that was Philip)

Constant treatment effects

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

Heterogenous treatment effects

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

Data requirements

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

Picking a good instrument

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

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

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

Does family size reduce labor supply or is it selection?

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

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

Angrist and Evans cont.

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

Good instruments must be a bit strange

- On its face, it's puzzling that the first two kids' gender predicts labor market participation
- Instrumental variables strategies formalize *strangeness of the instrument*, which is the inference drawn by an intelligent layperson with no particular knowledge of the phenomena or background in statistics.
- You need more information, in other words, otherwise the layperson can't understand what same gender of your children has to do with working

When a good IV strategy finally makes sense

- But then the researchers point out that women whose first two children are of the same gender are more likely to have additional children than women whose first two children are of different genders
- The layperson then asks himself, “Hm. I wonder if the labor market differences are due *solely* to the differences in the number of kids the woman has...”

Sunday Candy is a good instrument

- Let's listen to a few lines from "Ultralight Beam" by Kanye West. Chance the Rapper sings on it and says
*"I made Sunday Candy, I'm never going to hell
I met Kanye West, I'm never going to fail."*
- *Chance the Rapper*
- What does making a song have to do with hell? What does meeting Kanye West have to do with success? Let's consider each in order

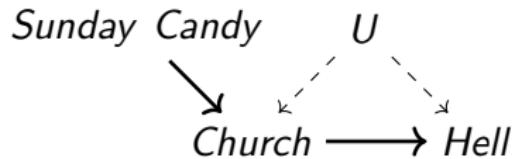
What are we missing?

*"I made Sunday Candy,
I'm never going to hell",*

- There must be more to this story, right?
- So what if it's something like this

*"I made Sunday Candy
this pastor invited me to church on Sunday,
I'm never going to hell"*

Sunday Candy DAG



Kanye West is a bad instrument

- Chance long idolized and was inspired by Kanye West – both Chicago, both very creative hip hop artists
- Kanye West is not a good instrument for Chance's inspiration, though, because Kanye West can singlehandedly make a person's career
- Kanye is not strange enough

Kanye West DAG



Foreshadowing the questions you need to be asking

- ① Is our instrument highly correlated with the treatment? With the outcome? Can you test that?
- ② Are there random elements within the treatment? Why do you think that?
- ③ Is the instrument exogenous? Why do you think that?
- ④ Could the instrument affect outcomes directly? Why do you think that?
- ⑤ Could the instrument be associated with anything that causes the outcome even if it doesn't directly? Why do you think that?

Our causal model: Returns to schooling again

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

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

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

How can IV be used to obtain consistent estimates?

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

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

Consistency

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

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

IV is Consistent if IV Assumptions are Satisfied

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

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

Identifying assumptions and consistency

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

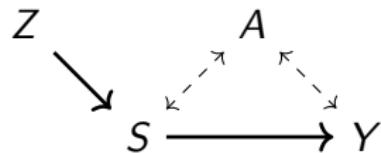
$$\begin{aligned}\text{plim } \widehat{\delta}_{IV} &= \text{plim } \delta + \gamma \frac{\text{Cov}(\eta, Z)}{\text{Cov}(S, Z)} \\ &= \delta\end{aligned}$$

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

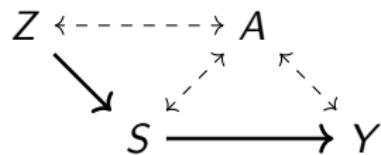
IV Assumptions

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

One of these DAGs is not like the other



(a)



(b)

Notice - the top DAG, *a*, satisfies both exclusion and relevance (i.e., non-zero first stage), but the bottom DAG, *b*, satisfies relevance but not exclusion.

Hidden curriculum	Intuition
Foundational causality stuff	Two stage least squares
Regression discontinuity designs	Weak instruments
Instrumental variables	Practical IV Tips
Two-way fixed effects estimator	Heterogeneity and the LATE
Difference-in-differences	Sub IV: Lottery designs
Comparative case studies	Sub IV: Judge fixed effects
Matching and weighting	Sub IV: Bartik
Concluding remarks	Sub IV: Fuzzy design

Two-stage least squares

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

Two Sample IV

- In a pinch, you can even get by with two different data sets
 - ➊ Dataset 1 needs information on the outcome and the instrument
 - ➋ Dataset 1 needs information on the treatment and the instrument.
- This is known as “Two sample IV” because there are two *samples* involved, rather than the traditional one sample.
- Once we define what IV is measuring carefully, you will see why this works.

Two-stage least squares concepts

- Causal model. Sometimes called the structural model:

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

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

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

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

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

Reduced form

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

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

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

Two-stage least squares

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

$$\begin{aligned} Y_i &= \alpha + \delta S_i + \eta_i \\ S_i &= \gamma + \rho Z_i + \zeta_i \end{aligned}$$

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

Two-stage least squares

Plug in covariance and write out the following:

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

Two-stage least squares

Substitute the causal model definition of Y to get:

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

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

Two-stage least squares

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

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

- Rewrite $\hat{\rho}$ as

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

Two-stage least squares

Then rewrite $\hat{\delta}_{2sls}$

$$\begin{aligned}\hat{\delta}_{2sls} &= \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, S)} = \frac{\hat{\rho}\text{Cov}(Z, Y)}{\hat{\rho}\text{Cov}(Z, S)} = \frac{\hat{\rho}\text{Cov}(Z, Y)}{\hat{\rho}^2\text{Var}(Z)} \\ &= \frac{\text{Cov}(\hat{\rho}Z, Y)}{\text{Var}(\hat{\rho}Z)}\end{aligned}$$

Two-stage least squares

Recall

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

Then

$$\hat{S} = \hat{\gamma} + \hat{\rho} Z$$

Then

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

Proof.

We will show that $\widehat{\delta} \text{Cov}(Y, Z) = \text{Cov}(\widehat{S}, Y)$. I will leave it to you to show that $\text{Var}(\widehat{\delta}Z) = \text{Var}(\widehat{S})$

$$\begin{aligned}\text{Cov}(\widehat{S}, Y) &= E[\widehat{S}Y] - E[\widehat{S}]E[Y] \\&= E(Y[\widehat{\rho} + \widehat{\delta}Z]) - E(Y)E(\widehat{\rho} + \widehat{\delta}Z) \\&= \widehat{\rho}E(Y) + \widehat{\delta}E(YZ) - \widehat{\rho}E(Y) - \widehat{\delta}E(Y)E(Z) \\&= \widehat{\delta}[E(YZ) - E(Y)E(Z)] \\ \text{Cov}(\widehat{S}, Y) &= \widehat{\delta} \text{Cov}(Y, Z)\end{aligned}$$



Intuition of 2SLS

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

Intuition of IV in 2SLS

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

Estimation with software

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

Estimation with software

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

Estimation with software

- Estimate this in Stata using -ivregress 2sls-.
- Estimate this in R -ivreg()- which is in the AER package
- Let's review Card and Graddy.

Weak instruments

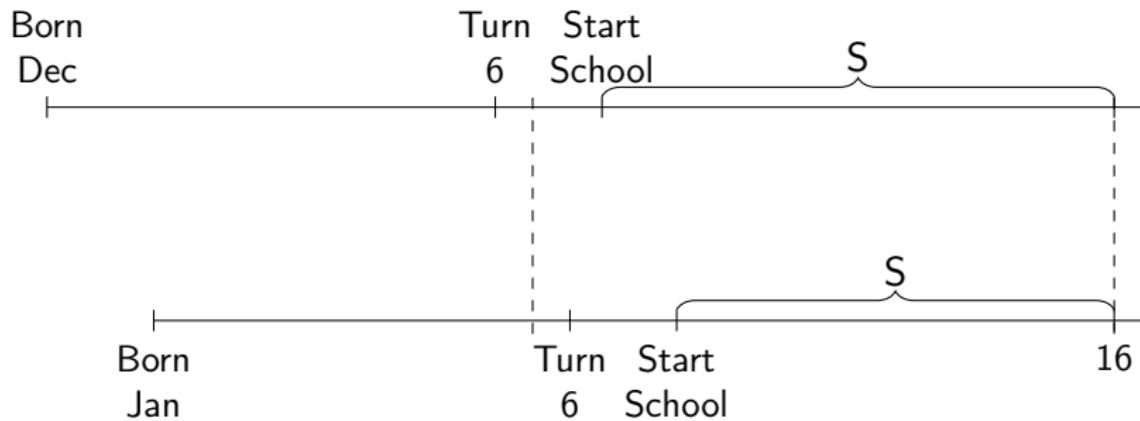
- A weak instrument is one that is not strongly correlated with the endogenous variable in the first stage
- This can happen if the two variables are independent or the sample is small
- If you have a weak instrument, then the bias of 2SLS is centered on the bias of OLS and the cure ends up being worse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

Angrist and Krueger (1991)

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

Compulsory schooling

- In the US, you could drop out of school once you turned 16
- "School districts typically require a student to have turned age six by January 1 of the year in which he or she enters school" (Angrist and Krueger 1991, p. 980)
- Children have different ages when they start school, though, and this creates different lengths of schooling at the time they turn 16 (potential drop out age):



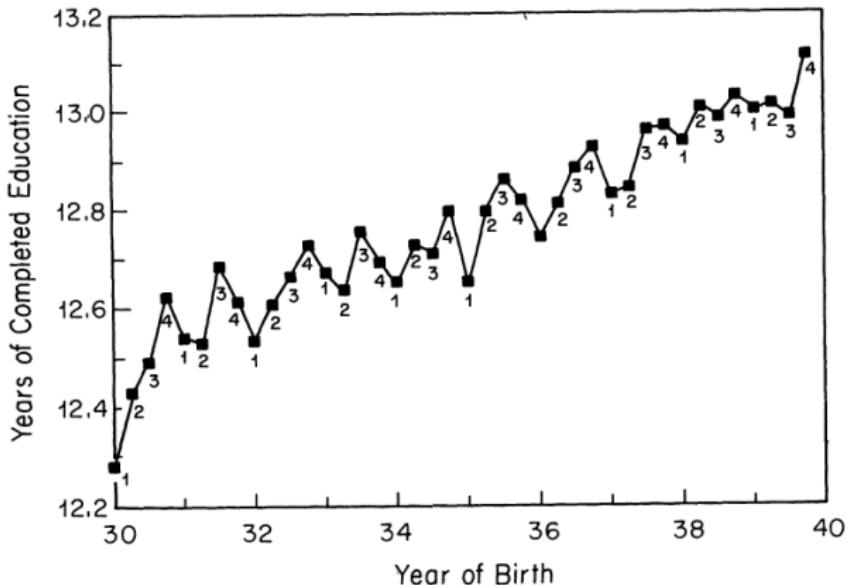
If you're born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

Visuals

- You need good data visualization for IV partly because of the scrutiny around the design
- The two pieces you should be ready to build pictures for are the first stage and the reduced form
- Angrist and Krueger (1991) provide simple, classic and compelling pictures of both

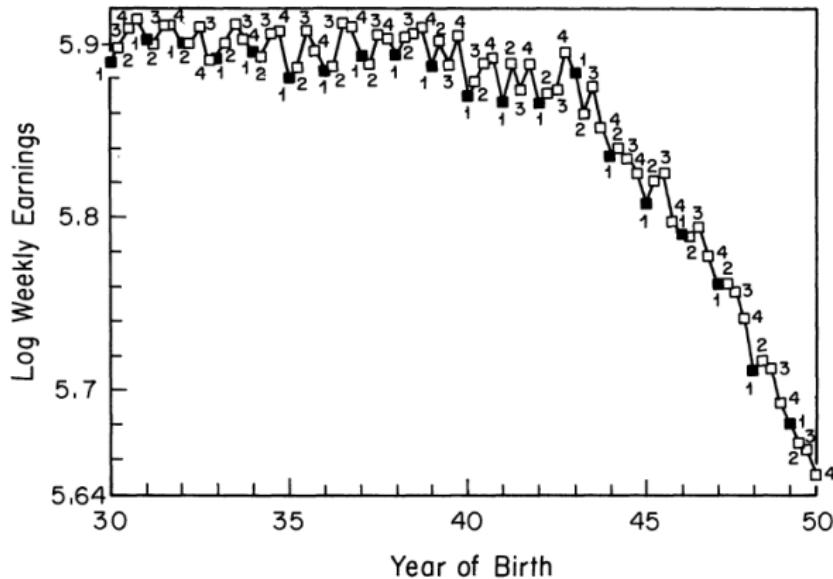
First Stage

Men born earlier in the year have lower schooling. This indicates that there is a first stage. Notice all the 3s and 4s at the top. But then notice how it attenuates over time . . .



Reduced Form

Do differences in schooling due to different quarter of birth translate into different earnings?



Two Stage Least Squares model

- The causal model is

$$Y_i = \delta S_i + \varepsilon$$

- The first stage regression is:

$$S_i = X\pi_{10} + \pi_{11}Z_i + \eta_{1i}$$

- The reduced form regression is:

$$Y_i = X\pi_{20} + \pi_{21}Z_i + \eta_{2i}$$

- The covariate adjusted IV estimator is the sample analog of the ratio, $\frac{\pi_{21}}{\pi_{11}}$

Two Stage Least Squares

- Angrist and Krueger instrument for schooling using three quarter of birth dummies: a dummies for 2nd, 3rd and 4th qob
- Their estimated first-stage regression is:

$$S_i = X\pi_{10} + Z_{1i}\pi_{11} + Z_{2i}\pi_{12} + Z_{3i}\pi_{13} + \eta_1$$

- The second stage is the same as before, but the fitted values are from the new first stage

First stage regression results

Quarter of birth is a strong predictor of total years of education

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect ^a			F-test ^b [P-value]
			I	II	III	
Total years of education	1930–1939	12.79	-0.124 (0.017)	-0.086 (0.017)	-0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	-0.085 (0.012)	-0.035 (0.012)	-0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	-0.019 (0.002)	-0.020 (0.002)	-0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	-0.015 (0.001)	-0.012 (0.001)	-0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	-0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	-0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	-0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	-0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

First stage regression results: Placebos

Completed master's degree	1930–1939	0.09	-0.001	0.002	-0.001	1.7
			(0.001)	(0.001)	(0.001)	[0.1599]
Completed doctoral degree	1930–1939	0.11	0.000	0.004	0.001	3.9
			(0.001)	(0.001)	(0.001)	[0.0091]
	1940–1949	0.03	0.002	0.003	0.000	2.9
			(0.001)	(0.001)	(0.001)	[0.0332]
	1940–1949	0.04	-0.002	0.001	-0.001	4.3
			(0.001)	(0.001)	(0.001)	[0.0050]

a. Standard errors are in parentheses. An $MA(+2, -2)$ trend term was subtracted from each dependent variable. The data set contains men from the 1980 Census, 5 percent Public Use Sample. Sample size is 312,718 for 1930–1939 cohort and is 457,181 for 1940–1949 cohort.

b. F-statistic is for a test of the hypothesis that the quarter-of-birth dummies jointly have no effect.

IV Estimates Birth Cohorts 20-29, 1980 Census

Independent variable	(1) OLS	(2) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)
Race (1 = black)	—	—
SMSA (1 = center city)	—	—
Married (1 = married)	—	—
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No
Age	—	—
Age-squared	—	—
χ^2 [dof]	—	25.4 [29]

Sidebar: Wald estimator

- Recall that 2SLS uses the predicted values from a first stage regression – but we showed that the 2SLS method was equivalent to $\frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}$
- The Wald estimator simply calculates the return to education as the ratio of the difference in earnings by quarter of birth to the difference in years of education by quarter of birth – it's a version of the above
- Formally, $IV_{Wald} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$

Mechanism

- In addition to log weekly wage, they examined the impact of compulsory schooling on log annual salary and weeks worked
- The main impact of compulsory schooling is on the log weekly wage – not on weeks worked

More instruments

To incorporate the cross-state seasonal variation in education, we computed TSLS estimates that use as instruments for education a set of three quarter-of-birth dummies interacted with fifty state-of-birth dummies, in addition to three quarter-of-birth dummies interacted with nine year-of-birth dummies.¹⁸ The estimates also include fifty state-of-birth dummies in the wage equation, so the variability in education used to identify the return to education in the TSLS estimates is solely due to differences by season of birth. Unlike the previous TSLS estimates, the seasonal differences are now allowed to vary by state as well as by birth year.

Problem enters with many quarter of birth interactions

- They want to increase the precision of their 2SLS estimates, so they load up their first stage with more instruments
- Specifications with 30 (quarter of birth \times year) dummy variables and 150 (quarter of birth \times state) instruments
 - What's the intuition here? The effect of quarter of birth may vary by birth year or by state
- It reduced the standard errors, but that comes at a cost of potentially having a weak instruments problem

More instruments

Table VII presents the TSLS and OLS estimates of the new specification for the sample of 40–49 year-old men in the 1980 Census. This is the same sample used in the estimates in Table V. Freeing up the instruments by state of birth and including 50 state-of-birth dummies in the wage equation results in approximately a 40 percent reduction in the standard errors of the TSLS estimates. Furthermore, in the specifications in each of the columns in Table VII, the estimated return to education in the TSLS model is slightly greater than the corresponding TSLS estimate in Table V, whereas in each of the OLS models the return is slightly smaller in Table VII than in Table V. As a consequence, the difference between the TSLS and OLS estimates is of greater significance. For example, the TSLS estimate in column (6) of Table VII is 0.083 with a standard error of 0.010, and the OLS estimate is 0.063 with a standard error of 0.0003: the TSLS estimate is nearly 30 percent greater than the OLS estimate.

More instruments

TABLE VII
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1930-1939: 1980 CENSUS^a

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0673 (0.0003)	0.0928 (0.0093)	0.0673 (0.0003)	0.0907 (0.0107)	0.0628 (0.0003)	0.0831 (0.0095)	0.0628 (0.0003)	0.0811 (0.0109)
Race (1 = black)	—	—	—	—	-0.2547 (0.0043)	-0.2333 (0.0109)	-0.2547 (0.0043)	-0.2354 (0.0122)
SMSA (1 = center city)	—	—	—	—	0.1705 (0.0029)	0.1511 (0.0095)	0.1705 (0.0029)	0.1531 (0.0107)
Married (1 = married)	—	—	—	—	0.2487 (0.0032)	0.2435 (0.0040)	0.2487 (0.0032)	0.2441 (0.0042)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
50 State-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	—	—	-0.0757 (0.0617)	-0.0880 (0.0624)	—	—	-0.0778 (0.0603)	-0.0876 (0.0609)
Age-squared	—	—	0.0008 (0.0007)	0.0009 (0.0007)	—	—	0.0008 (0.0007)	0.0009 (0.0007)
χ^2 [dof]	—	163 [179]	—	161 [177]	—	164 [179]	—	162 [177]

a. Standard errors are in parentheses. Excluded instruments are 30 quarter-of-birth times year-of-birth dummies and 150 quarter-of-birth times state-of-birth interactions. Age and age-squared are measured in quarters of years. Each equation also includes an intercept term. The sample is the same as in Table VI. Sample size is 329,509.

Weak Instruments

- For a long time, applied empiricists were not attentive to the small sample bias of IV
- But in the early 1990s, a number of papers highlighted that IV can be *severely* biased – in particular, when instruments have only a weak correlation with the endogenous variable of interest and when many instruments are used to instrument for one endogenous variable (i.e., there are many overidentifying restrictions).
- In the worst case, if the instruments are so weak that there is no first stage, then the 2SLS sampling distribution is centered on the probability limit of OLS

Causal model

- Let's consider a model with a single endogenous regressor and a simple constant treatment effect (i.e., "just identified")
- The causal model of interest is:

$$Y = \beta X + \nu$$

Matrices and instruments

- We'll sadly need some matrix notation, but I'll try to make it painless.
- The matrix of instrumental variables is Z with the first stage equation:

$$X = Z'\pi + \eta$$

- And let P_z be the project matrix producing residuals from population regression of X on Z

$$P_z = Z(Z'Z)^{-1}Z'$$

Weak instruments and bias towards OLS

- If ν_i and η_i are correlated, estimating the first equation by OLS would lead to biased results, wherein the OLS bias is:

$$E[\beta_{OLS} - \beta] = \frac{Cov(\nu, X)}{Var(X)}$$

- If ν_i and η_i are correlated the OLS bias is therefore: $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$

Deriving the bias of 2SLS

$$\begin{aligned}\hat{\beta}_{2sls} &= (X' P_z X)^{-1} X' P_z Y \\ &= \beta + (X' P_z X)^{-1} X' P_z \nu\end{aligned}$$

substitution of $Y = \beta X + \nu$

2SLS bias

$$\begin{aligned}\widehat{\beta}_{2SLS} - \beta &= (X' P_z X)^{-1} X' P_z \nu \\&= a X' P_z \nu \\&= a[\pi' Z' + \eta'] P_z \nu \\&= a\pi' Z' \nu + a\eta' P_Z \nu \\&= (X' P_Z X)^{-1} \pi' Z' \nu + (X' P_z X)^{-1} \eta' P_z \nu\end{aligned}$$

The bias of 2SLS comes from the non-zero expectation of terms on the right-hand-side even though Z and ν are not correlated.

Taking expectations

- Angrist and Pischke (ch. 4) note that taking expectations of that prior expression is hard because the expectation operator won't pass through $(X'P_zX)^{-1}$.
- However, the expectation of the ratios in the second term can be closely approximated

$$\begin{aligned}\widehat{\beta}_{2sls} - \beta &= (X'P_Z X)^{-1}\pi'Z'\nu + (X'P_z X)^{-1}\eta'P_z\nu \\ E[\widehat{\beta}_{2sls} - \beta] &\approx \left(E[X'P_Z X]\right)^{-1}E[\pi'Z'\nu] + \left(E[X'P_z X]\right)^{-1}E[\eta'P_z\nu]\end{aligned}$$

Approximate bias of 2SLS

We know $E[\pi' Z' \nu] = 0$ and $E[\pi' Z' \eta] = 0$. So letting $E[\eta' P_z \nu] = b$ bc this is hard for me otherwise

$$\begin{aligned} E[\widehat{\beta}_{2SLS} - \beta] &\approx E[X' P_z X]^{-1} b \\ &\approx E(X' Z (Z' Z)^{-1} Z' X)^{-1} b \\ &\approx E[(\pi Z + \eta)' P_z (\pi Z + \eta)]^{-1} b \\ &\approx \left(E(\pi' Z' Z \pi) + E(\eta' P_z \eta)^{-1} \right) b \\ &\approx \left(E(\pi' Z' Z \pi) + E(\eta' P_z \eta)^{-1} \right) E[\eta' P_z \nu] \end{aligned}$$

That last term is what creates the bias so long as η and ν are correlated – which it's because they are that you picked up 2SLS to begin with

First stage F

With some algebra and manipulation, Angrist and Pischke show that the bias of 2SLS is equal to

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\nu\eta}}{\sigma_{\eta}^2} \left[\frac{E(\pi' Z' Z \pi)/Q}{\sigma_{\eta}^2} + 1 \right]^{-1}$$

where the interior term is the population F-statistic for the joint significance of all regressions in the first stage

Weak instruments and bias towards OLS

- Substituting F for that big term, we can derive the approximate bias of 2SLS as:

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\nu\eta}}{\sigma_\eta^2} \frac{1}{F + 1}$$

- Consider the intuition all that work bought us now: if the first stage is weak (i.e, $F \rightarrow 0$), then the bias of 2SLS approaches $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$

Weak instruments and bias towards OLS

- This is the same as the OLS bias as for $\pi = 0$ in the second equation on the earlier slide (i.e., there is no first stage relationship) $\sigma_x^2 = \sigma_\eta^2$ and therefore the OLS bias $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$ becomes $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$.
- But if the first stage is very strong ($F \rightarrow \infty$) then the 2SLS bias is approaching 0.
- Cool thing is – you can test this with an F test on the joint significance of Z in the first stage
- It's absolutely critical therefore that you choose instruments that are strongly correlated with the endogenous regressor, otherwise the cure is worse than the disease

Weak Instruments - Adding More Instruments

- Adding more weak instruments will increase the bias of 2SLS
 - By adding further instruments without predictive power, the first stage F -statistic goes toward zero and the bias increases
 - We will see this more closely when we cover judge fixed effects
- If the model is “just identified” – mean the same number of instrumental variables as there are endogenous covariates – weak instrument bias is less of a problem

Weak instrument problem

- After Angrist and Krueger study, there were new papers highlighting issues related to weak instruments and finite sample bias
- Key papers are Nelson and Startz (1990), Buse (1992), Bekker (1994) and especially Bound, Jaeger and Baker (1995)
- Bound, Jaeger and Baker (1995) highlighted this problem for the Angrist and Krueger study.

Bound, Jaeger and Baker (1995)

Remember, AK present findings from expanding their instruments to include many interactions

- ① Quarter of birth dummies → 3 instruments
- ② Quarter of birth dummies + (quarter of birth) × (year of birth)
+ (quarter of birth) × (state of birth) → 180 instruments

So if any of these are weak, then the approximate bias of 2SLS gets worse

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV	(3) OLS	(4) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)
<i>F</i> (excluded instruments)		13.486		4.747
Partial <i>R</i> ² (excluded instruments, $\times 100$)		.012		.043
<i>F</i> (overidentification)		.932		.775
<i>Age Control Variables</i>				
Age, Age ²	x	x		
9 Year of birth dummies			x	x
<i>Excluded Instruments</i>				
Quarter of birth		x		x
Quarter of birth \times year of birth			x	
Number of excluded instruments	3		30	

Adding more weak instruments reduced the first stage *F*-statistic and increases the bias of 2SLS. Notice its also moved closer to OLS.

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV
Coefficient	.063 (.000)	.083 (.009)
<i>F</i> (excluded instruments)	2.428	
Partial <i>R</i> ² (excluded instruments, ×100)	.133	
<i>F</i> (overidentification)	.919	
<i>Age Control Variables</i>		
Age, Age ²		
9 Year of birth dummies	x	x
<i>Excluded Instruments</i>		
Quarter of birth	x	
Quarter of birth × year of birth	x	
Quarter of birth × state of birth	x	
Number of excluded instruments	180	

More instruments increase precision, but drive down *F*, therefore we know the problem has gotten worse

Guidance on working around weak instruments

- Use a just identified model with your strongest IV
- Use a limited information maximum likelihood estimator (LIML) as it is approximately median unbiased for over identified constant effects models and provides the same asymptotic distribution as 2SLS (under constant effects) with a finite-sample bias reduction.
- Find stronger instruments – easier said than done

Hidden curriculum	Intuition
Foundational causality stuff	Two stage least squares
Regression discontinuity designs	Weak instruments
Instrumental variables	Practical IV Tips
Two-way fixed effects estimator	Heterogeneity and the LATE
Difference-in-differences	Sub IV: Lottery designs
Comparative case studies	Sub IV: Judge fixed effects
Matching and weighting	Sub IV: Bartik
Concluding remarks	Sub IV: Fuzzy design

Look at the reduced form

① Look at the reduced form

- The reduced form is estimated with OLS and is therefore unbiased
- If you can't see the causal relationship of interest in the reduced form, it is probably not there

Report the first stage

- ② Report the first stage (preferably in the same table as your main results)
 - Does it make sense?
 - Do the coefficients have the right magnitude and sign?
 - Please make beautiful IV tables – you'll be celebrated across the land if you do

Report F statistic and OLS

- ③ Report the F -statistic on the excluded instrument(s).
 - Stock, Wright and Yogo (2002) suggest that F -statistics > 10 indicate that you do not have a weak instrument problem – this is not a proof, but more like a rule of thumb
 - If you have more than one endogenous regressor for which you want to instrument, reporting the first stage F -statistic is not enough (because 1 instrument could affect both endogenous variables and the other could have no effect – the model would be under identified). In that case, you want to report the Cragg-Donald EV statistic.
- ④ Report OLS – you said it was biased, but we want to still see it

Table: OLS and 2SLS regressions of Log Earnings on Schooling

Dependent variable	Log wage	
	OLS	2SLS
educ	0.071*** (0.003)	0.124** (0.050)
exper	0.034*** (0.002)	0.056*** (0.020)
black	-0.166*** (0.018)	-0.116** (0.051)
south	-0.132*** (0.015)	-0.113*** (0.023)
married	-0.036*** (0.003)	-0.032*** (0.005)
smsa	0.176*** (0.015)	0.148*** (0.031)
<hr/>		
First Stage Instrument		
College in the county		0.327***
Robust standard error		0.082
F statistic for IV in first stage		15.767
N	3,003	3,003
Mean Dependent Variable	6.262	6.262
Std. Dev. Dependent Variable	0.444	0.444

Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Practical Tips for IV Papers

- ⑤ If you have many IVs, pick your best instrument and report the just identified model (weak instrument problem is much less problematic)
- ⑥ Check over identified 2SLS models with LIML

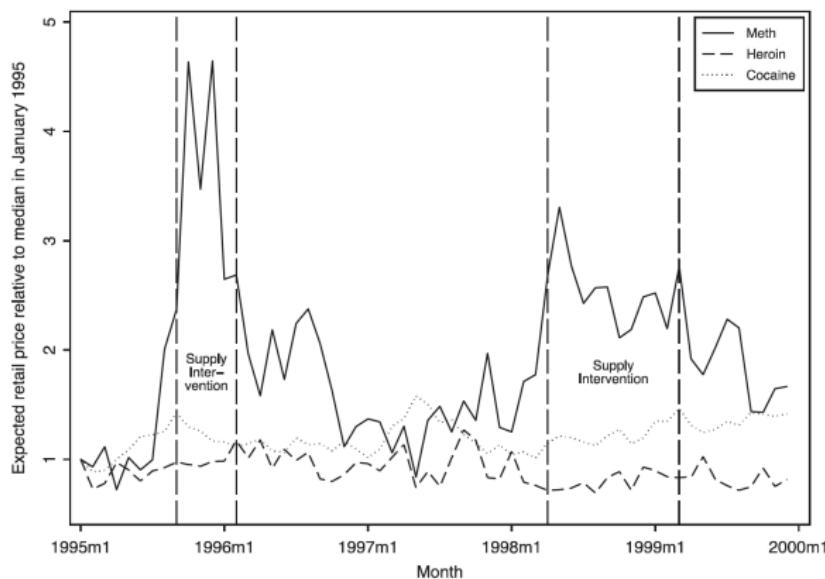
Make beautiful pictures of first stage and reduced form

- ⑦ This cannot be overstated: you must present your main results in beautiful pictures
 - Show pictures of the first stage. Convince the reader something is there. The eyeball is underrated
 - You can't show a second stage with raw data, so instead show pictures of the reduced form.

Visualizing the instrument: supply shocks on meth prices

FIGURE 3

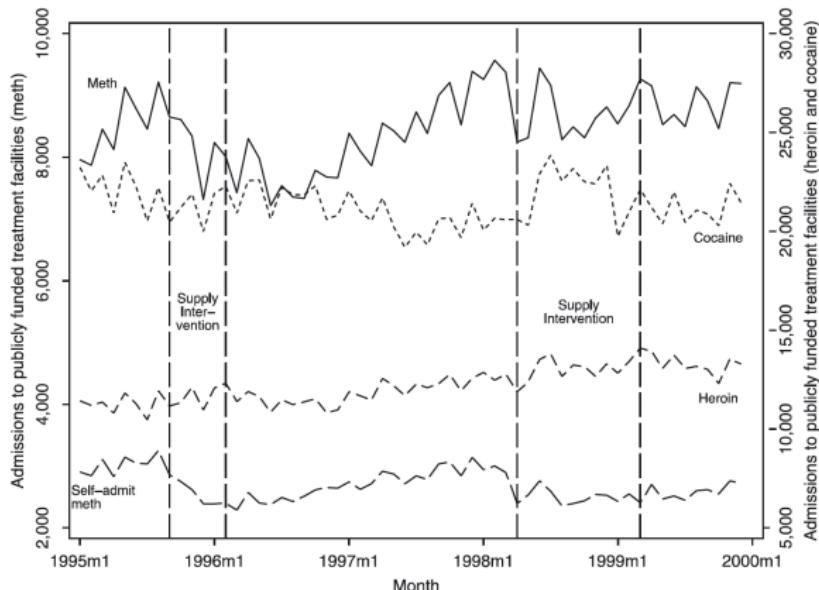
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



Visualizing the first stage

FIGURE 5

Total Admissions to Publicly Funded Treatment Facilities by Drug and Month, Selected States,
Whites, TEDS, Seasonally Adjusted, 1995–1999

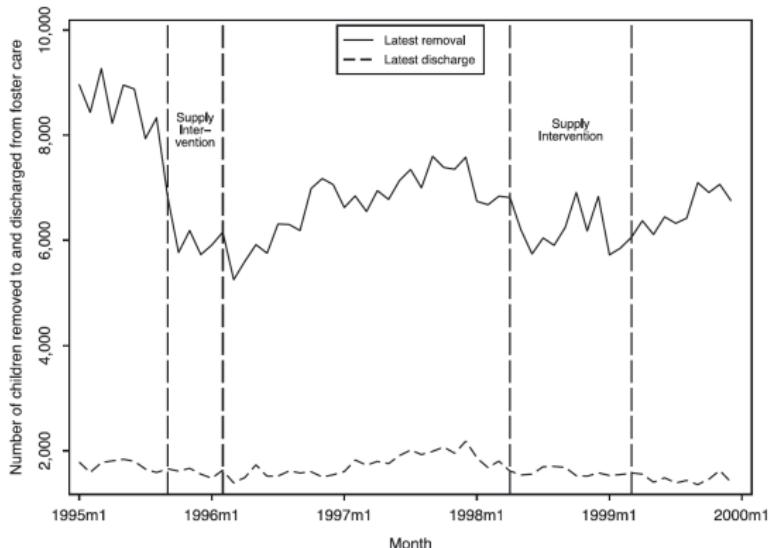


Notes: Authors' calculations from TEDS. Arizona, the District of Columbia, Kentucky, Mississippi, West Virginia, and Wyoming are excluded because of poor data quality. Patients can report the use of more than one drug.

Visualizing the reduced form

FIGURE 4

Number of Children Removed to and Discharged from Foster Care in a Set of Five States by Month, AFCARS, Seasonally Adjusted, 1995–1999



Sources: Authors' calculations from AFCARS. This figure contains AFCARS data only from California, Illinois, Massachusetts, New Jersey, and Vermont. These states form a balanced panel through the entire sample period.

Hidden curriculum	Intuition
Foundational causality stuff	Two stage least squares
Regression discontinuity designs	Weak instruments
Instrumental variables	Practical IV Tips
Two-way fixed effects estimator	Heterogeneity and the LATE
Difference-in-differences	Sub IV: Lottery designs
Comparative case studies	Sub IV: Judge fixed effects
Matching and weighting	Sub IV: Bartik
Concluding remarks	Sub IV: Fuzzy design

Heterogeneous Treatment Effects

- Up to this point, we only considered models where the causal effect was the same for all individuals
 - Constant treatment effects where $Y_i^1 - Y_i^0 = \delta$ for all i units)
- Let's now try to understand what instrumental variables estimation is measuring if treatment effects are *heterogenous*
 - $Y_i^1 - Y_i^0 = \delta_i$ which varies across the population

Why do we care about heterogeneity?

- Heterogeneity, it turns out, makes life interesting and challenging
- There are two issues here:
 - ① We care about internal validity: Does the design successfully uncover causal effects for the population that we are studying?
 - ② We care about external validity: Does the study's results inform us about different populations?
- What parameter did we even estimate using IV when there were heterogenous treatment effects?

Potential outcome notation

“Potential treatment status” (D^j) versus “observed” treatment status (D)

- $D_i^1 = i$ ’s treatment status when $Z_i = 1$
- $D_i^0 = i$ ’s treatment status when $Z_i = 0$

We’ll represent outcomes as a function of both treatment status and instrument status. In other words, $Y_i(D_i = 0, Z_i = 1)$ is represented as $Y_i(0, 1)$

Switching equation

Move from potential treatment status to observed treatment status

$$\begin{aligned} D_i &= D_i^0 + (D_i^1 - D_i^0)Z_i \\ &= \pi_{0i} + \pi_{1i}Z_i + \zeta_i \end{aligned}$$

$$\pi_{0i} = E[D_i^0]$$

$\pi_{1i} = (D_i^1 - D_i^0)$ is the heterogenous causal effect of the IV
on D_i .

$E[\pi_{1i}]$ = The average causal effect of Z_i on D_i

Identifying assumptions under heterogenous treatment effects

- ① Stable Unit Treatment Value Assumption (SUTVA)
- ② Random Assignment
- ③ Exclusion Restriction
- ④ Nonzero First Stage
- ⑤ Monotonicity

Stable Unit Treatment Value Assumption (SUTVA)

Stable Unit Treatment Value Assumption (SUTVA)

If $Z_i = Z'_i$, then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$

If $Z_i = Z'_i$ and $D_i = D'_i$, then $Y_i(\mathbf{D}, \mathbf{Z}) = Y_i(\mathbf{D}', \mathbf{Z}')$

- Potential outcomes for each person i are unrelated to the treatment status of other individuals.
- Example: Your instrument is a randomly generated draft number. If you being drafted makes someone less likely to be drafted, then SUTVA is violated
- In which case, the instrument is related to treatment status of other individuals.

Independence assumption

Independence assumption (e.g., “as good as random assignment”)

$$\{Y_i(D_i^1, 1), Y_i(D_i^0, 0), D_i^1, D_i^0\} \perp\!\!\!\perp Z_i$$

- The IV is independent of the vector of potential outcomes and potential treatment assignments (i.e. “as good as randomly assigned”)
- Example: If your draft number is randomly generated, then your instrument satisfies independence in a way that is trivially true
- It's all about the *randomness* of the instrument, in other words, not the instrument's effect.

Independence

Independence means that the first stage measures the causal effect of Z_i on D_i :

$$\begin{aligned} E[D_i|Z_i = 1] - E[D_i|Z_i = 0] &= E[D_i^1|Z_i = 1] - E[D_i^0|Z_i = 0] \\ &= E[D_i^1 - D_i^0] \end{aligned}$$

Independence

The independence assumption is sufficient for a causal interpretation of the reduced form:

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_i(D_i^1, 1)|Z_i = 1] \\ &\quad - E[Y_i(D_i^0, 0)|Z_i = 0] \\ &= E[Y_i(D_i^1, 1)] - E[Y_i(D_i^0, 0)] \end{aligned}$$

Exclusion Restriction

$$Y(D, Z) = Y(D, Z') \text{ for all } Z, Z', \text{ and for all } D$$

- Any effect of Z on Y must be via the effect of Z on D . In other words, $Y_i(D_i, Z_i)$ is a function of D only. Or formally:

$$Y_i(D_i, 0) = Y_i(D_i, 1) \text{ for } D = 0, 1$$

- Sometimes called the “only through” assumption because you’re assuming the effect of Z on Y is “only through” its effect on D .
- Example: If your draft number (Z) is correlated with earnings only via the instrument’s effect on enrollment in the military, then exclusion holds. If your draft number causes you to invest in more schooling to avoid the draft, then exclusion is violated.
- Recall the DAG and the *missing arrows* from Z to u and from Z to Y .

Exclusion restriction

- Use the exclusion restriction to define potential outcomes indexed solely against treatment status:

$$Y_i^1 = Y_i(1, 1) = Y_i(1, 0)$$

$$Y_i^0 = Y_i(0, 1) = Y_i(0, 0)$$

- Rewrite the switching equation:

$$Y_i = Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)]D_i$$

$$Y_i = Y_i^0 + [Y_i^1 - Y_i^0]D_i$$

- Random coefficients notation for this is:

$$Y_i = \alpha_0 + \delta_i D_i$$

with $\alpha_0 = E[Y_i^0]$ and $\delta_i = Y_i^1 - Y_i^0$

Spotting violations of exclusion is a sport

Watch the gears turn:

- We are interested in causal effect of military service on earnings, and so use draft number are instrument for military service.
- Draft number is generated by a random number generator. Therefore independence is met as draft number is independent of potential outcomes and potential treatment status.
- But, people with higher draft numbers evade draft by investing in schooling. Earnings change for reasons other than military service. Exclusion is violated
- In other words, random lottery numbers (independence) do not imply that the exclusion restriction is satisfied

Strong first stage

Nonzero Average Causal Effect of Z on D

$$E[D_i^1 - D_i^0] \neq 0$$

- D^1 means instrument is turned on, and D^0 means it is turned off. We need treatment to change when instrument changes.
- Z has to have some statistically significant effect on the average probability of treatment
- Example: Check whether a high draft number makes you more likely to get drafted and vice versa
- Finally – a testable assumption. We have data on Z and D

Monotonicity

Monotonicity

Either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all $i = 1, \dots, N$

- Recall that π_{1i} is the reduced form causal effect of the instrumental variable on an individual i 's treatment status.
- Monotonicity requires that the instrumental variable (weakly) operate in the same direction on all individual units.
- Example of a violation: People with high draft number dodge the draft but would have volunteered had they gotten a low number
- In other words, while the instrument may have no effect on some people, all those who are affected are affected *in the same direction* (i.e., positively or negatively, but not both).

Local average treatment effect

If all 1-5 assumptions are satisfied, then IV estimates the **local average treatment effect (LATE)** of D on Y :

$$\delta_{IV,LATE} = \frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D}$$

Estimand

Instrumental variables (IV) estimand:

$$\begin{aligned}\delta_{IV,LATE} &= \frac{E[Y_i(D_i^1, 1) - Y_i(D_i^0, 0)]}{E[D_i^1 - D_i^0]} \\ &= E[(Y_i^1 - Y_i^0)|D_i^1 - D_i^0 = 1]\end{aligned}$$

Local Average Treatment Effect

- The LATE parameters is the average causal effect of D on Y for those whose treatment status was changed by the instrument, Z
- For example, IV estimates the average effect of military service on earnings for the subpopulation who enrolled in military service because of the draft but would not have served otherwise.
- LATE does not tell us what the causal effect of military service was for patriots (volunteers) or those who were exempted from military service for medical reasons

LATE cont.

- We have reviewed the properties of IV with heterogenous treatment effects using a very simple dummy endogenous variable, dummy IV, and no additional controls example.
- The intuition of LATE generalizes to most cases where we have continuous endogenous variables and instruments, and additional control variables.

LATE and subpopulations

The instrument partitions any population into 4 distinct groups:

- ① Compliers: The subpopulation with $D_i^1 = 1$ and $D_i^0 = 0$. Their treatment status is affected by the instrument in the “correct direction”.
- ② Always takers: The subpopulation with $D_i^1 = D_i^0 = 1$. They always take the treatment independently of Z .
- ③ Never takers: The subpopulation with $D_i^1 = D_i^0 = 0$. They never take the treatment independently of Z .
- ④ Defiers: The subpopulation with $D_i^1 = 0$ and $D_i^0 = 1$. Their treatment status is affected by the instrument in the “wrong direction”.

Subpopulations of soldiers

Examples of subpopulations:

- ① Compliers: I only enrolled in the military because I was drafted otherwise I wouldn't have served
- ② Always takers: My family have always served, so I serve regardless of whether I am drafted
- ③ Never takers: I'm a contentious objector so under no circumstances will I serve, even if drafted
- ④ Defiers: When I was drafted, I dodged. But had I not been drafted, I would have served. I can't make up my mind.

Never-Takers

$$D_i^1 - D_i^0 = 0$$

$$Y_i(0, 1) - Y_i(0, 0) = 0$$

By **Exclusion Restriction**, causal effect of Z on Y is zero.

Complier

$$D_i^1 - D_i^0 = 1$$

$$Y_i(1, 1) - Y_i(0, 0) = Y_i(1) - Y_i(0)$$

Average Treatment Effect among Compliers

Defier

$$D_i^1 - D_i^0 = -1$$

$$Y_i(0, 1) - Y_i(1, 0) = Y_i(0) - Y_i(1)$$

By **Monotonicity**, no one in this group

Always-taker

$$D_i^1 - D_i^0 = 0$$

$$Y_i(1, 1) - Y_i(1, 0) = 0$$

By **Exclusion Restriction**, causal effect of Z on Y is zero.

Monotonicity Ensures that there are no defiers

- Why is it important to not have defiers?
 - If there were defiers, effects on compliers could be (partly) canceled out by opposite effects on defiers
 - One could then observe a reduced form which is close to zero even though treatment effects are positive for everyone (but the compliers are pushed in one direction by the instrument and the defiers in the other direction)
- Monotonicity assumes there are no defiers

What Does IV (Not) Estimate?

- As said, with all 5 assumptions satisfied, IV estimates the average treatment effect for *compliers*, or LATE
- Without further assumptions (e.g., constant causal effects), LATE is not informative about effects on never-takers or always-takers because the instrument does not affect their treatment status
- So what? Well, it matters because in most applications, we would be mostly interested in estimating the average treatment effect on the whole population:

$$ATE = E[Y_i^1 - Y_i^0]$$

- But that's not possible usually with IV

Sensitivity to assumptions: exclusion restriction

- Someone at risk of draft (low lottery number) changes education plans to retain draft deferments and avoid conscription.
- Increased bias to IV estimand through two channels:
 - Average direct effect of Z on Y for compliers
 - Average direct effect of Z on Y for noncompliers multiplied by odds of being a non-complier
- Severity depends on:
 - Odds of noncompliance (smaller → less bias)
 - “Strength” of instrument (stronger → less bias)
 - Effect of the alternative channel on Y

Sensitivity to assumptions: Monotonicity violations

- Someone who would have volunteered for Army when not at risk of draft (high lottery number) chooses to avoid military service when at risk of being drafted (low lottery number)
- Bias to IV estimand (multiplication of 2 terms):
 - Proportion defiers relative to compliers
 - Difference in average causal effects of D on Y for compliers and defiers
- Severity depends on:
 - Proportion of defiers (small \rightarrow less bias)
 - “Strength” of instrument (stronger \rightarrow less bias)
 - Variation in effect of D on Y (less \rightarrow less bias)

Summarizing

- The potential outcomes framework gives a more subtle interpretation of what IV is measuring
 - In the constant coefficients world, IV measures δ which is “the” causal effect of D_i on Y_i , and assumed to be the same for all i units
 - In the random coefficients world, IV measures instead an average of heterogeneous causal effects across a particular population – $E[\delta_i]$ for some group of i units
 - IV, therefore, measures the *local average treatment effect* or LATE parameter, which is the average of causal effects across the subpopulation of *compliers*, or those units whose covariate of interest, D_i , is influenced by the instrument.

Summarizing

- Under heterogeneous treatment effects, Angrist and Evans (1996) identify the causal effect of the gender composition of the first two kids on labor supply
- This is not the same thing as identifying the causal effect of children on labor supply; the former is a LATE whereas the latter might be better described as an ATE
- *Ex post* this is probably obvious, but like many obvious things, it wasn't obvious until it was worked out. This was a real breakthrough (see Angrist, Imbens and Rubin 1996; Imbens and Angrist 1994)

IV in Randomized Trials

- In many randomized trials, participation is nonetheless voluntary among those randomly assigned to treatment
- Consequently, noncompliance is not uncommon and without correcting for it, creates selection biases
- IV designs may even be helpful when evaluating a randomized trial, even though treatment was randomly assigned
- The solution is to instrument for treatment with whether you “won the lottery” and estimate LATE

Lottery designs

- The instrument is your randomized lottery
- Examples might be randomized lottery for attending charter schools to study effect of charter schools on educational outcomes, or a randomized voucher to encourage the collection of health information
- Recall Thornton (2008) instrumented for getting HIV results to estimate causal effect of learning one was HIV+ on condom purchases
- We'll discuss two papers from 2012 and 2014 evaluating a lottery-based expansion of Medicaid health insurance on Oregon on numerous health and financial outcomes

Overarching question

- What are the effects of expanding access to public health insurance for low income adults?
 - Magnitudes, and even the signs, associated with that question were uncertain
- Limited existing evidence
 - Institute of Medicine review of evidence was suggestive, but a lot of uncertainty
 - Observational studies are confounded by selection into health insurance
 - Quasi-experimental work often focuses on elderly and children
 - Only one randomized experiment in a developed country: the RAND health insurance experiment
 - 1970s experiment on a general population
 - Randomized cost-sharing, not coverage itself

The Oregon Health Insurance Experiment

Setting: Oregon Health Plan Standard

- Oregon's Medicaid expansion program for poor adults
- Eligibility
 - Poor (<100% federal poverty line) adults 19-64
 - Not eligible for other programs
 - Uninsured > 6 months
 - Legal residents
- Comprehensive coverage (no dental or vision)
- Minimum cost-sharing
- Similar to other states in payments, management
- Closed to new enrollment in 2004

The Oregon Medicaid Experiment

Oregon held a lottery

- Waiver to operate lottery
- 5-week sign-up period, heavy advertising (January to February 2008)
- Low barriers to sign up, no eligibility pre-screening
- Limited information on list
- Randomly drew 30,000 out of 85,000 on list (March–October 2008)
- Those selected given chance to apply
 - Treatment at household level
 - Had to return application within 45 days
 - 60% applied; 50% of those deemed eligible → 10,000 enrollees

Oregon Health Insurance Experiment

- Evaluate effects of Medicaid using lottery as randomized controlled trial (RCT)
 - Intent-to-treat: Reduced form comparison of outcomes between treatment group (lottery selected individuals) and controls (not selected)
 - LATE: IV using lottery as instrument for insurance coverage
 - First stage: about a 25 percentage point increase in insurance coverage
 - Archived analysis plan
 - Massive data collect effort – primary and secondary
- Similar to ACA expansion but limits to generalizability
 - Partial equilibrium vs. General equilibrium
 - Mandate and external validity
 - Oregon vs. other states
 - Short vs. Long-run

Examine Broad Range of Outcomes

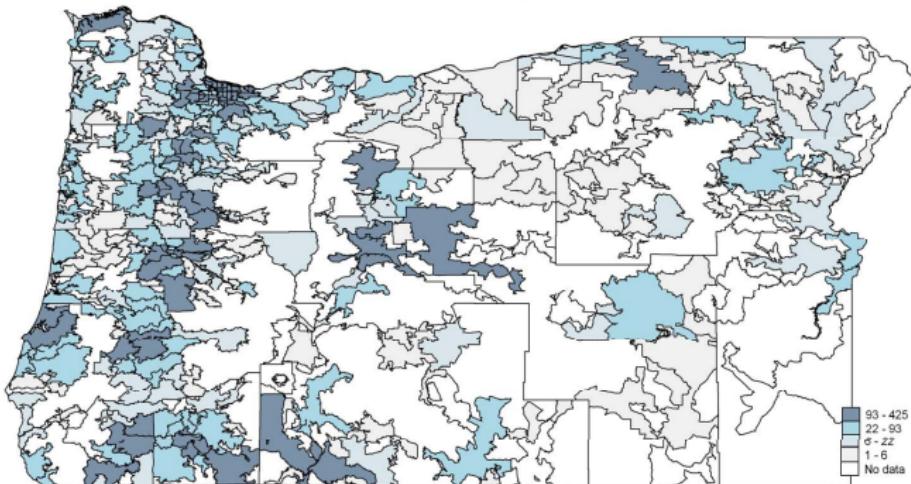
- Costs: Health care utilization
 - Insurance increases resources (income) and lowers price, increasing utilization
 - But improved efficiency (and improved health), decreasing utilization ("offset")
 - Additional uncertainty when comparing Medicaid to no insurance
- Benefits I: Financial risk exposure
 - Insurance supposed to smooth consumption
 - But for very low income, is most care *de jure* or *de facto* free?
- Benefits II: Health
 - Expected to improve (via increased quantity / quality of care)
 - But could discourage health investments ("ex ante moral hazard")

Data

- Pre-randomization demographic information
 - From lottery sign-up
- State administrative records on Medicaid enrollment
 - Primary measure of first stage (i.e., insurance coverage)
- Outcomes
 - Administrative data (~16 months post-notification): Hospital discharge data, mortality, credit reports
 - Mail surveys (~15 months): some questions ask 6-month look-back; some ask current
 - In-person survey and measurements (~25 months): Detailed questionnaires, blood samples, blood pressure, body mass index

Lottery List

Distribution Across Zip Codes



Empirical Framework

- They present reduced form estimates of the causal effect of lottery selection

$$Y_{ihj} = \beta_0 + \beta_1 LOTTERY_h + X_{ih}\beta_2 + V_{ih}\beta_3 + \varepsilon_{ihj}$$

- Validity of experimental design: randomization; balance on treatment and control. This is what readers expect

Empirical framework

- They also present IV results because they want to isolate the causal effect of insurance coverage

$$\begin{aligned} INSURANCE_{ihj} &= \delta_0 + \delta_1 LOTTERY_{ih} + X_{ih}\delta_2 + V_{ih}\delta_3 + \mu_{ihj} \\ y_{ihj} &= \pi_0 + \pi_1 \widehat{INSURANCE}_{ih} + X_{ih}\pi_2 + V_{ih}\pi_3 + v_{ihj} \end{aligned}$$

- Effect of lottery on coverage: about 25 percentage points
- We have independence guaranteed; now we need exclusion: the primary pathway of the lottery must be via being on Medicaid
 - Could affect participation in other programs, but actually small
 - “Warm glow” of winning – especially early
- Analysis plan, multiple inference adjustment

Effect of lottery on coverage (first stage)

	Full sample		Credit subsample		Survey respondents	
	Control mean	Estimated FS	Control mean	Estimated FS	Control mean	Estimated FS
Ever on Medicaid	0.141 (0.004)	0.256 (0.004)	0.135 (0.004)	0.255 (0.004)	0.135 (0.007)	0.290 (0.007)
Ever on OHP Standard	0.027 (0.003)	0.264 (0.003)	0.028 (0.004)	0.264 (0.004)	0.026 (0.005)	0.302 (0.005)
# of Months on Medicaid	1.408 (0.045)	3.355 (0.045)	1.352 (0.055)	3.366 (0.055)	1.509 -0.09	3.943 -0.09
On Medicaid, end of study period	0.106 (0.003)	0.148 (0.003)	0.101 (0.004)	0.151 (0.004)	0.105 (0.006)	0.189 (0.006)
Currently have any insurance (self report)					0.325 0.179 (0.008)	
Currently have private ins. (self report)					0.128 -0.008 (0.005)	
Currently on Medicaid (self report)					0.117 0.197 (0.006)	
Currently on Medicaid					0.093 0.177 (0.006)	

Amy Finkelstein, et al. (2012). "The Oregon Health Insurance Experiment: Evidence from the First Year", *Quarterly Journal of Economics*, vol. 127, issue 3, August.

Effects of Medicaid

Use primary and secondary data to gauge 1-year effects

- Mail surveys: 70,000 surveys at baseline, 12 months
- Administrative data
 - Medicaid enrollment records
 - Statewide Hospital discharge data, 2007-2010
 - Credit report data, 2007-2010
 - Mortality data, 2007-2010

Mail survey data

- **Fielding protocol**
 - ~70,000 people, surveyed at baseline and 12 months later
 - Basic protocol: three-stage mail survey protocol, English/Spanish
 - Intensive protocol on a 30% subsample included additional tracking, mailings, phone attempts (done to adjust for non-response bias)
- **Response rate**
 - Effective response rate = 50%
 - Non-response bias always possible, but response rate and pre-randomization measures in administrative data were balanced between treatment and control

Administrative data

- **Medicaid records**
 - Pre-randomization demographics from list
 - Enrollment records to assess “first stage” (how many of the selected got insurance coverage)
- **Hospital discharge data**
 - Probabilistically matched to list, de-identified at Oregon Health Plan
 - Includes dates and source of admissions, diagnoses, procedures, length of stay, hospital identifier
 - Includes years before and after randomization
- **Other data**
 - Mortality data from Oregon death records
 - Credit report data, probabilistically matched, de-identified

Sample

- 89,824 unique individuals on the waiting list
- Sample exclusions (based on pre-randomization data only)
 - Ineligible for OHP Standard (out of state address, age, etc.)
 - Individuals with institutional addresses on list
- Final sample: 79,922 individuals out of 66,385 households
 - 29,834 treated individuals (surveyed 29,589)
 - 40,088 control individuals (surveyed 28,816)

Sample characteristics

Variable	Mean	Variable	Mean
Panel A: Full sample			
% Female	0.56	Average Age	41
Panel B: Survey responders only			
<i>Demographics:</i>		<i>Health Status: Ever diagnosed with:</i>	
% White	0.82	Diabetes	0.18
% Black	0.04	Asthma	0.28
% Spanish/Hispanic/Latino	0.12	High Blood Pressure	0.40
% High school or less	0.67	Emphysema or Chronic Bronchitis	0.13
% don't currently work	0.55	Depression	0.56
<i>Determinants of eligibility:</i>			
Average hh income (2008)	13,050	% with any insurance	0.33
% below Federal poverty line	0.68	% with private insurance	0.13

Outcomes

- **Access and use of care**
 - Is access to care improved? Do the insured use more care? Is there a shift in the types of care being used?
 - Mail surveys and hospital discharge data
- **Financial strain**
 - How much does insurance protect against financial strain?
 - What are the out-of-pocket implications?
 - Mail surveys and credit reports
- **Health**
 - What are the short-term impacts on self-reported physical and mental health?
 - Mail surveys and vital statistics (mortality)

Effect of lottery on coverage

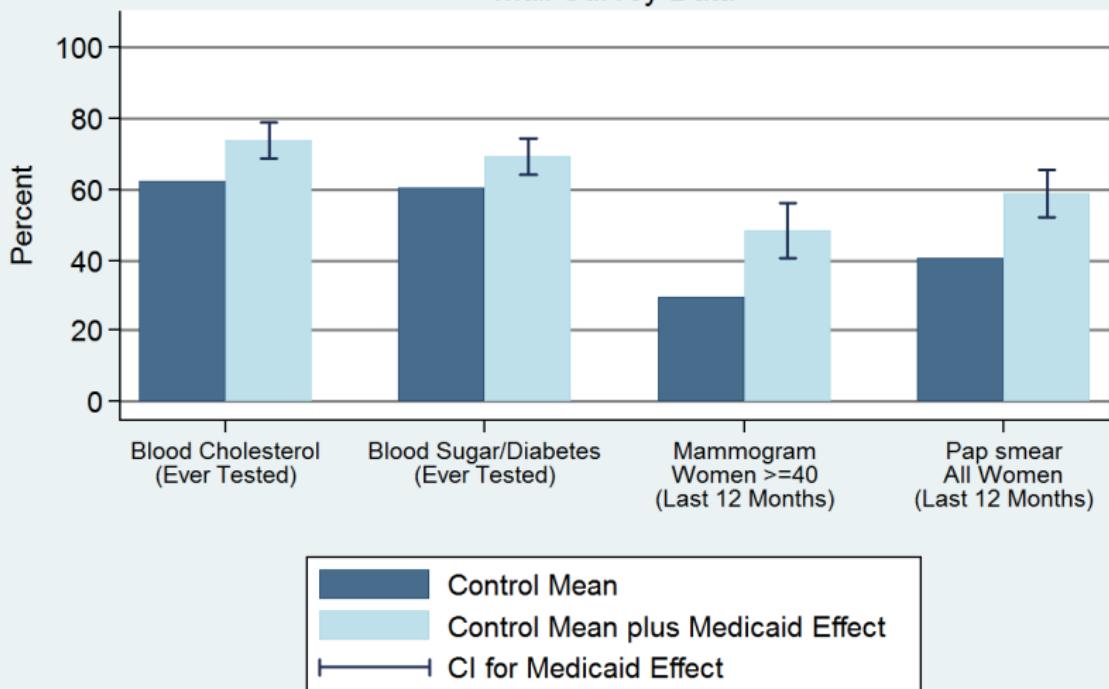
Gaining insurance resulted in better access to care and higher satisfaction with care (conditional on actually getting care)

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Have a usual place of care	49.9%	+9.9%	+33.9%	.0001
Have a personal doctor	49.0%	+8.1%	+28.0%	.0001
Got all needed health care	68.4%	+6.9%	+23.9%	.0001
Got all needed prescriptions	76.5%	+5.6%	+19.5%	.0001
Satisfied with quality of care	70.8%	+4.3%	+14.2%	.001

SOURCE: Survey data

Preventive Care

Mail Survey Data



Effect of lottery on coverage

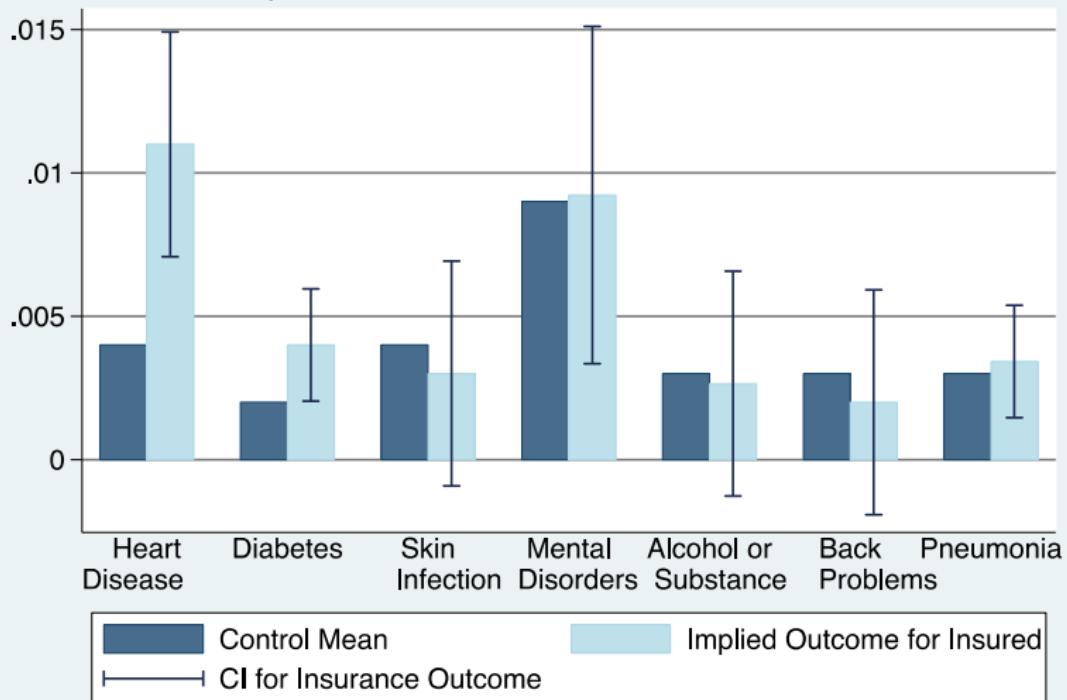
Gaining insurance resulted in increased probability of hospital admissions, primarily driven by non-emergency department admissions

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Any hospital admission	6.7%	+.50%	+2.1%	.004
--Admits through ED	4.8%	+.2%	+.7%	.265
--Admits NOT through ED	2.9%	+.4%	+1.6%	.002

SOURCE: Hospital Discharge Data

Overall, this represents a 30% higher probability of admission, although admissions are still rare events

Hospital Utilization for Selected Conditions



Summary: Access and use of care

- Overall, utilization and costs went up relative to controls
 - 30% increase in probability of an inpatient admission
 - 35% increase in probability of an outpatient visit
 - 15% increase in probability of taking prescription medications
 - Total \$777 increase in average spending (a 25% increase)
- With this increased spending, those who gained insurance were
 - 35% more likely to get all needed care
 - 25% more likely to get all needed medications
 - Far more likely to follow preventive care guidelines, such as mammograms (60%) and PAP tests (45%)

Results: Financial Strain

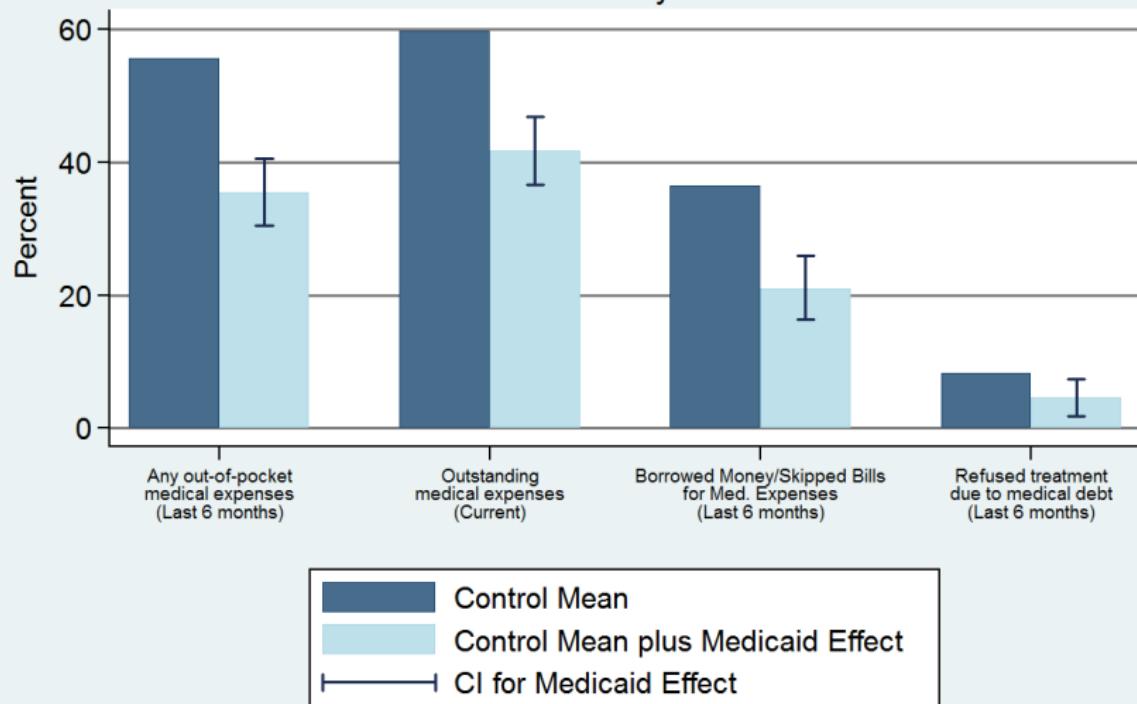
Gaining insurance resulted in a reduced probability of having medical collections in credit reports, and in lower amounts owed

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Had a bankruptcy	1.4%	+0.2%	+0.9%	.358
Had a collection	50.0%	-1.2%	-4.8%	.013
--Medical collections	28.1%	-1.6%	-6.4%	.0001
--Non-medical collections	39.2%	-0.5	-1.8%	.455
\$ owed medical collections	\$1,999	-\$99	-\$390	.025

Source: Credit report data

Self-reported Financial Strain

Mail Survey Data



Summary: Financial Strain

- Overall, reductions in collections on credit reports were evident
 - 25% decrease in probability of a medical collection
 - Those with a collection owed significantly less
- Household financial strain related to medical costs was mitigated
 - Substantial reduction across all financial strain measures
 - Captures “informal channels” people use to make it work
- Implications for both patients and providers
 - Only 2% of bills sent to collections are ever paid

Results: Self-reported health

Self-reported measures showed significant improvements one year after randomization

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Health good, v good, excellent	54.8%	+3.9%	+13.3%	.0001
Health stable or improving	71.4%	+3.3%	+11.3%	.0001
Depression screen NEGATIVE	67.1%	+2.3%	+7.8%	.003
CDC Healthy Days (physical)	21.86	+.381	+1.31	.018
CDC Healthy Days (mental)	18.73	+.603	+2.08	.003

Source: Survey data

Summary: Self-reported health

- Overall, big improvements in self-reported physical and mental health
 - 25% increase in probability of good, very good or excellent health
 - 10% decrease in probability of screening for depression
- Physical health measures open to several interpretations
 - Improvements consistent with findings of increased utilization, better access, and improved quality
 - BUT in their baseline surveys, results appeared shortly after coverage ($\sim 2/3$ rds magnitude of full result)
 - May suggest increase in *perception* of well-being rather than physical health
- Biomarker data can shed light on this issue

Discussion

- At 1 year, found increases in utilization, reductions in financial strain, and improvements in self-reported health
 - Medicaid expansion had benefits and costs – didn't "pay for itself"
 - Confirmed biases inherent in observational studies – would have estimated bigger increases in use and smaller improvements in outcomes
- Policy-makers may have different views on value of different aspects of improved well-being
 - "I have an incredible amount of fear because I don't know if the cancer has spread or not."
 - "A lot of times I wanted to rob a bank so I could pay for the medicine I was just so scared . . . People with cancer either have a good chance or no chance. In my case it's hard to recover from lung cancer but it's possible. Insurance took so long to kick in that I didn't think I would get it. Now there is a big bright light shining on me." (Anecdotes)
- Important to have broad evidence on multifaceted effects of Medicaid expansions

Baicker, Katherine, et al. (2014). "The Oregon Experiment – Effects of Medicaid on Clinical Outcomes", *The New England Journal of Medicine*.

In-person data collection

- Questionnaire and health examination including
 - Survey questions
 - Anthropometric and blood pressure measurement
 - Dried blood spot collection
 - Catalog of all medications
- Fielded between September 2009 and December 2010
 - Average response ~25 months after lottery began
- Limited to Portland area: 20,745 person sample
- 12,229 interviews for effective response rate of 73%

Analytic approach

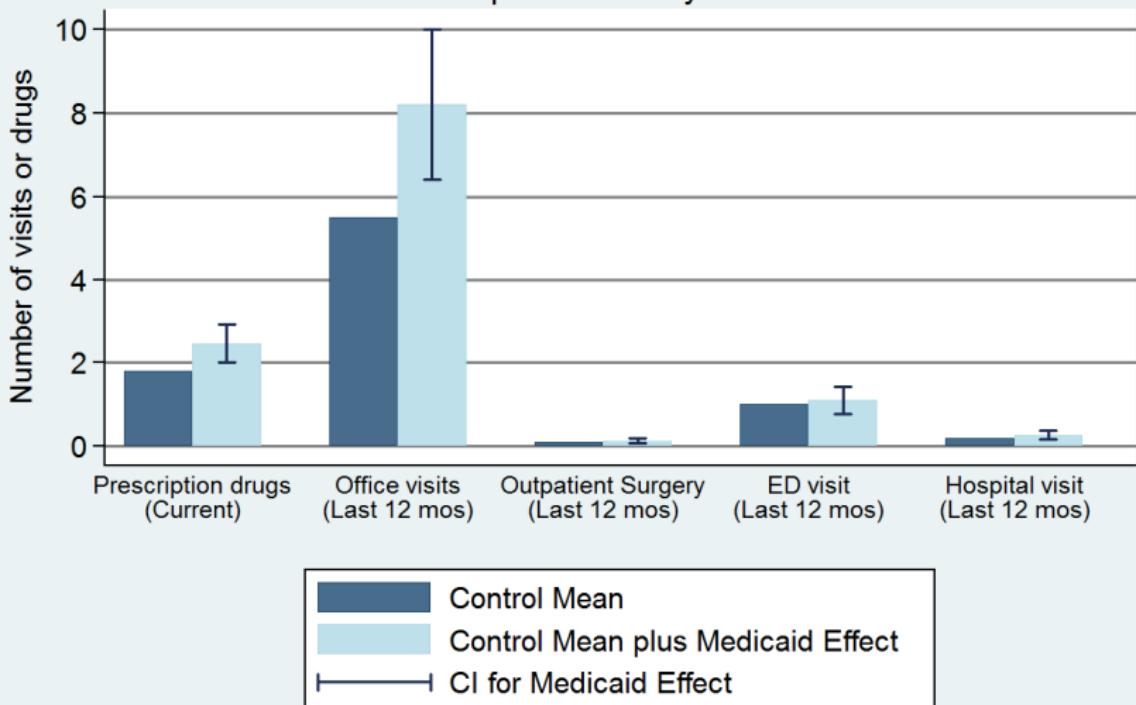
- Intent to treat effect of *lottery selection*
 - Comparing all selected with all not selected
 - Random treatment assignment
 - No differential selection for outcome measurement
- Local average treatment effect on *Medicaid coverage*
 - Using lottery selection as an instrument for coverage
 - ~24 percentage point increase in Medicaid enrollment
 - No change in private insurance (no crowd-out)
 - No effect of lottery except via Medicaid coverage
- Statistical inference is the same for both

Results

- ① *Health care use*
- ② Financial strain
- ③ Clinical health outcomes

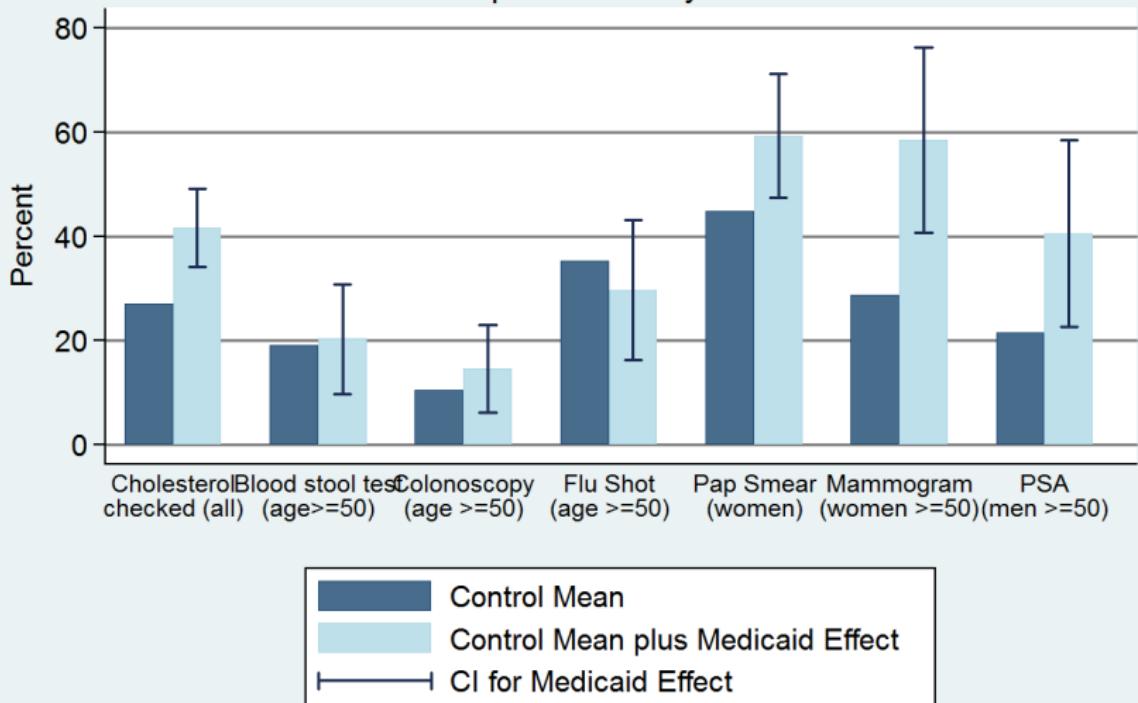
Health Care Utilization

Inperson Survey Data



Preventive Care (Last 12 Months)

Inperson Survey Data



Health care use results

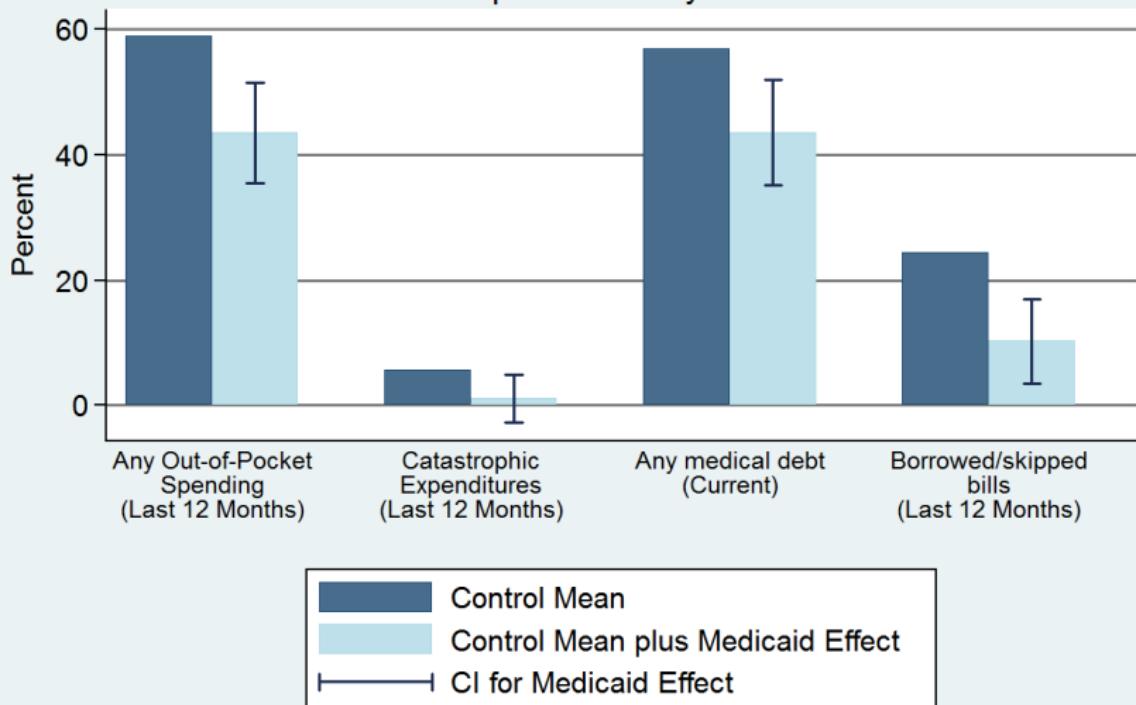
- Increases in use in various settings
 - Increases in probability and number of outpatient visits
 - Increases in probability and number of prescription drugs
 - No discernible change in hospital or ED use (imprecise)
- Increases in preventive care across range of services
- Increases in perceived access and quality
- Implied 35% increase in spending for insured

Results

- ① Health care use
- ② *Financial strain*
- ③ Clinical health outcomes

Financial Hardship

Inperson Survey Data



Financial Hardship Results

- Reduction in strain, out-of-pocket (OOP), money owed
 - Substantial reduction across measures
 - Elimination of catastrophic OOP health spending
- Implications for distribution of burden/benefits
 - Some borne by patients, some by providers
 - Non-financial burden of medical expenses and debt

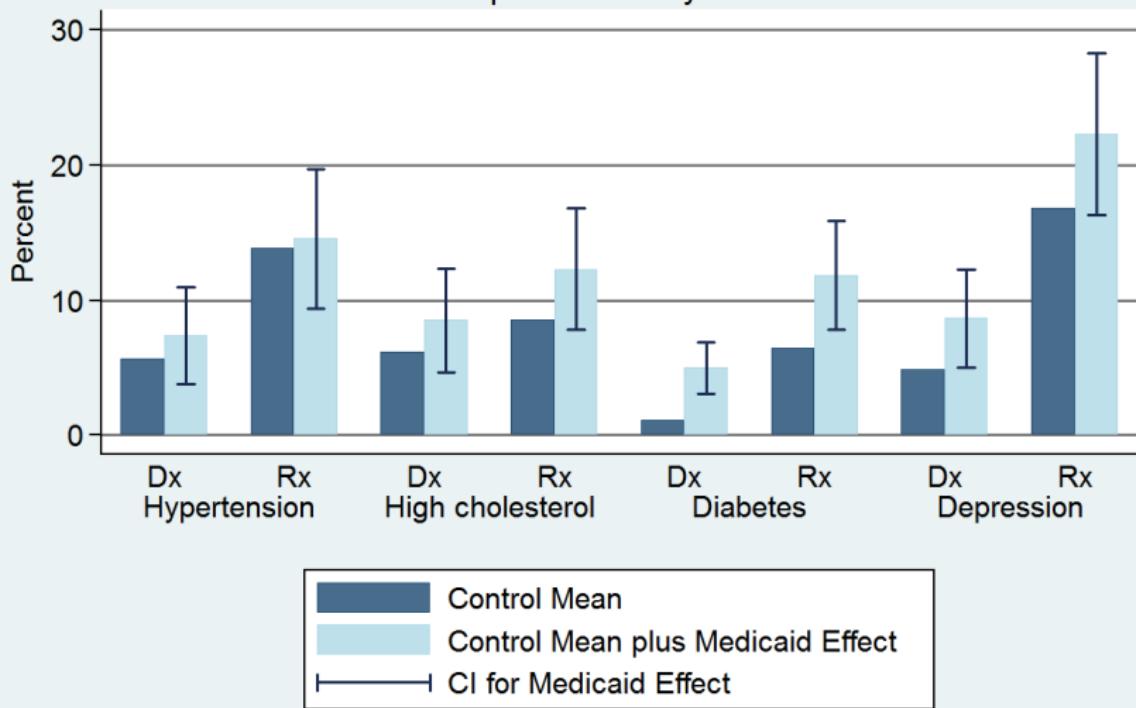
Results

- ① Health care use
- ② Financial strain
- ③ *Clinical health outcomes*

Focusing on specific conditions

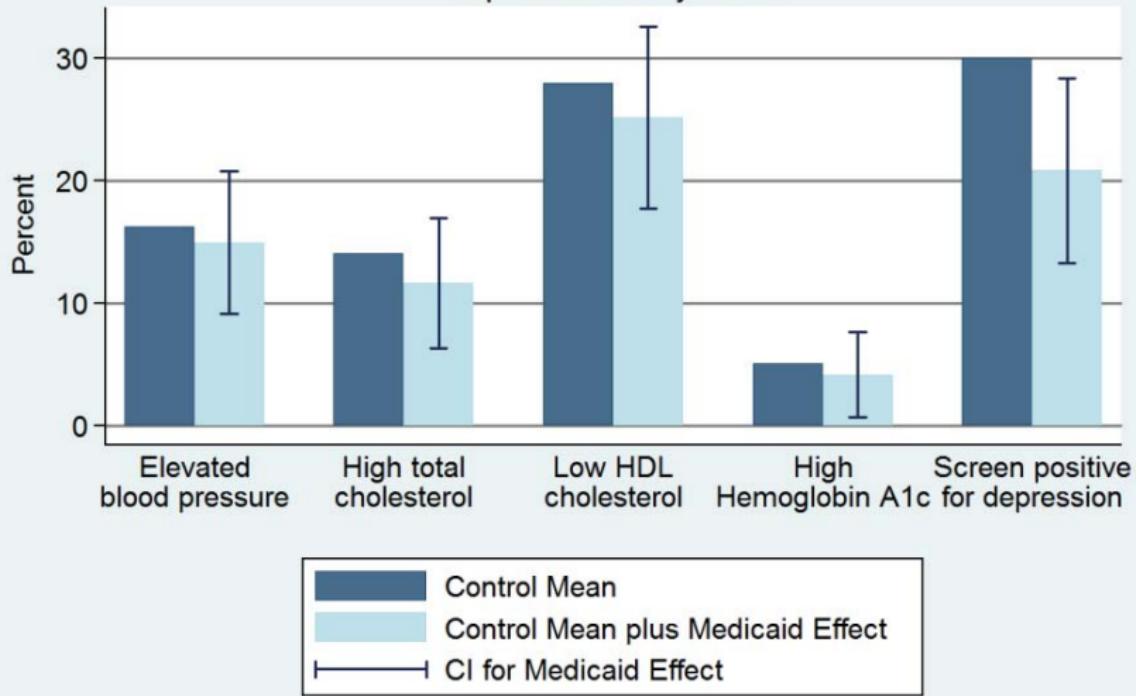
- Measured:
 - Blood pressure
 - Cholesterol levels
 - Glycated hemoglobin
 - Depression
- Reasons for selecting these:
 - Reasonably prevalent conditions
 - Clinically effective medications exist
 - Markers of longer term risk of cardiovascular disease
 - Can be measured by trained interviewers and lab tests
- A limited window into health status

Post-lottery Diagnosis (Dx) and Current Medication (Rx) Inperson Survey Data

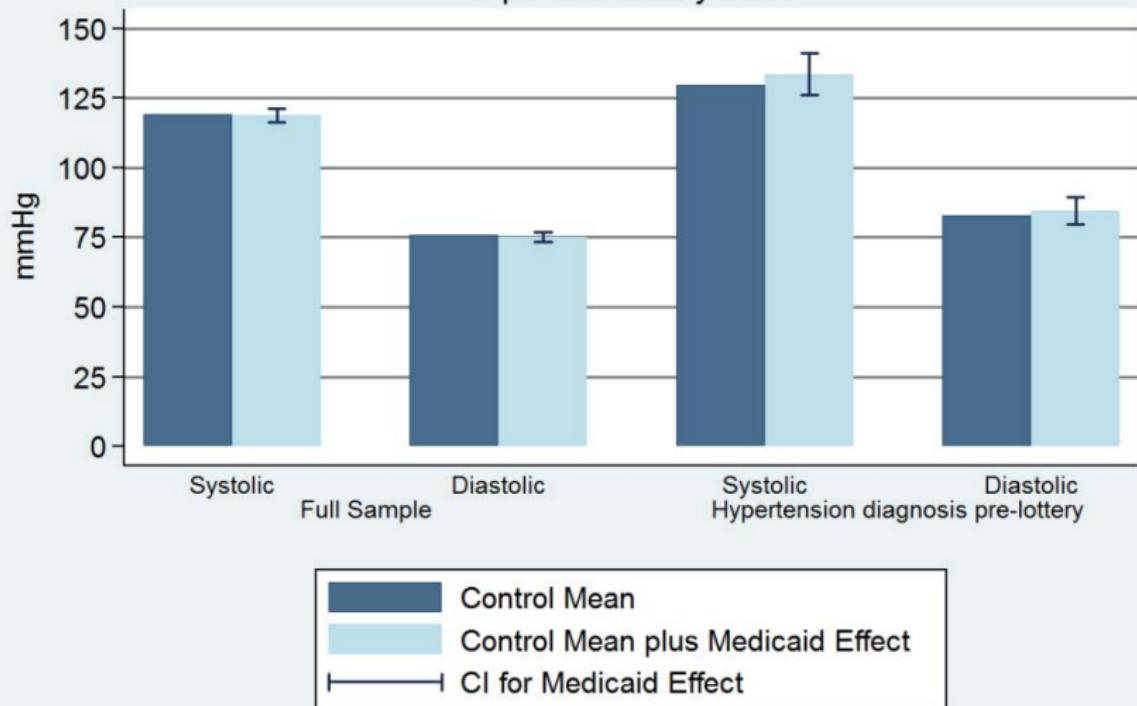


Current Clinical Measures

Inperson Survey Data

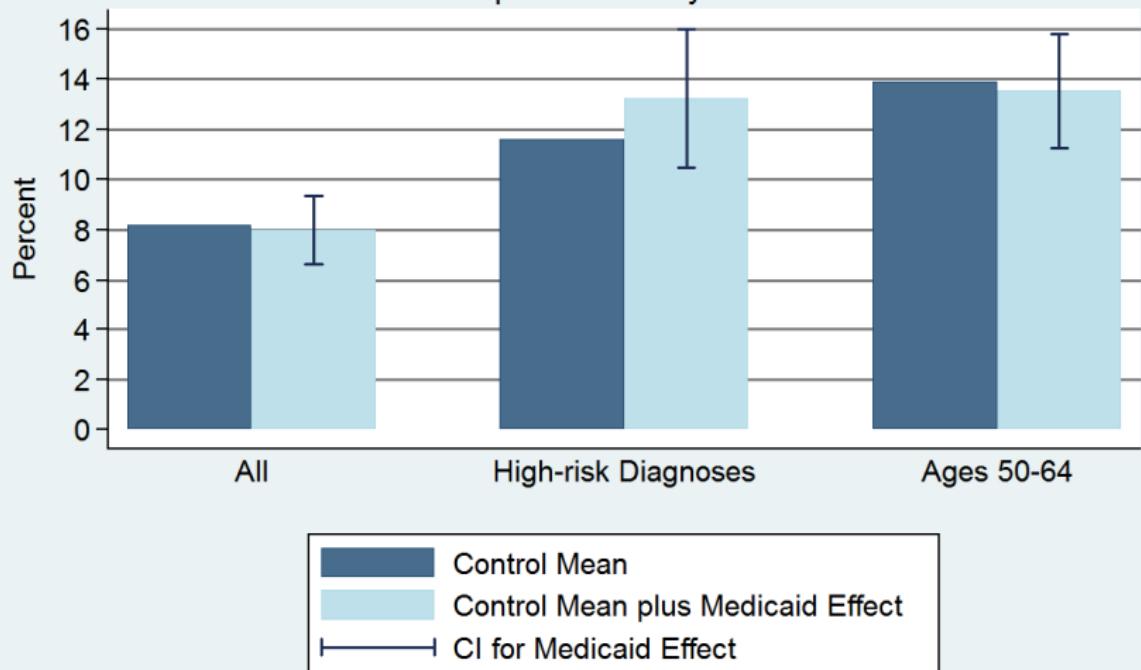


Blood Pressure Inperson Survey Data



Framingham Risk Scores

Inperson Survey Data



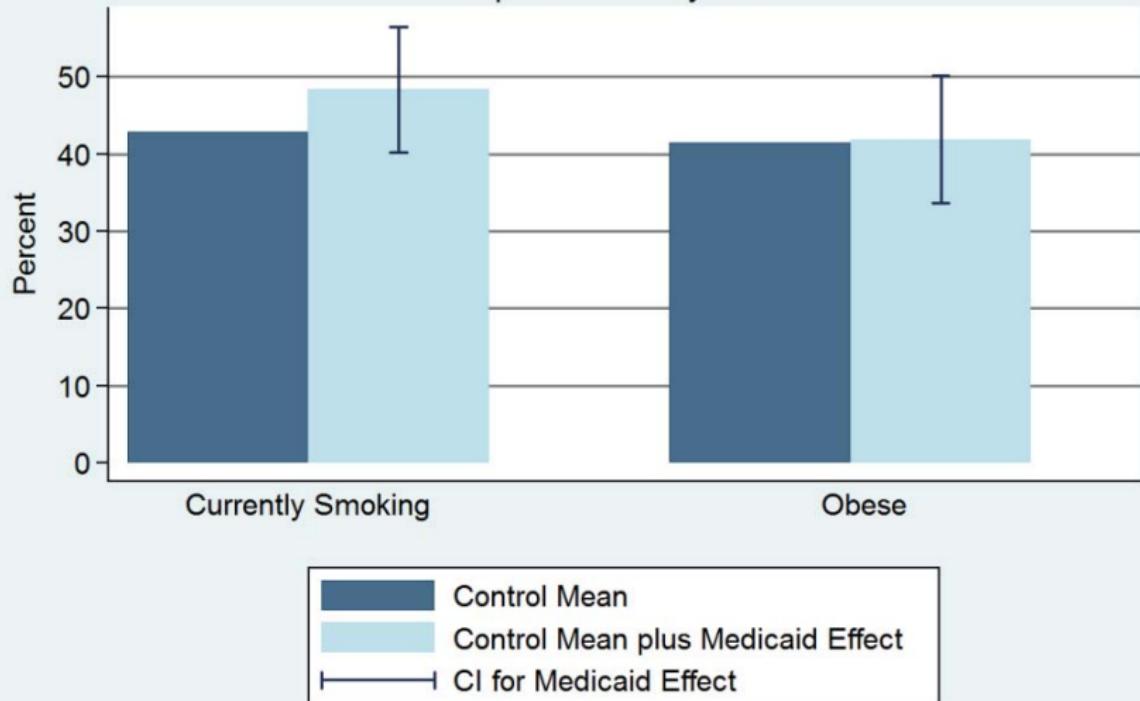
Framingham Risk Score gives the 10 year predicted risk of cardiovascular disease.

Results on specific conditions

- Large reductions in depression
 - Increases in diagnoses and medication
 - In-person estimate of –9 percentage points in being depressed
- Glycated hemoglobin
 - Increases in diagnosis and medication
 - No significant effect on HbA1c; wide confidence intervals
- Blood pressure and cholesterol
 - No significant effects on diagnosis or medication
 - No significant effects on outcomes
- Framingham risk score
 - No significant effect (in general or sub-populations)

Smoking and Obesity

Inperson Survey Data



Summary

- One to two years after expanded access to Medicaid:
 - Increases in health care use and associated costs
 - Increases in compliance with recommended preventive care
 - Improvements in quality and access
 - Reductions in financial strain
 - Improvements in self-reported health
 - Improvements in depression
 - No significant change in specific physical measures
- Sense of the relative magnitude of the effects
 - Use and access, financial benefits, general health, depression
 - Physical measures of specific chronic conditions

Extrapolation to Obamacare (ACA) Expansion

- Context quite relevant for health care reform:
 - States can choose to cover a similar population in planned 2014 Medicaid expansions (up to 138% of federal poverty line)
- But important caveats to bear in mind
 - Oregon and Portland vs. US generally
 - Voluntary enrollment vs. mandate
 - Partial vs. general equilibrium effects
 - Short-run (1-2 years) vs. medium or long run
- We will revisit this again later in the difference-in-differences section when discussing Miller, et al. (2019)

Updating Priors based on Study's Findings

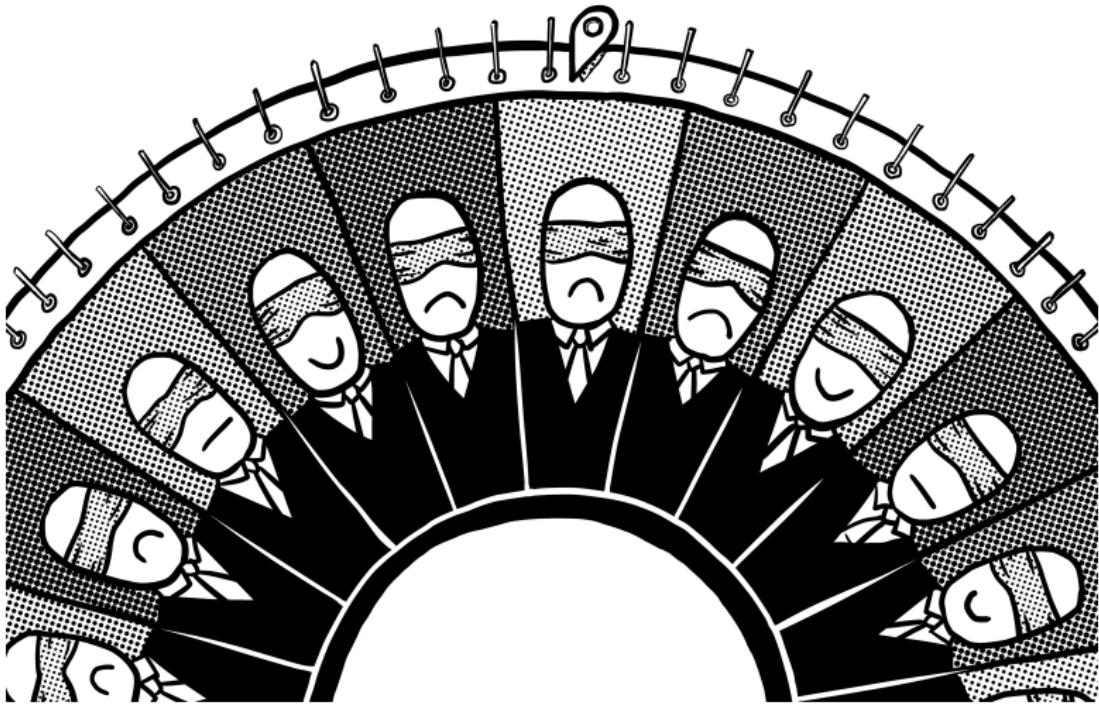
- "Medicaid is worthless or worse than no insurance'"
 - Studies found increases in utilization and perceived access and quality
 - Reductions in financial strain, improvement in self-reported health
 - Improvement in depression
 - Can reject large declines in several physical measures
- "Health insurance expansion saves money"
 - In short run, studies showed increases in utilization and cost and no change in ED use
 - Increases in preventive care, improvements in self-reported health, improvements in depression

Conclusion

- Effects of expanding Medicaid likely to be manifold
 - Hard to establish with observational data and often misleading
- Expanding Medicaid generates both costs and benefits
 - Increased spending
 - Measurably improves *some* aspects of health but not others
 - Important caveats about generalizability
 - Weighing them depends on policy priorities
- Further research on alternative policies needed
 - Many steps in pathway between insurance and outcome
 - Role for innovation in insurance coverage
 - Complements to health care (e.g., social determinants)

Judge fixed effects designs

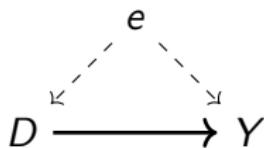
- Imagine the following:
 - ① A person moves through a pipeline and hits a critical point where treatment occurs as a result of some decision-maker
 - ② There are many different decision-makers and you're assigned randomly to one of them
 - ③ Each decision-maker differs in terms of their *leniency* in assigning the treatment
- Very popular in criminal justice bc of how often judges are randomly assigned to defendants (Kling 2006; Mueller-Smith 2015; Dobbie, et al. 2018) or even children to foster care case workers (Doyle 2007; Doyle 2008)



Juvenile incarceration

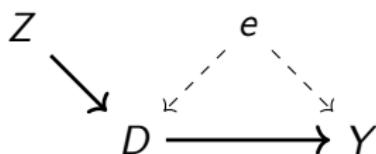
- Aizer and Doyle (2015) were interested in the causal effect of juvenile imprisonment on future crime and human capital accumulation
- Extremely important policy question given the US has the world's highest incarceration rate and prison population of any country in the world by a significant margin (500 prisoners per 100,000, over 2 million adults imprisoned, 4.8 million under supervision)
- High rates of incarceration extend to juveniles: in 2010, the stock of juvenile detainees stood at 70,792, a rate of 2.3 per 1,000 aged 10-19.
- Including supervision, US has a juvenile corrections rate 5x higher than the next highest country, South Africa

Confounding



- We are interested in the causal effect of juvenile incarceration (D) on life outcomes, like adult crime and high school completion
- But youth *choose* to commit crimes, and that choice may be due to unobserved criminogenic factors like poverty or underlying criminal propensities which are themselves causing those future outcomes

Leniency as an instrument



- Aizer and Doyle (2015) propose an instrument - the propensity to convict by the judge the youth is randomly assigned
- If judge assignment is random, and the various assumptions hold, then the IV strategy identifies the local average treatment effect of juvenile incarceration on life outcomes

The Main Idea

- “Plausibly exogenous” variation in juvenile detention stemming from the random assignment of cases to judges who vary in their sentencing
- Consider two juveniles randomly assigned to two different judges with different incarceration tendencies (Scott and Bob)
- Random assignment ensures that differences in incarceration between Scott and Bob are due to the judge, not themselves, because remember, they’re identical

Data

- 35,000 juveniles administrative records over 10 years who came before a juvenile court in Chicago (Juvenile Court of Cook County Delinquency Database)
- Data were linked to public school data for Chicago (Chicago Public Schools) and adult incarceration data for Illinois (Illinois Dept. of Corrections Adult Admissions and Exits)
- They wanted to know the effect of juvenile incarceration on high school completion (2nd data needed) and adult crime (3rd data needed) using randomized judge assignment (1st data needed)
- They need personal identifying information in each data set to make this link (i.e., name, DOB, address)

Preview of findings

- Juvenile incarceration decreased high school graduation by 13 percentage points (vs. 39pp in OLS)
- Increased adult incarceration by 23 percentage points (vs. 41pp in OLS)
- Marginal cases are high risk of adult incarceration and low risk of high school completion as a result of juvenile custody
- Unlikely to ever return to school after incarcerated, but when they do return, they are more likely to be classified as special ed students, and more likely to be classified for special ed services due to behavioral/emotional disorders (as opposed to cognitive disability)

“Plausibly” exogenous

- Very common in these studies for the assignment to some decision-maker to be *arbitrary* but not clearly random (i.e., not random no. generator)
- In this case, juveniles charged with a crime are assigned to a calendar corresponding to their neighborhood and calendars have 1-2 judges who preside over them
- 1/5 of hearings are presided over by judges who cover the calendar when the main judge can't, known as swing judges
- Judge assignment is a function of the sequence with which cases happen to enter into the system and judge availability that is set in advance
- No scope for which judge you see first; conversations with court administrators confirm its random

Structural equation

$$Y_i = \beta_0 + \beta_1 JI_i + \beta_2 X_i + \varepsilon_i$$

where X_i is controls and ε_i is an error term. In this, juvenile incarceration is likely correlated with the error term.

This is the “long” causal model. But note, from the prior DAG, we cannot control for e because it is unobserved. But it is confounding the estimation of juvenile incarceration’s effect on outcomes.

Incarceration Propensity as an Instrument

- The instrument is based on the randomized judge equalling the propensity to incarcerate from the randomly assigned judge
- “Leave-one-out mean”

$$Z_{j(i)} = \left(\frac{1}{n_{j(i)} - 1} \right) \left(\sum_{k \neq i}^{n_{j(i)} - 1} \tilde{J}_{l_k} \right)$$

- The $n_{j(i)}$ terms is the total number of cases seen by judge k , and \tilde{J}_{l_k} is equal to 1 if the juvenile was incarcerated during their first case
- Thus the instrument is the judge's incarceration among first cases based on all their other cases
- It's basically a judge fixed effect given the likelihood two judges have precisely the same propensity is small

Information about the instrument

- There are 62 judges in the data, and the average number of initial cases per judge is 607
- Substantial variation in the data - raw measure ranges from 4% to 21%
- Residualized measure based on controls still has substantial variation from 6% to 18%
- Variation comes from two sources: variation among the regular (nonswing) judges (80% of cases) and variation from the swing judges (20% of cases)

Distribution of IV

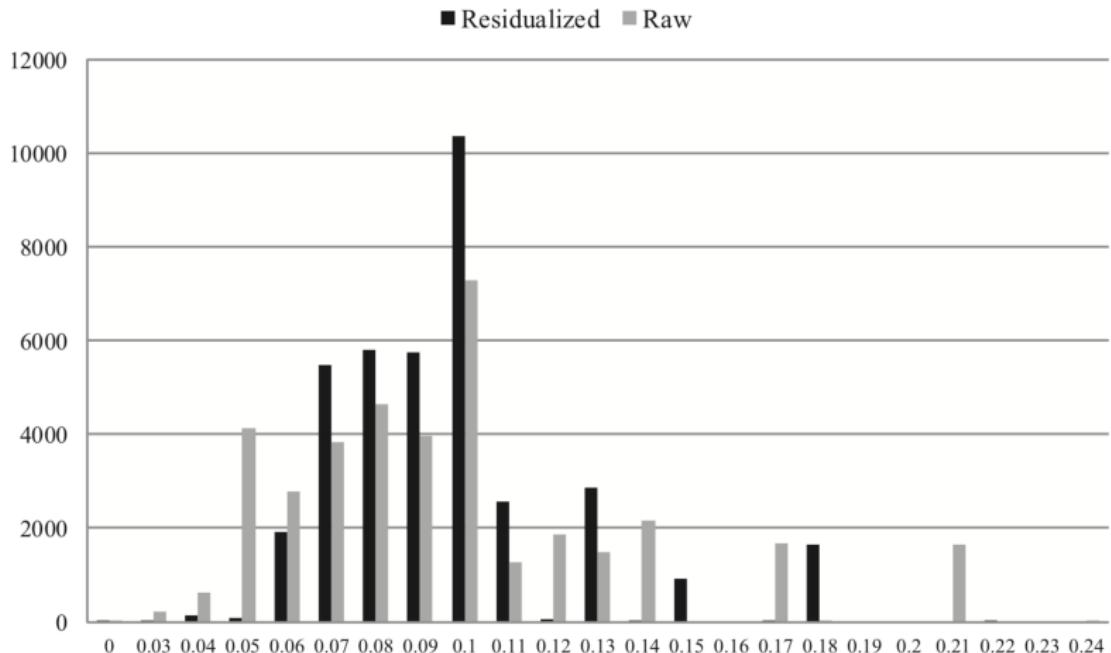


FIGURE I
Distribution of Z: Judge Incarceration Rate

Balance test

TABLE II
INSTRUMENT VERSUS JUVENILE CHARACTERISTICS

	Z distribution			Middle vs.	Top vs.
	Bottom tercile	Middle tercile	Top tercile	bottom p-value	bottom p-value
Z: first judge's leave-out mean incarceration rate in first cases	0.062	0.094	0.147	(.000)	(.000)
Juvenile characteristics					
Male	0.827	0.830	0.833	(.561)	(.311)
African American	0.724	0.737	0.742	(.096)	(.249)
Hispanic	0.189	0.176	0.172	(.061)	(.272)
White	0.078	0.079	0.078	(.833)	(.957)
Other race/ethnicity	0.009	0.008	0.007	(.352)	(.345)
Special education	0.241	0.237	0.252	(.549)	(.130)
U.S. census tract poverty rate	0.264	0.265	0.265	(.572)	(.696)
Age at offense	14.8	14.8	14.8	(.437)	(.434)
P(Juvenile incarceration X)	0.219	0.221	0.220	(.251)	(.516)
Observations	37,692				

First stage

TABLE III
FIRST STAGE

	(1)	(2)	(3)
Dependent variable: juvenile incarcerations		OLS	
First judge's leave-out mean incarceration rate among first cases	1.103 (0.102)	1.082 (0.095)	1.060 (0.097)
Demographic controls	No	Yes	Yes
Court controls	No	No	Yes
Observations	37,692		
Mean of dependent variable	0.227		

High school completion

TABLE IV
JUVENILE INCARCERATION AND HIGH SCHOOL GRADUATION

	Dependent variable: graduated high school						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample			Juvenile court sample			
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	-0.389 (0.0066)	-0.292 (0.0065)	-0.391 (0.0055)	-0.088 (0.0043)	-0.073 (0.0041)	-0.108 (0.044)	-0.125 (0.043)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440,797	440,797	420,033	37,692			
Mean of dependent variable	0.428	0.428	0.433	0.099			

Adult crime

TABLE V
JUVENILE INCARCERATION AND ADULT CRIME

	Dependent variable: entered adult prison by age 25						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample				Juvenile court sample		
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	0.407 (0.0082)	0.350 (0.0064)	0.219 (0.013)	0.200 (0.0072)	0.155 (0.0073)	0.260 (0.073)	0.234 (0.076)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440797	440797	420033	37692			
Mean of dependent variable	0.067	0.067	0.057	0.327			

Crime type

TABLE VI
JUVENILE INCARCERATION AND ADULT CRIME TYPE

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: entered adult prison by age 25 for crime type					
	Homicide			Violent		
	OLS	OLS	2SLS	OLS	OLS	2SLS
Juvenile incarceration	0.051 (0.0031)	0.021 (0.0030)	0.035 (0.030)	0.138 (0.0046)	0.061 (0.0050)	0.149 (0.041)
Sample	Full CPS	Juvenile court	Juvenile court	Full CPS	Juvenile court	Juvenile court
Mean of dep. var.: JI = 0	0.008	0.043	0.043	0.024	0.121	0.121
Observations	440,797	37,692	37,692	440,797	37,692	37,692
	Property			Drug		
Juvenile incarceration	0.079 (0.0040)	0.047 (0.0038)	0.142 (0.044)	0.183 (0.011)	0.078 (0.0068)	0.097 (0.052)
Sample	Full CPS	Juvenile Court	Juvenile Court	Full CPS	Juvenile Court	Juvenile Court
Mean of dep. var.	0.013	0.060	0.060	0.034	0.176	0.176
Observations	440,797	37,692	37,692	440,797	37,692	37,692

High school transfers

TABLE VIII
INTERMEDIATE SCHOOLING OUTCOMES: HIGH SCHOOL TRANSFERS

Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)
	Ever present in CPS school at least 1 year after Initial hearing	Transferred to another CPS high school in years after hearing	Ultimate transfer: adult correctional facility	OLS	2SLS	OLS
Juvenile incarceration	-0.025 (0.0063)	-0.215 (0.069)	0.055 (0.010)	-0.115 (0.243)	0.127 (0.006)	0.243 (0.060)
Mean of dependent variable	0.666		0.242		0.175	
Observations	37,692		18,195		37,692	

Developing emotional problems

TABLE IX
INTERMEDIATE SCHOOLING OUTCOMES: SPECIAL EDUCATION STATUS

Dependent variable:	Special education type observed in years after initial hearing						
	(1)	(2)	(3)	(4)	(5)	(6)	
	Any Special Education			Emotional/ behavioral disorder		Learning disability	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	
Juvenile incarceration	-0.024 (0.004)	-0.003 (0.037)	0.027 (0.003)	0.133 (0.043)	-0.040 (0.004)	-0.097 (0.039)	
Mean of dependent variable	0.193		0.082		0.085		
Observations	29,794						

Concluding remarks

- Sad, but important, paper - the marginal kid shouldn't have been incarcerated
- More generally, leniency designs are very powerful and very common if you know how to look for them
- Bottleneck, influential decision-makers, discretion - these are the three elements of the design

Comments on judge fixed effects

- Leave-one-out average propensity of the decision-maker, or some residualized instrument, is very common
- More often you'll see jackknife IV (JIVE) which drops observations while running regressions to improve finite sample bias
- The biggest threats aren't exclusion probably (though sometimes), but monotonicity
- Might judges be harsh in some situations (violent crimes) but lenient in others (female defendants, first time offenders)

Tests for violations

- New paper by Frandsen, Lefgren and Leslie (2019) proposes a test
- They show that the identifying assumptions imply a conditional expectation of the outcome of interest given the judge assignment is a continuous function of the judge propensity
- They propose a two-part test that generalizes the Sargan-Hansen over identification test and assesses whether treatment effects across judge propensities are possible
- Software available on Emily Leslie's website

Multi-dimensional instrument

- Peter Hull in a cautionary note notes that while combining judge fixed effects into a single propensity is numerically equivalent, it's still a series of dummies
- Therefore it's very important to keep in mind the lessons we learned from weak instruments – the more weak instruments you have when a parameter is overidentified, the larger the bias
- It's ongoing at the moment to think about ways to improve instrument selection, but not settled
- I encourage you to read Peter's note on his website and begin thinking about this yourself

Discussion questions

- When working on a judge fixed effects project, write down an IV DAG
- Whereas monotonicity cannot be visualized to my knowledge on a DAG, exclusion can – so what does an exclusion violation mean in this context?
- Use logic and conversations with those administering the program to answer the following – what does monotonicity mean in this context and how might it be violated?

Empirical exercise

- Let's estimate the effect of cash bail on defendant outcomes using 2SLS and JIVE
- Excellent paper by Megan Stevenson
- -bail.do- and -bail.r- in dropbox and github

Hidden curriculum	Intuition
Foundational causality stuff	Two stage least squares
Regression discontinuity designs	Weak instruments
Instrumental variables	Practical IV Tips
Two-way fixed effects estimator	Heterogeneity and the LATE
Difference-in-differences	Sub IV: Lottery designs
Comparative case studies	Sub IV: Judge fixed effects
Matching and weighting	Sub IV: Bartik
Concluding remarks	Sub IV: Fuzzy design

Bartik instruments

- Bartik instruments are very common in trade, immigration, and labor
- They may be a bit challenging to defend
- But new work by Goldsmith-Pinkham, et al. (2020) and Borusyak, et al. (2019) provide insight into how these work

Bartik instruments

- Perloff (1957) showed industry shares could be used to predict income levels
- Freeman (1980) also used the change in industry composition as an instrument for labor demand
- Timothy Bartik's (1991) study of regional labor markets in which he used the instrument
- Due to Bartik (1991) detailed exposition of the logic of how the national growth shares created variation in labor market demand gets it named after him
- Also called shift-share instruments

Shift-share

"Obvious candidates for instruments are variables shifting MSA labor demand. In this book, only one type of demand shifter is used to form instrumental variables: the share effect from a shift-share analysis of each metropolitan area and year-to-year employment change. A shift-share analysis decomposes MSA growth into three components: a national growth component, which calculates what growth would have occurred if all industries in the MSA had grown at the all-industry national average; a share component, which calculates what extra growth would have occurred if each industry in the MSA had grown at that industry's national average; and a shift component, which calculates the extra growth that occurs because industries grow at different rates locally than they do nationally."

So what is it?

The idea behind a Bartik instrument is to measure the change in a region's labor demand due to changes in the national demand for different industries' products

$$Y_{I,t} = \alpha + \delta I_{I,t} + \rho X_{I,t} + \varepsilon_{I,t}$$

where $Y_{I,t}$ is log wages in location I (e.g., Detroit) in time period t (e.g., 2000) among native workers, $I_{I,t}$ are immigration flows in region I at time period t and $X_{I,t}$ are controls that include region and time fixed effects among other things.

Endogeneity

$$Y_{I,t} = \alpha + \delta I_{I,t} + \rho X_{I,t} + \varepsilon_{I,t}$$

- The parameter δ as elsewhere is some average treatment effect of the immigration flows' effect on native wages.
- The problem is that it is almost certainly the case that immigration flows are highly correlated with the disturbance term such as the time varying characteristics of location I (e.g., changing amenities)

Shares

- Bartik instruments are created by interacting initial “shares” of geographic regions, prior to the immigration flow you’re studying, with national growth rates
- The deviations of a region’s growth from the US national are explained by deviations from national averages
- These deviations are due to shares because the national growth is the same otherwise

Shift-Share Instrument

$$B_{l,t} = \sum_{k=1}^K z_{l,k,t^0} m_{k,t}$$

- The first term is the share variable and the second term is the shift variable.
 - **Share:** z_{l,k,t^0} are the “initial” t^0 share of immigrants from source country k (e.g., Mexico) in location l (e.g., Detroit)
 - **Shift:** $m_{k,t}$ is the change in immigration from country k (e.g., Mexico) into the US as a whole.
- B is the predicted flow of immigrants into destination l (e.g., Detroit) and is a weighted average of the national inflow rates from each country
- Weights depend on the initial distribution of immigrants.

Estimator

- 2SLS: Regress endogenous $I_{I,t}$ (immigration) onto the controls and our Bartik instrument (B)
- Using the fitted values from that regression, \hat{B} , regress $Y_{I,t}$ onto $\hat{I}_{I,t}$
- Under what assumptions does this recover the impact of immigration flows onto log wages?

Two views

- Goldsmith-Pinkham, et al. (2020): the share view
- Borusyak, et al. (2019): the shift view

Exogenous share view

"The Bartik instrument is 'equivalent' to using local industry shares as instruments, and so the exogeneity condition should be interpreted in terms of the shares."

- GMM results showing that the shifts affect the strength of the first stage, but the shares provide the exogenous variation
- If you are exploiting differential exposure to a common shock, then probably the variation comes from shares not shifts
- Strict exogeneity conditional on observables, which means you have to argue why initial shares exogenous to error

Many shifting values

- Bartik is practically challenging to interpret because of the large number of shifting values; US has over 400 industries for instance
- Multiply these over many time periods and the exclusion restriction becomes hard to defend
- Goldsmith-Pinkham, et al. (2020) provide several suggestions for evaluating the central identifying assumption:
over-identification tests for instance (but this fails with heterogenous treatment effects)

Rotemberg weights

- Decomposition of the Bartik estimator into a weighted combination of estimates where each share is an instrument
- Weights are called “Rotemberg weights” which sum to one
- Higher valued weights indicate that those instruments are responsible for more of the identifying variation in the design itself
- Now you can focus on the top weights rather than, for instance, 400 shares
- If high weighted areas pass some basic specific tests, then confidence may be more defensible
- <https://github.com/paulgp/shift-share>

Exogenous shift view

"Ultimately, the plausibility of the exogenous shocks framework, as with the alternative framework of Goldsmith-Pinkham, et al. (2020) based on exogenous shares, depends on the shift-share IV application. We encourage practitioners to use shift-share instruments based on an a priori argument supporting the plausibility of either one of these approaches; various diagnostics and tests of the framework that is most suitable for the setting may then be applied. While Borusyak, et al. (2019) develops such procedures for the "shocks" view, Goldsmith-Pinkham, et al. (2020) provide different tools for the "shares" view."

Exogenous shift view

- Exogenous shares are sufficient, but not necessary, for identifying causal effects from Bartik designs
- Temporal shocks may provide exogenous sources of variation
- Borusyak, et al. (2019) show that exogenous independent shocks to many industries allow Bartik designs to identify causal effects regardless of whether the shrars are exogenous
- Shocks must be uncorrelated with the bias of the shares

Hidden curriculum	Intuition
Foundational causality stuff	Two stage least squares
Regression discontinuity designs	Weak instruments
Instrumental variables	Practical IV Tips
Two-way fixed effects estimator	Heterogeneity and the LATE
Difference-in-differences	Sub IV: Lottery designs
Comparative case studies	Sub IV: Judge fixed effects
Matching and weighting	Sub IV: Bartik
Concluding remarks	Sub IV: Fuzzy design

Fuzzy RDD, IV and ITT

- Fuzzy RDD is an IV estimator, and requires those assumptions
- You may be more comfortable with presenting the intent-to-treat (ITT) parameter which is just the reduced form regression of Y on Z , therefore
- Many papers will not present an IV-style parameter, but rather a blizzard of ITT parameters, out of a “fear” that the exclusion restrictions may not hold

Probability of treatment jumps at discontinuity

Probabilistic treatment assignment (i.e. “fuzzy RDD”)

The probability of receiving treatment changes discontinuously at the cutoff, c_0 , but need not go from 0 to 1

$$\lim_{X_i \rightarrow c_0} Pr(D_i = 1 | X_i = c_0) \neq \lim_{c_0 \leftarrow X_i} Pr(D_i = 1 | X_i = c_0)$$

Examples: Incentives to participate in some program may change discontinuously at the cutoff but are not powerful enough to move everyone from non participation to participation.

Deterministic (sharp) vs. probabilistic (fuzzy)

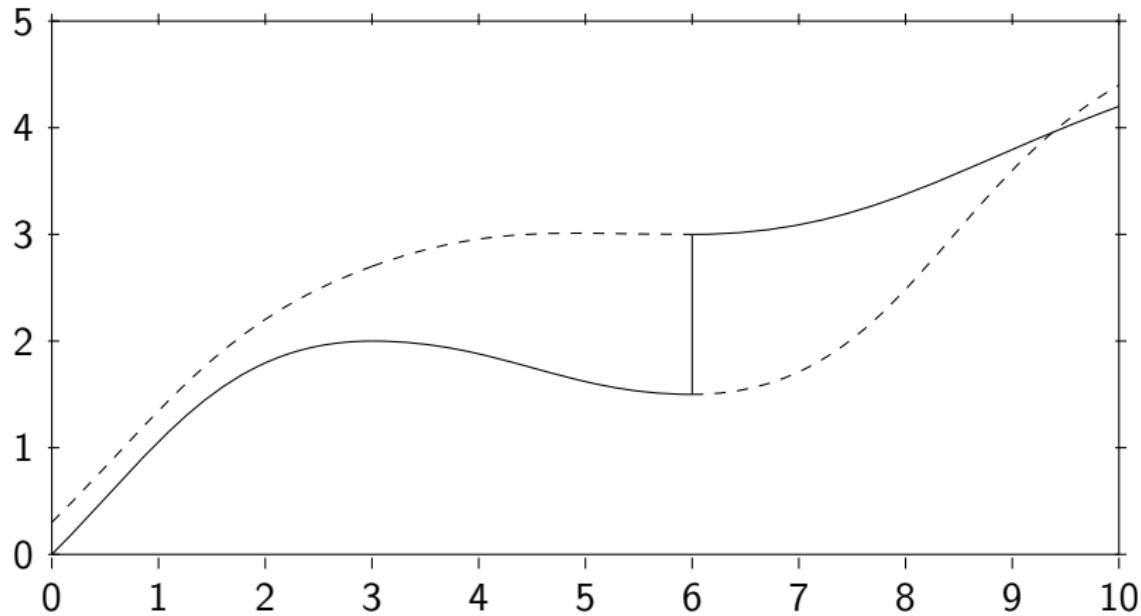
- In the sharp RDD, D_i was *determined* by $X_i \geq c_0$
- In the fuzzy RDD, the *conditional probability* of treatment *jumps* at c_0 .
- The relationship between the conditional probability of treatment and X_i can be written as:

$$P[D_i = 1 | X_i] = g_0(X_i) + [g_1(X_i) - g_0(X_i)]Z_i$$

where $Z_i = 1$ if $(X_i \geq c_0)$ and 0 otherwise.

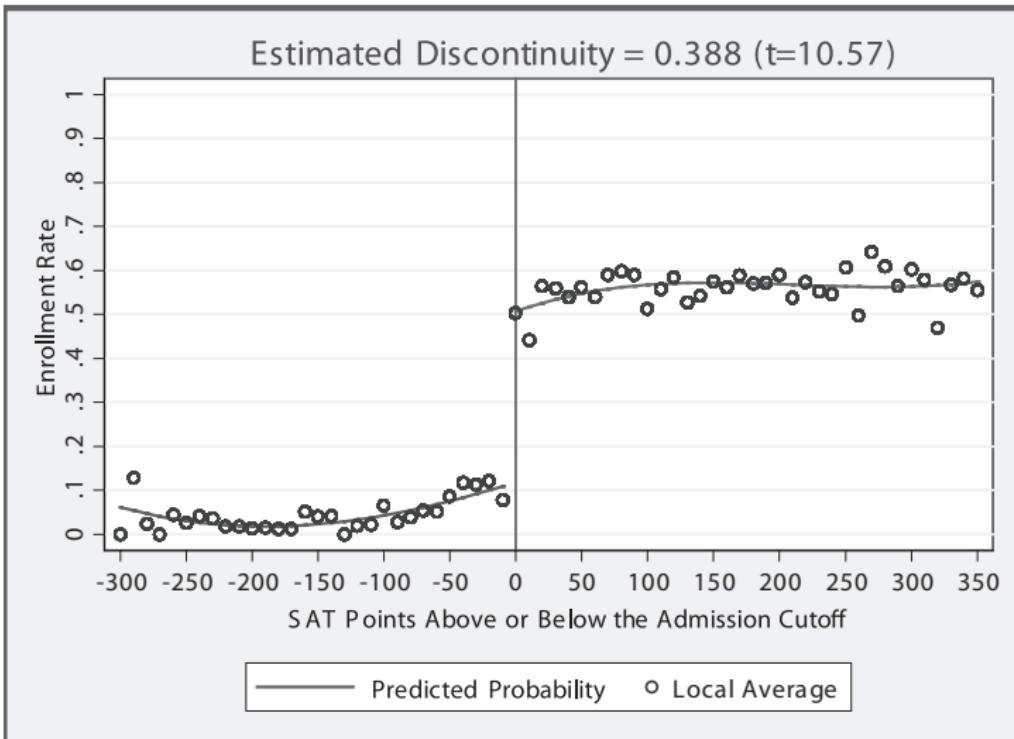
Visualization of identification strategy (i.e. smoothness)

- $E[Y^0|X]$ and $E[Y^1|X]$ for $D = 0, 1$ are the dashed/solid continuous functions
- $E[Y|X]$ is the solid which jumps at $X = 6$



Hoekstra flagship school

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Instrumental variables

- As said, fuzzy designs are numerically equivalent and conceptually similar to IV
 - “Reduced form” Numerator: “jump” in the regression of the outcome on the running variable, X .
 - “First stage” Denominator: “jump” in the regression of the treatment indicator on the running variable X .
- Same IV assumptions, caveats about compliers vs. defiers, and statistical tests that we will discuss in next lecture with instrumental variables apply here – e.g., check for weak instruments using F test on instrument in first stage, etc.

Wald estimator

Wald estimator of treatment effect under Fuzzy RDD

Average causal effect of the treatment is the Wald IV parameter

$$\delta_{\text{Fuzzy RDD}} = \frac{\lim_{X \rightarrow c_0} E[Y|X = c_0] - \lim_{c_0 \leftarrow X} E[Y|X = c_0]}{\lim_{X \rightarrow c_0} E[D|X = c_0] - \lim_{c_0 \leftarrow X} E[D|X = c_0]}$$

RDD's Relationship to IV

- Center X it's equal to zero at c_0 and define $Z = \mathbf{1}(X \geq 0)$
- The coefficient on Z in a regression like

```
. reg Y Z X X2 X3
```

is the reduced form discontinuity, and

```
. reg D Z X X2 X3
```

is the first stage discontinuity

- Ratio of discontinuities is estimate of $\delta_{\text{Fuzzy RDD}}$
- Simple way to implement is IV

```
. ivregress 2sls Y (D=Z) X X2 X3
```

First stage relationship between X and D

- One can use both Z_i as well as the interaction terms as instruments for D_i .
- If one uses only Z_i as IV, then it is a “just identified” model which usually has good finite sample properties.
- In the just identified case, the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \cdots + \gamma_p X_i^p + \pi Z_i + \varepsilon_{1i}$$

where π is the causal effect of Z on the conditional probability of treatment.

- The fuzzy RD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \cdots + \kappa_p X_i^p + \rho \pi Z_i + \varepsilon_{2i}$$

Fuzzy RDD with varying Treatment Effects - Second Stage

- As in the sharp RDD case one can allow the smooth function to be different on both sides of the discontinuity.
- The second stage model with interaction terms would be the same as before:

$$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p + \rho D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \eta_i$$

- Where \tilde{x} are now not only normalized with respect to c_0 but are also fitted values obtained from the first stage regression.

Fuzzy RDD with Varying Treatment Effects - First Stages

- Again one can use both Z_i as well as the interaction terms as instruments for D_i
- Only using Z the estimated first stages would be:

$$D_i = \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p + \pi Z_i + \gamma_1^* \tilde{X}_i Z_i + \gamma_2^* \tilde{X}_i^2 Z_i + \cdots + \gamma_p^* Z_i + \varepsilon_{1i}$$

- We would also construct analogous first stages for $\tilde{X}_i D_i$, $\tilde{X}_i^2 D_i, \dots, \tilde{X}_i^p D_i$.

Limitations of the LATE

- Fuzzy RDD has assumptions of all standard IV framework (exclusion, independence, nonzero first stage, and monotonicity)
- As with other binary IVs, the fuzzy RDD is estimating LATE: the local average treatment effect for the group of *compliers*
- In RDD, the compliers are those whose treatment status changed as we moved the value of x_i from just to the left of c_0 to just to the right of c_0
- Means we can use Medicare age cutoff to estimate the effect of public insurance on mortality (LATE) and still not know the effect of public insurance on mortality (ATE)

Two-way fixed effects

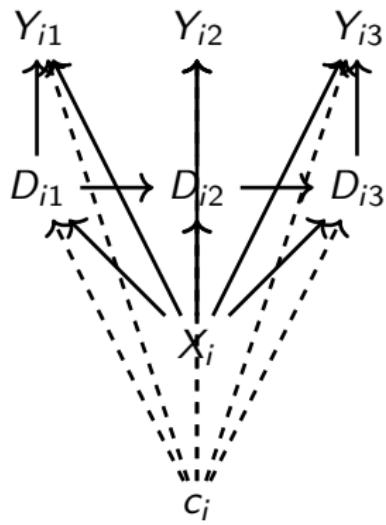
- When working with panel data, the so-called “two-way fixed effects” (TWFE) estimator is the workhorse estimator
- It’s easy to run, a version of OLS, and many people are just interested in mean effects anyway
- It’s the most common model for estimating treatment effects in a difference-in-differences, and so for all these reasons, we need to spend some time understanding what it is

Panel Data

- Panel data: we observe the same units (individuals, firms, countries, schools, etc.) over several time periods
- Often our outcome variable depends on unobserved factors which are also correlated with our explanatory variable of interest
- If these omitted variables are constant over time, we can use panel data estimators to consistently estimate the effect of our explanatory variable

What I will cover

- I will cover pooled OLS and twoway fixed effects
- But I won't be covering random effects, Arrelano and Bond and any number of important panel estimators because the purpose here is to present the modal regression model used in difference-in-differences



Sorry - drawing the DAG for a simple panel model is somewhat messy!

When to use this

- Traditionally, this was used for estimating constant treatment effects with unobserved time-invariant heterogeneity – recall the c_i was constant across all time periods
- It's a linear model, so you'll be estimating conditional mean treatment effects – if you want the median, you can't use this
- Once you enter into a world with dynamic treatment effects and differential timing, this loses all value

Problems that fixed effects cannot solve

- Reverse causality: Becker predicted police reduce crime, but when you regress crime onto police, it's usually positive
 - $\hat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
 - $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - implies ε_{it} uncorrelated with past, current and future regressors
- Time-varying unobserved heterogeneity
 - It's the time-varying unobservables you have to worry about in fixed effects
 - Can include time-varying controls, but as always, don't condition on a collider

Formal panel notation

- Let y and $x \equiv (x_1, x_2, \dots, x_k)$ be observable random variables and c be an unobservable random variable
- We are interested in the partial effects of variable x_j in the population regression function

$$E[y|x_1, x_2, \dots, x_k, c]$$

Formal panel notation cont.

- We observe a sample of $i = 1, 2, \dots, N$ cross-sectional units for $t = 1, 2, \dots, T$ time periods (a balanced panel)
 - For each unit i , we denote the observable variables for all time periods as $\{(y_{it}, x_{it}) : t = 1, 2, \dots, T\}$
 - $x_{it} \equiv (x_{it1}, x_{it2}, \dots, x_{itk})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{y_i, x_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
 - $y_i \equiv (y_{i1}, y_{i2}, \dots, y_{iT})'$ and $x_i \equiv (x_{i1}, x_{i2}, \dots, x_{iT})$
 - Consider asymptotic properties with T fixed and $N \rightarrow \infty$

Formal panel notation

Single unit:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \quad X_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & X_{i,1,j} & \dots & X_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,t,1} & X_{i,t,2} & X_{i,t,j} & \dots & X_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,T,1} & X_{i,T,2} & X_{i,T,j} & \dots & X_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{pmatrix}_{NT \times 1} \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix}_{NT \times K}$$

Unobserved heterogeneity

- For a randomly drawn cross-sectional unit i , the model is given by

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- y_{it} : log wages i in year t
- x_{it} : $1 \times K$ vector of variable events for person i in year t , such as education, marriage, etc. plus an intercept
- β : $K \times 1$ vector of marginal effects of events
- c_i : sum of all time-invariant inputs known to people i (but unobserved for the researcher), e.g., ability, beauty, grit, etc., often called unobserved heterogeneity or fixed effect
- ε_{it} : time-varying unobserved factors, such as a recession, unknown to the farmer at the time the decision on the events x_{it} are made, sometimes called idiosyncratic error

Pooled OLS

- When we ignore the panel structure and regress y_{it} on x_{it} we get

$$y_{it} = x_{it}\beta + v_{it}; \quad t = 1, 2, \dots, T$$

with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- What happens when we regress y_{it} on x_{it} if x is correlated with c_i ?
- Then x ends up correlated with v , the composite error term.
- Somehow we need to eliminate this bias, but how?

Pooled OLS

- Main assumption to obtain consistent estimates for β is:
 - $E[v_{it}|x_{i1}, x_{i2}, \dots, x_{iT}] = E[v_{it}|x_{it}] = 0$ for $t = 1, 2, \dots, T$
 - x_{it} are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current and future regressors
 - But: education x_{it} likely depends on grit and ability c_i and so we have omitted variable bias and $\hat{\beta}$ is not consistent
 - No correlation between x_{it} and v_{it} implies no correlation between unobserved effect c_i and x_{it} for all t
 - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)
 - Additional problem: v_{it} are serially correlated for same i since c_i is present in each t and thus pooled OLS standard errors are invalid

Pooled OLS

- Always ask: is there a time-constant unobserved variable (c_i) that is correlated with the regressors?
- If yes, then pooled OLS is problematic
- This is how we motivate a fixed effects model: because we believe unobserved heterogeneity is the main driving force making the treatment variable endogenous

Fixed effect regression

- Our unobserved effects model is:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \dots, T$$

- If we have data on multiple time periods, we can think of c_i as **fixed effects** to be estimated
- OLS estimation with fixed effects yields

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it}b - m_i)^2$$

this amounts to including N individual dummies in regression of y_{it} on x_{it}

Derivation: fixed effects regression

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it} b - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^N \sum_{t=1}^T x'_{it} (y_{it} - x_{it} \hat{\beta} - \hat{c}_i) = 0$$

and

$$\sum_{t=1}^T (y_{it} - x_{it} \hat{\beta} - \hat{c}_i) = 0$$

for $i = 1, \dots, N$.

Derivation: fixed effects regression

Therefore, for $i = 1, \dots, N$,

$$\hat{c}_i = \frac{1}{T} \sum_{t=1}^T (y_{it} - x_{it}\hat{\beta}) = \bar{y}_i - \bar{x}_i\hat{\beta},$$

where

$$\bar{x}_i \equiv \frac{1}{T} \sum_{t=1}^T x_{it}; \bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^T y_{it}$$

Plug this result into the first FOC to obtain:

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i) \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(y_{it} - \bar{y}) \right)$$

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{x}_{it} \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{y}_{it} \right)$$

with time-demeaned variables $\ddot{x}_{it} \equiv x_{it} - \bar{x}$, $\ddot{y}_{it} \equiv y_{it} - \bar{y}$

Fixed effects regression

Running a regression with the time-demeaned variables
 $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{x}_{it} \equiv x_{it} - \bar{x}$ is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $Cov[x_{it}, c_i] \neq 0$ because time-demeaning eliminates the unobserved effects

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}$$

$$\bar{y}_i = \bar{x}_i\beta + c_i + \bar{\varepsilon}_i$$

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x})\beta + (c_i - \bar{c}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}$$

Fixed effects regression: main results

- Identification assumptions:

- ① $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - regressors are strictly exogenous conditional on the unobserved effect
 - allows x_{it} to be arbitrarily related to c_i
- ② $\text{rank}\left(\sum_{t=1}^T E[\ddot{x}'_{it} \ddot{x}_{it}]\right) = K$
 - regressors vary over time for at least some i and not collinear

- Fixed effects estimator

- ① Demean and regress \ddot{y}_{it} on \ddot{x}_{it} (need to correct degrees of freedom)
- ② Regress y_{it} on x_{it} and unit dummies (dummy variable regression)
- ③ Regress y_{it} on x_{it} with canned fixed effects routine
 - Stata: `xtreg y x, fe i(PanelID)`

FE main results

- Properties (under assumptions 1-2):
 - $\widehat{\beta}_{FE}$ is consistent: $\plim_{N \rightarrow \infty} \widehat{\beta}_{FE,N} = \beta$
 - $\widehat{\beta}_{FE}$ is unbiased conditional on \mathbf{X}

Fixed effects regression: main issues

- Inference:
 - Standard errors have to be “clustered” by panel unit (e.g., farm) to allow correlation in the ε_{it} ’s for the same i .
 - Yields valid inference as long as number of clusters is reasonably large
- Typically we care about β , but unit fixed effects c_i could be of interest
 - \hat{c}_i from dummy variable regression is unbiased but not consistent for c_i (based on fixed T and $N \rightarrow \infty$)

Application: SASP

- From 2008-2009, I fielded a survey of Internet sex workers (685 respondents, 5% response rate)
- I asked two types of questions: static provider-specific information (e.g., age, weight) and dynamic session information over last 5 sessions
- Let's look at the panel aspect of this analysis together

Risk premium equation

$$\begin{aligned}Y_{is} &= \beta_i X_i + \delta D_{is} + \gamma_{is} Z_{is} + u_i + \varepsilon_{is} \\ \ddot{Y}_{is} &= \gamma_{is} \ddot{Z}_{is} + \ddot{\eta}_{is}\end{aligned}$$

where Y is log price, D is unprotected sex with a client in a session, X are client and session characteristics, Z is unobserved heterogeneity, and u_i is both unobserved and correlated with Z_{is} .

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	POLS	FE	Demeaned OLS
Unprotected sex with client of any kind	0.013 (0.028)	0.051* (0.028)	0.051* (0.026)
Ln(Length)	-0.308*** (0.028)	-0.435*** (0.024)	-0.435*** (0.019)
Client was a Regular	-0.047* (0.028)	-0.037** (0.019)	-0.037** (0.017)
Age of Client	-0.001 (0.009)	0.002 (0.007)	0.002 (0.006)
Age of Client Squared	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Client Attractiveness (Scale of 1 to 10)	0.020*** (0.007)	0.006 (0.006)	0.006 (0.005)
Second Provider Involved	0.055 (0.067)	0.113* (0.060)	0.113* (0.048)
Asian Client	-0.014 (0.049)	-0.010 (0.034)	-0.010 (0.030)
Black Client	0.092 (0.073)	0.027 (0.042)	0.027 (0.037)
Hispanic Client	0.052 (0.080)	-0.062 (0.052)	-0.062 (0.045)
Other Ethnicity Client	0.156** (0.068)	0.142*** (0.049)	0.142*** (0.045)
Met Client in Hotel	0.133*** (0.029)	0.052* (0.027)	0.052* (0.024)
Gave Client a Massage	-0.134*** (0.029)	-0.001 (0.028)	-0.001 (0.024)
Age of provider	0.003 (0.012)	0.000 (.)	0.000 (.)
Age of provider squared	-0.000 (0.000)	0.000 (.)	0.000 (.)

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	POLS	FE	Demeaned OLS
Body Mass Index	-0.022*** (0.002)	0.000 (.)	0.000 (.)
Hispanic	-0.226*** (0.082)	0.000 (.)	0.000 (.)
Black	0.028 (0.064)	0.000 (.)	0.000 (.)
Other	-0.112 (0.077)	0.000 (.)	0.000 (.)
Asian	0.086 (0.158)	0.000 (.)	0.000 (.)
Imputed Years of Schooling	0.020** (0.010)	0.000 (.)	0.000 (.)
Cohabitating (living with a partner) but unmarried	-0.054 (0.036)	0.000 (.)	0.000 (.)
Currently married and living with your spouse	0.005 (0.043)	0.000 (.)	0.000 (.)
Divorced and not remarried	-0.021 (0.038)	0.000 (.)	0.000 (.)
Married but not currently living with your spouse	-0.056 (0.059)	0.000 (.)	0.000 (.)
N	1,028	1,028	1,028
Mean of dependent variable	5.57	5.57	0.00

Heteroskedastic robust standard errors in parenthesis clustered at the provider level. * p<0.10,

** p<0.05, *** p<0.01

Unit specific time trends often eliminate “results”

Table: Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers with provider specific trends

Depvar:	FE w/provider trends
Unprotected sex with client of any kind	0.004 (0.046)
Ln(Length)	-0.450*** (0.020)
Client was a Regular	-0.071** (0.023)
Age of Client	0.008 (0.005)
Age of Client Squared	-0.000 (0.000)
Client Attractiveness (Scale of 1 to 10)	0.003 (0.003)
Second Provider Involved	0.126* (0.055)
Asian Client	-0.048*** (0.007)
Black Client	0.017 (0.043)
Hispanic Client	-0.015 (0.022)
Other Ethnicity Client	0.135*** (0.031)
Met Client in Hotel	0.073*** (0.019)
Gave Client a Massage	0.022 (0.012)

Concluding remarks

- This is not a review of panel econometrics; for that see Wooldridge and other excellent options
- We reviewed POLS and TWFE because they are commonly used with individual level panel data and difference-in-differences
- Their main value is how they control for unobserved heterogeneity through a simple demeaning
- Now let's discuss difference-in-differences which will at various times use the TWFE model

John Snow

- John Snow was a practicing anesthesiologist in the mid 19th century London
- He was then famous for inventing a machine that would carefully deliver chloroform to patients in homogenous dosage which reduced mortality from anesthesia
- But he is now famous for providing convincing evidence that cholera was a waterborne disease during the 1854 outbreak
- Published two works on cholera – an essay in 1849, and a book in 1855
- Died of a stroke in 1858

Deaths from Cholera, each day in 1854

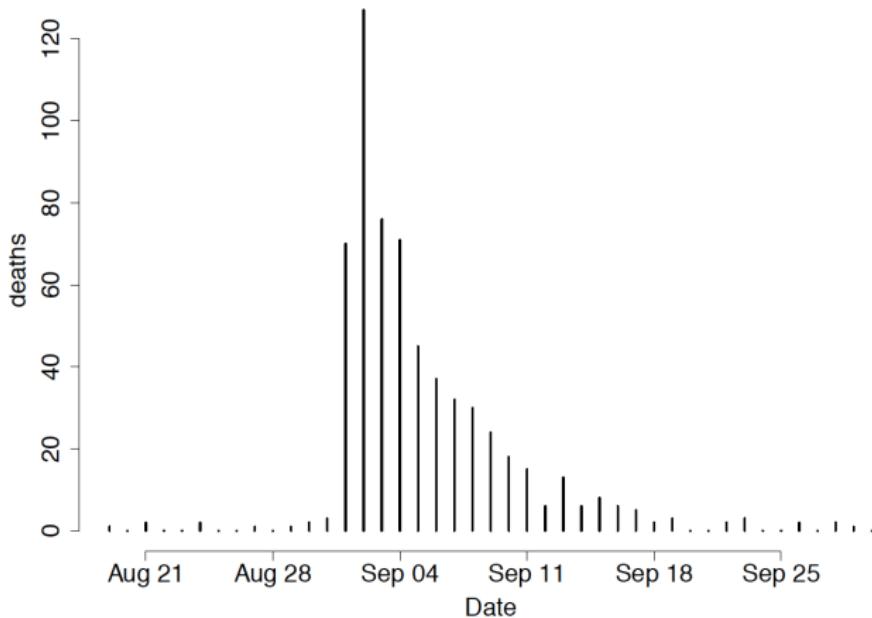


Figure: Daily cholera deaths, London (Coleman 2019)

Cholera background

- Cholera hits London three times in the early to mid 1800s causing large waves of tens of thousands of deaths
- Three London epidemics – 1831-1832, 1848-1849, 1853-1854
- Cholera attacked victims suddenly, with a 50% survival rate, and very painful symptoms included vomiting and acute diarrhea

Miasmis

- 19th century London was a filthy place with waste collecting in cesspools under houses or emptied into open ditches and sewers
- Majority opinion about disease was *miasmis*
- Miasmis hypothesized that disease transmission was caused by vapors and smells; unclear its relevance for person-to-person

Never before seen microorganism

- Microscopes were around but had horrible resolution
- Most human pathogens couldn't be seen
- Johnson (2007) reports Snow did track down a microscope but could only see blurry things moving around
- Isolating these microorganisms wouldn't occur for half a century

Snow's hypothesis

- Snow (as well as a few others like Rev. Henry Whitehead) believe miasms is not relevant for explaining cholera
- Snow hypothesizes that the active agent was a living organism that entered the body, got into the alimentary canal with food or drink, multiplied in the body, and generated some poison that caused the body to expel water
- The organism passed out of the body with these evacuations, entered the water supply and infected new victims
- The process repeated itself, growing rapidly through the common water supply, causing an epidemic

Thought Experiment

- How will he convince anyone that cholera is waterborne and not due to “bad air”?
- Consider the ideal experiment: randomize households by coin flip to receive water from runoff (control) vs. water without runoff (treatment)
- Unethical, impractical and unrealistic
- Even if the randomized experiment is not possible, the thought experiment suggests the observational equivalent

Multiple sources of evidence, not just one

Snow makes his argument with many pieces of evidence that when taken together are very compelling that water, not air, is the cause of the cholera epidemics. These can be categorized as:

- ① Observation
- ② Broad Street Pump
- ③ Grand Experiment

Observation

- Observed progression of the disease for years
- Tracked Patient Zero
- Treatments didn't work: Snow would cover with burlap sacks, which did nothing
- Strange irregular patterns – higher deaths in close proximity to a public pump on Broad Street, fewer deaths at a pub
"cholera extended to nearly all the houses in which the water was thus tainted, and to no others." (Snow 1849)

Broad street outbreak

"The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street [now Lexington St.] joins Broad Street [now Broadwick], there were upwards of five hundred fatal attacks of cholera in ten days." (Snow 1855)

How he argues for the Broad street pump

- Famous map showing unusual mass of cholera deaths near the public Broad street pump
- He was looking for the source, but he was not inductively forming his theory with this map because he already knew the mechanism
- He was assembling evidence that would further refute the explanations of those who advocated an alternative explanation of the outbreak

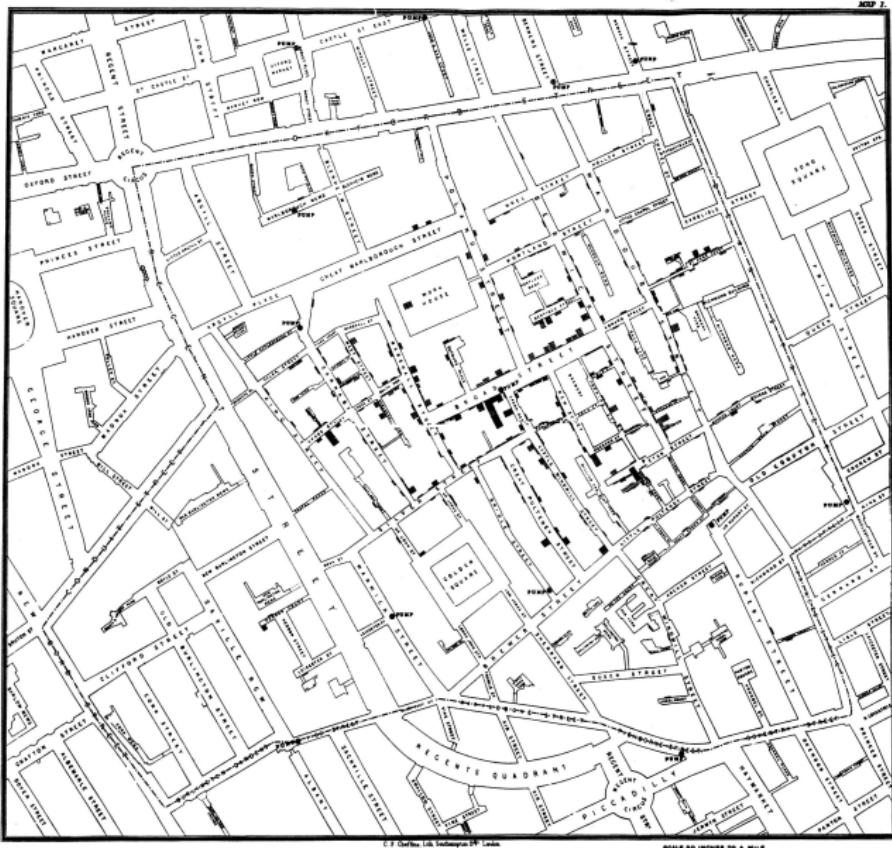


Figure: Cholera deaths laid over a small area of London near Broad Street

Map was important but not enough on its own

"[Snow] could see at a glance that he'd be able to demonstrate that the outbreak was clustered around the pump, yet he knew from experience that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over that part of Soho, something emanating from the gully holes or cesspools – or perhaps even from the pump itself. Snow knew that the case would be made in the exceptions from the norm. Pockets of life where you could expect death, pockets of death where you would expect life." Johnson (2007) p. 140

Two companies fight for customers

- Southwark and Vauxhall Waterworks Company and the Lambeth Water Company competed over some of the regions south of the Thames
- In 16 sub-districts, with a population of 300,000, they competed directly, even supplying customers side-by-side

"In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies." Snow (1855) p 75

Lambeth moves its pipe

- During the 1849 epidemic, both companies drew water from Thames which was polluted with sewage and cholera
- London passes legislation requiring utility companies to move their pipes above the city
- In 1852, the Lambeth Company, a water utility company, changed supply from Hungerford Bridge
- It moved its intake pipe upstream to cleaner water and in response to legislation (SV delayed)
- This created a natural experiment because Southwark and Vauxhall left its intake pipe in place

Meticulous Data Collection

- Two types of data: DD uses aggregate deaths bc of mixing of customers whereas his Broad Street evidence focused on individuals
- Collected detailed information from households with cholera deaths on utility subscription (Lambeth or SV)
- Many residents didn't know their water company – distant landlords paid for it
- He knew Lambeth water was four times saltier, so he'd take a sample and test it using a saline test back at his office

Shoelather and knowledge of institutional details

- Careful balance checks – “the pipes of each Company go down all the streets into nearly all the courts and alleys”
- Concern for sample selection bias –“No fewer than 3000 people of both sexes [of all types affected]”
- Treatment assignment was arbitrary – “a few houses supplied by one Company and a few by the other”

Table XII
Modified Table XII (Snow 1854)

Company name	1849	1854
Southwark and Vauxhall	135	147
Lambeth	85	19

Estimated ATT using DD is 78 fewer deaths per 10,000

Failure to convince

"In spite of what has since been recognized as a classic exercise in data, analysis, and argument, Snow failed to convince the medical profession, the policy-making establishment, or the public." (Coleman 2019)

Final victory

- Another cholera outbreak in 1866, east of London, is when Snow's ideas were gradually and reluctantly accepted by public officials and the scientific community
- 1866 outbreak was confined only to the east of London, which was the last area not yet connected to the newly constructed sewage system which discharged sewage below the Thames
- The rest of London didn't have an outbreak
- This was the final piece of evidence that swayed skeptics and led to a more reasoned assessment of Snow's data and analysis

Merits of Snow's work

- Long commitment to the topic led him to reject unsound hypotheses and form new ones based on observation and experience (shoe leather)
- Expert handling of data analysis, data visualization, and a framing of evidence with a ladder of reasoning

Layered rhetoric of research

"The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself." (Johnson, Ghost Map)

Hidden curriculum
Foundational causality stuff
Regression discontinuity designs
Instrumental variables
Twoway fixed effects estimator
Difference-in-differences
Comparative case studies
Matching and weighting
Concluding remarks

Two group case
Event study
Covariates
Differential timing
Revisiting event studies
Alternative DD estimators
Conclusion

Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

Interrupted time series design

Table: Lambeth, 1849 and 1854

Company	Time	Cholera mortality
Lambeth	1854	$Y = L$
	1849	$Y = L + (T + D)$

Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

Sample averages

$$\hat{\delta}_{kU}^{2 \times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Population expectations

$$\widehat{\delta}_{kU}^{2\times 2} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

Potential outcomes and the switching equation

$$\widehat{\delta}_{kU}^{2x2} = \underbrace{\left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)}_{\text{Switching equation}} + \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\widehat{\delta}_{kU}^{2\times 2} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\left[E[Y_k^0|Post] - E[Y_k^0|Pre] \right] - \left[E[Y_U^0|Post] - E[Y_U^0|Pre] \right]}_{\text{Non-parallel trends bias in } 2\times 2 \text{ case}}$$

Another famous DD study

- Card and Krueger (1994) was a seminal study on the minimum wage both for the result and for the design
- Not the first time we saw DD in the modern period - there's Ashenfelter (1978) and Card (1991) - but got a lot of attention

Competitive vs noncompetitive markets

- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In a competitive input market, increases in the minimum wage would move us up a downward sloping labor demand curve → employment would fall
- Monopsony (imperfect labor markets) suggest the opposite effect whereby raising the minimum wage increases employment

Monopsony's minimum wage predictions

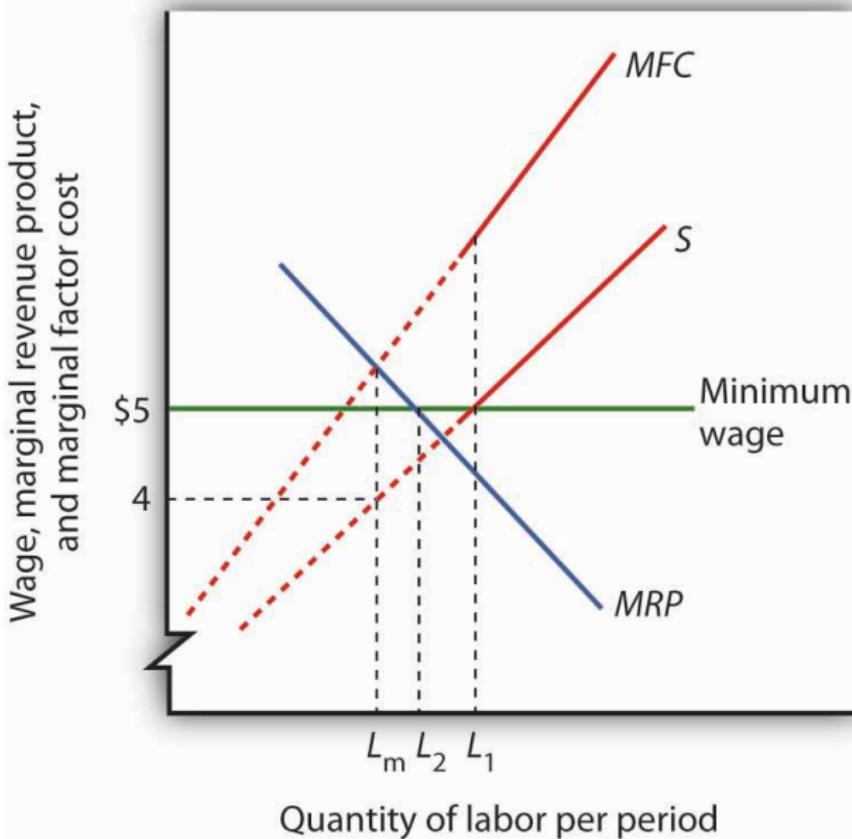
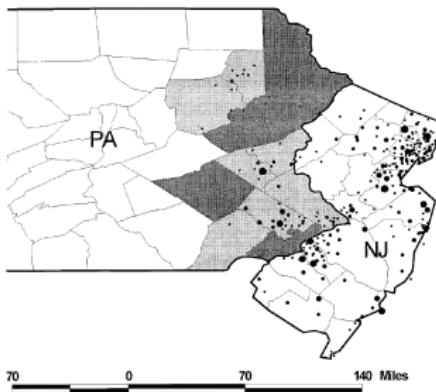


Figure: Monopsony predictions

Card and Krueger (1994)

- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey - shoe leather!

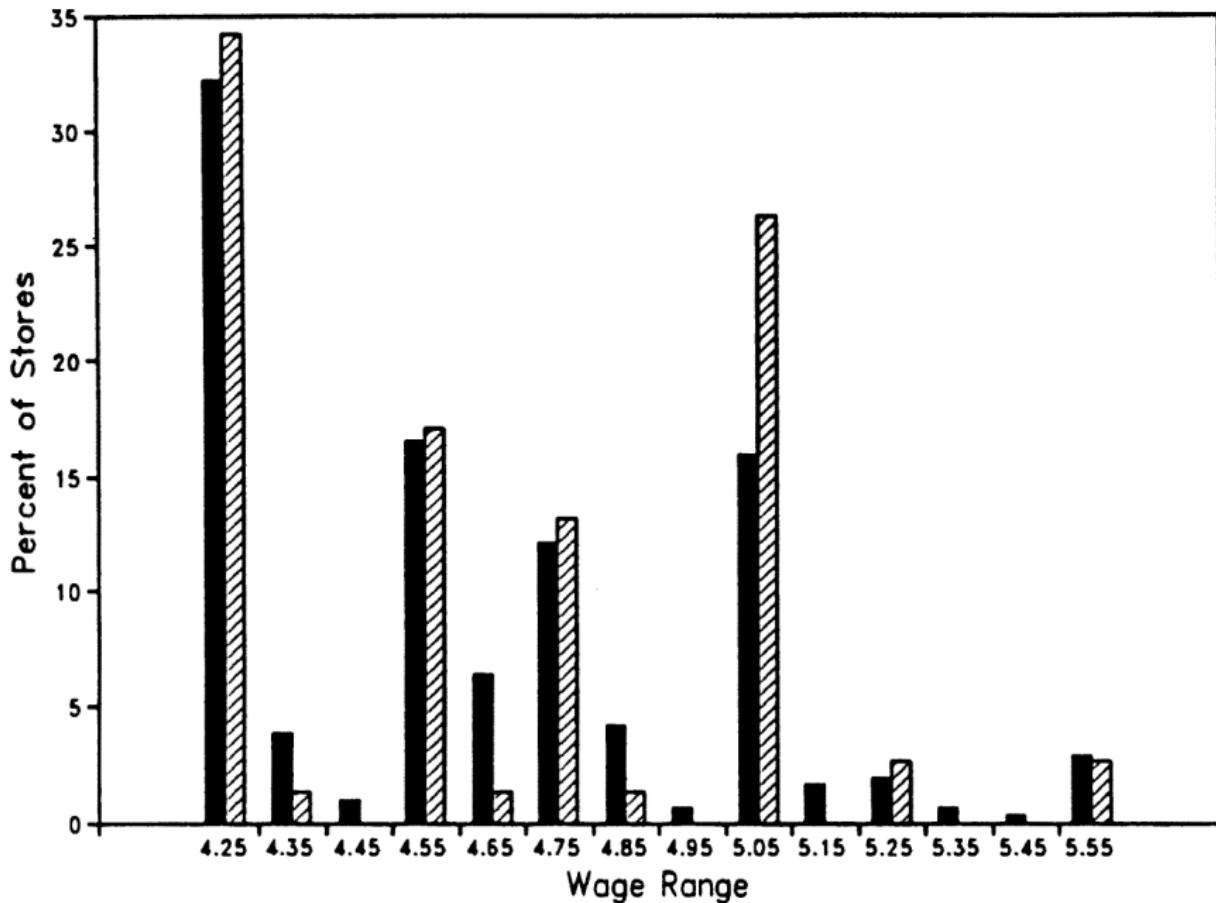
Parallel trends assumption

- Key identifying assumption is the “parallel trends” assumption

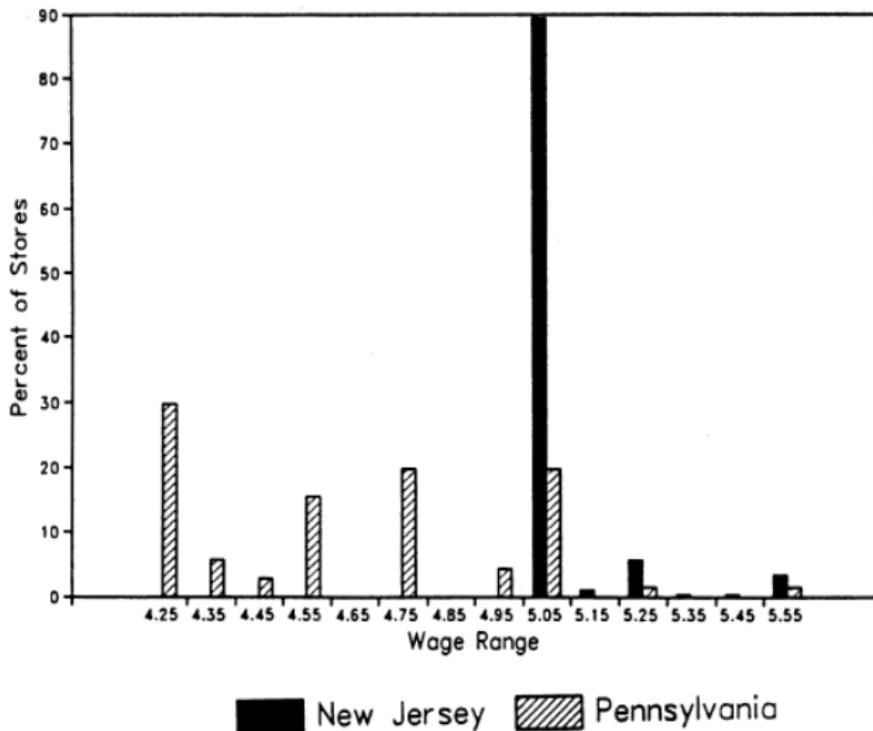
$$\underbrace{[E[Y_{NJ}^0 | Post] - E[Y_{NJ}^0 | Pre]] - [E[Y_{PA}^0 | Post] - E[Y_{PA}^0 | Pre]]}_{\text{Non-parallel trends bias}}$$

- Note the counterfactual - it is *not testable* no matter what someone tells you, bc New Jersey's post period potential employment in a world with a lower minimum wage is unobserved
- Let's look at this a couple of different ways, including a graphic showing the binding minimum wage

February 1992



November 1992



Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment *rose* in NJ relative to PA after the minimum wage change - consistent with monopsony theory

Regression DD

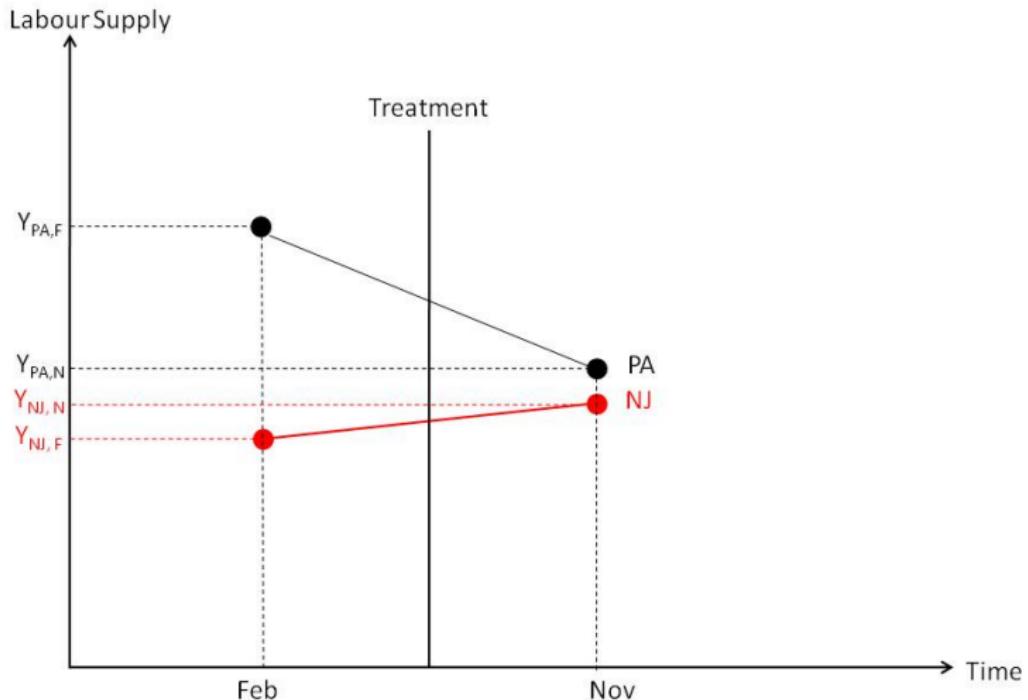
- There are several good reasons to use TWFE
 - It estimates the ATT under parallel trends
 - It's easy to calculate the standard errors
 - It's easy to include multiple periods
 - We can study treatments with different treatment intensity.
(e.g., varying increases in the minimum wage for different states)
- But there are bad reasons, too, which I'll discuss under differential timing and covariates

Regression DD - Card and Krueger

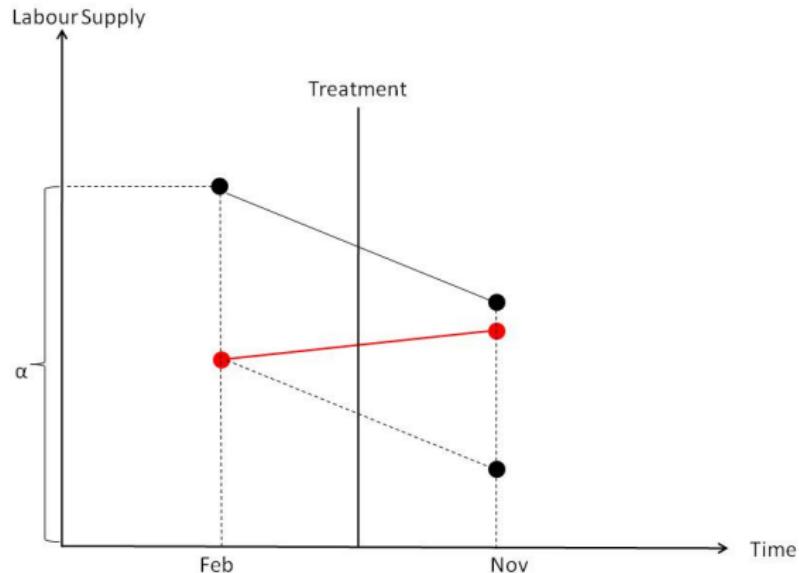
- In the Card and Krueger case, the equivalent regression would be:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

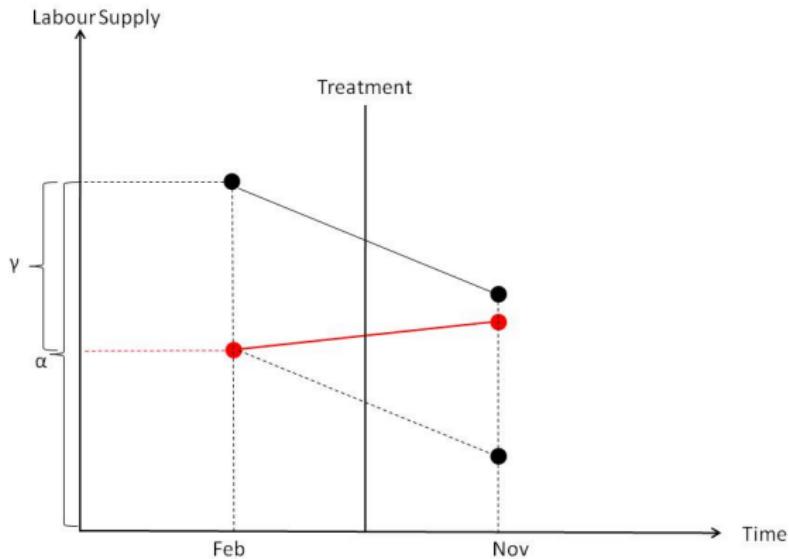
- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$



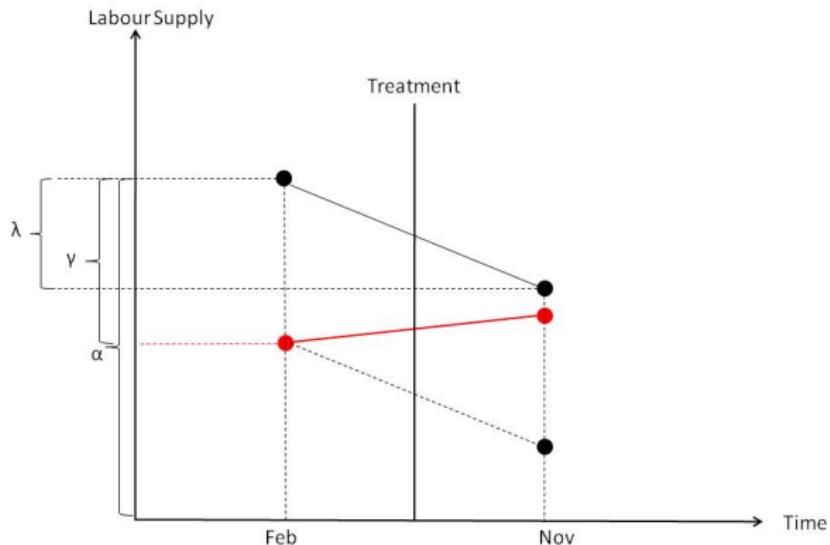
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



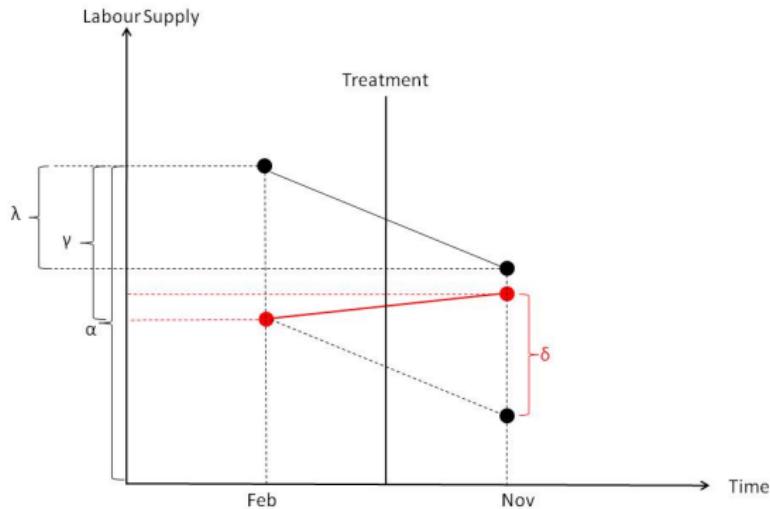
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$



Losing parallel trends

- If parallel trends doesn't hold, then ATT is not identified
- But, regardless of whether ATT is identified, OLS always estimates the same thing
- That's because OLS uses the slope of the control group to estimate the DD parameter, which is only unbiased if that slope is the correct counterfactual trend for the treatment group

Labor Supply

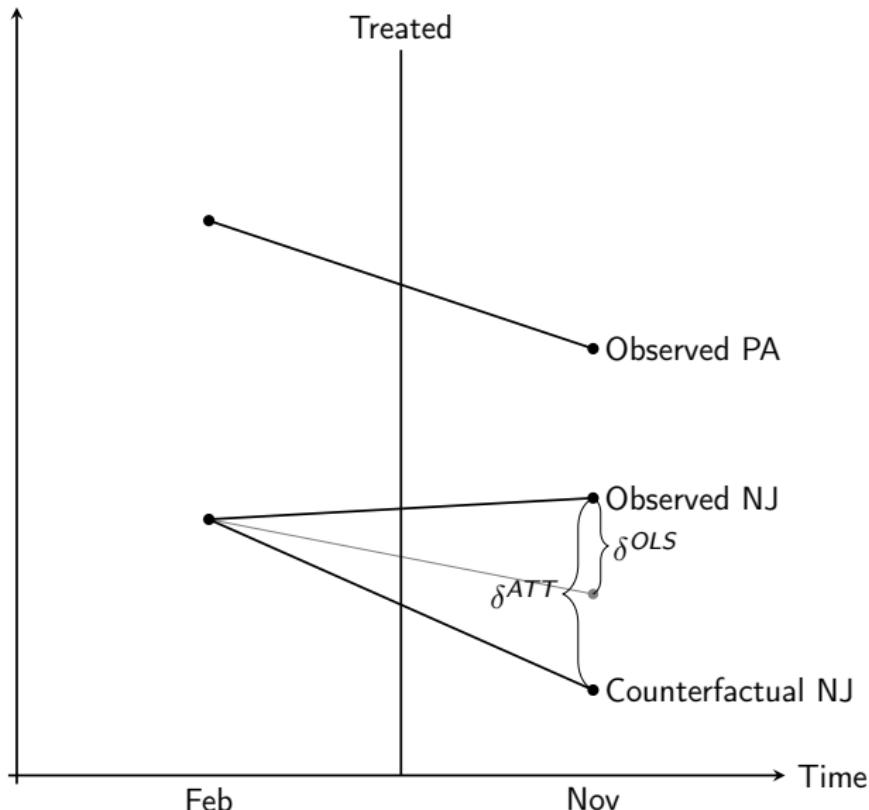


Figure: DD regression diagram without parallel trends

Compositional differences violate parallel trends

- One of the risks of a repeated cross-section is that the composition of the sample may have changed between the pre and post period
- Hong (2011) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napster (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX

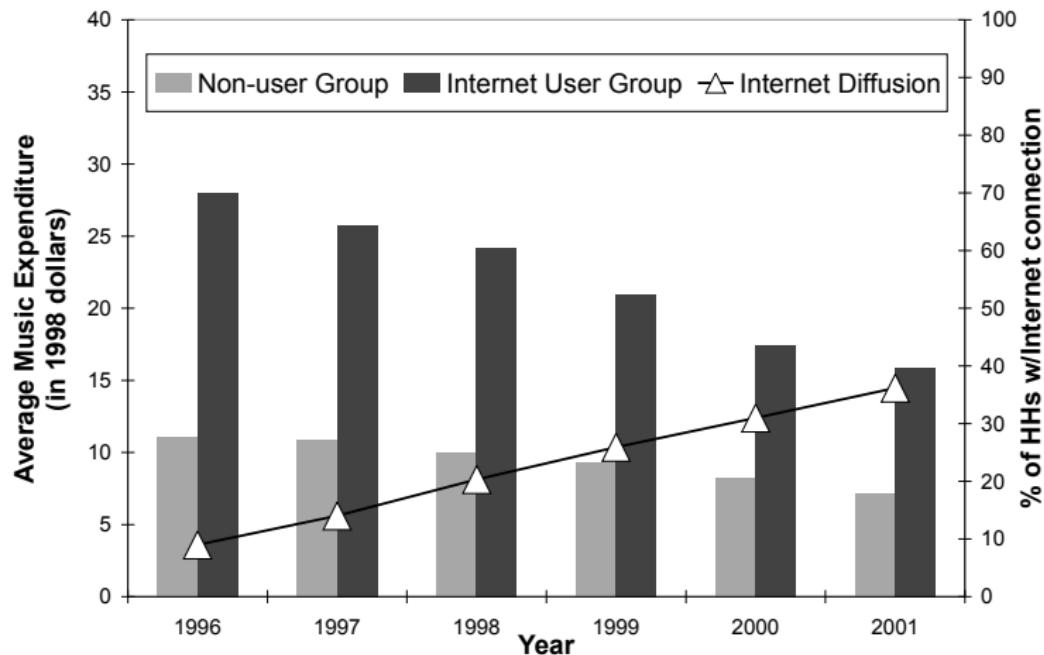


Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

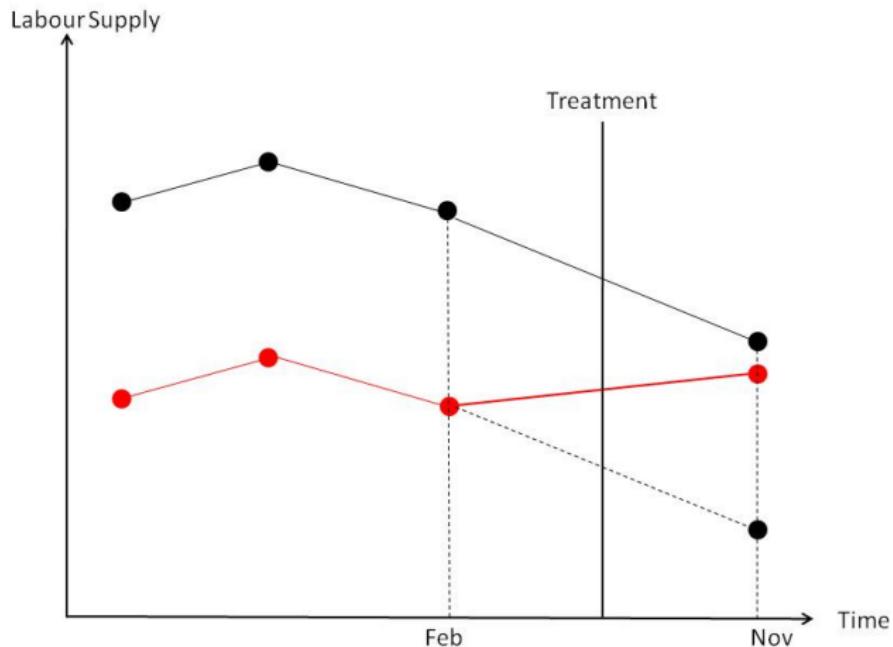
Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	\$52,887	\$30,459	\$51,995	\$28,169	\$49,970	\$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the Internet changes samples (e.g., younger music fans are early adopters)

Pre-trends

- The identifying assumption for all DD designs is parallel trends
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check a hunch for parallel trends using pre-trends
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

Plot the raw data when there's only two groups



Event study regression

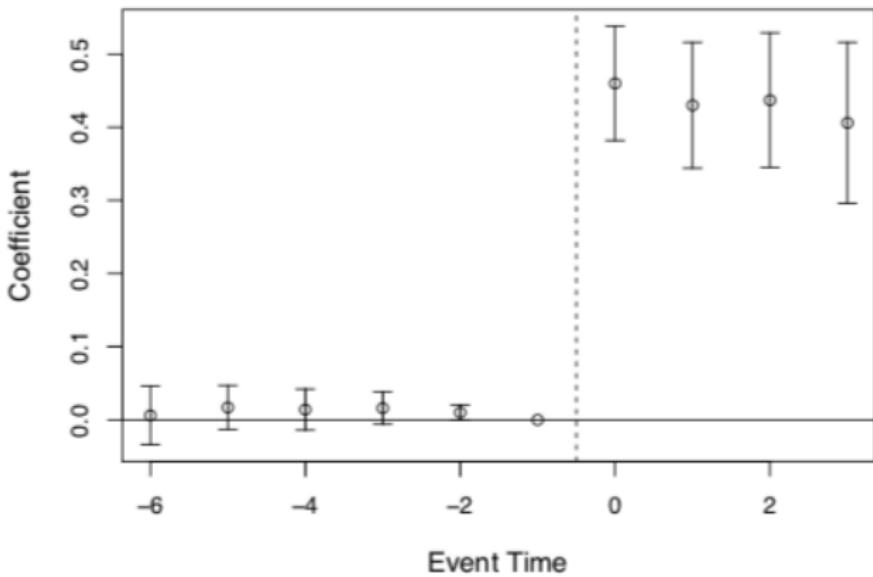
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-2}^{-q} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

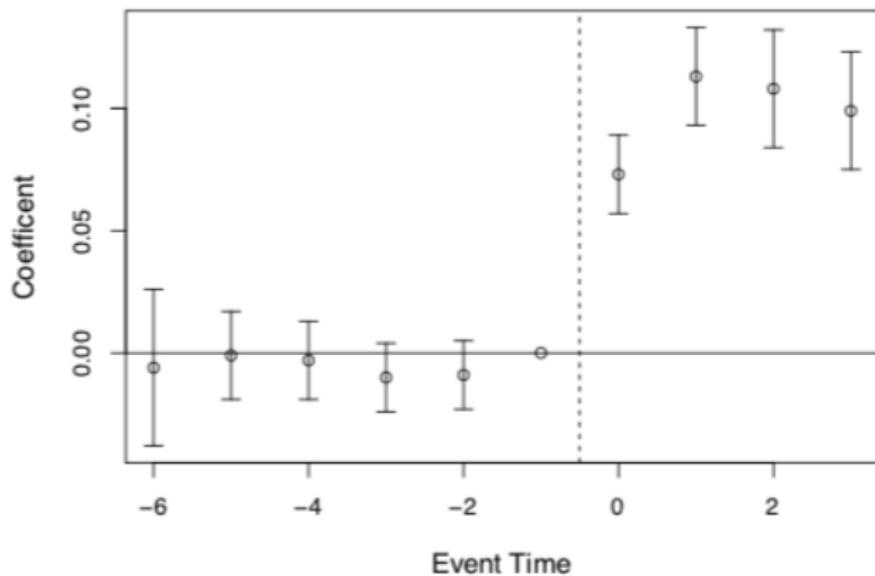
- Treatment occurs in year 0
- Includes q leads or anticipatory effects
- Includes m leads or post treatment effects

Medicaid and Affordable Care Act example

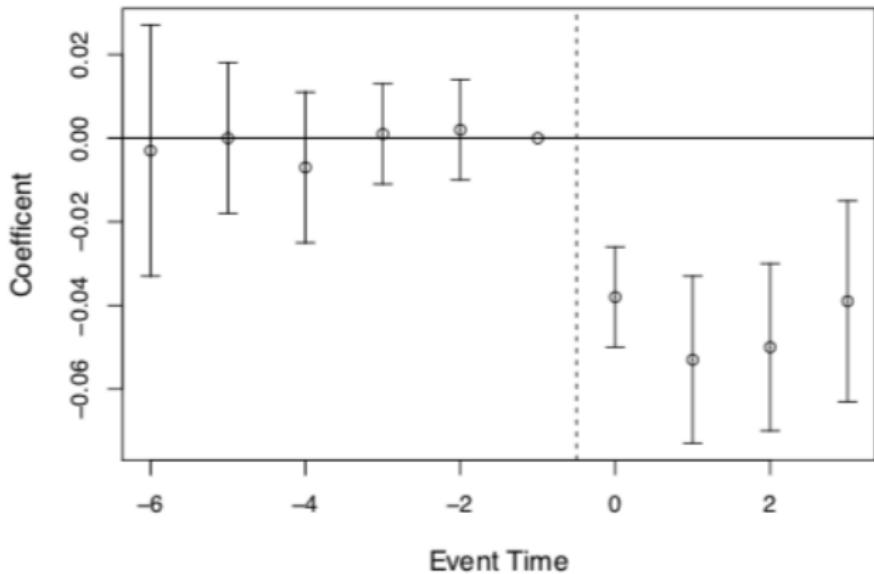
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

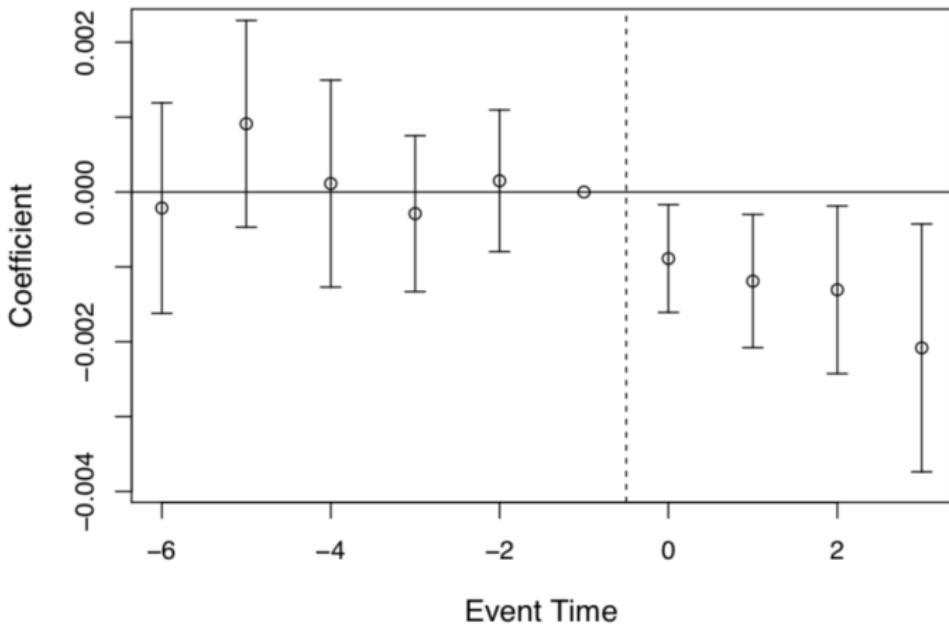


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

Standard errors in DD strategies

- Many papers using DD strategies use data from many years – not just 1 pre and 1 post period
- The variables of interest in many of these setups only vary at a group level (say a state level) and outcome variables are often serially correlated
- As Bertrand, Duflo and Mullainathan (2004) point out, conventional standard errors often severely underestimate the standard deviation of the estimators – standard errors are biased downward (i.e., too small, over reject)

Standard errors in DD – practical solutions

- Bertrand, Duflo and Mullainathan propose the following solutions:
 - ① Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
 - ② Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)
 - ③ Aggregating the data into one pre and one post period.
Literally works if there is only one treatment data. With staggered treatment dates one should adopt the following procedure:
 - Regress Y_{st} onto state FE, year FE and relevant covariates
 - Obtain residuals from the treatment states only and divide them into 2 groups: pre and post treatment
 - Then regress the two groups of residuals onto a post dummy

Note about groups

- Correct treatment of standard errors sometimes makes the number of groups very small: in the Card and Krueger study the number of groups is only 2.

DD Robustness

- Very common for readers and others to request a variety of “robustness checks” from a DD design
- Think of these as along the same lines as the leads and lags we already discussed
 - Event study (already discussed)
 - Falsification test using data for alternative control group
 - Falsification test using alternative “placebo” outcome that should not be affected by the treatment

Within group controls - triple diff

Table: Difference-in-Difference-in-differences

States	Group	Period	Outcomes	D_1	D_2	D_3
NJ	Low wage employment	After	$NJ + T + NJ_t + I_t + D$	$T + NJ_t + I_t + D$	$D + I_t - s_t$	$D - s_t$
		Before	NJ			
	High wage employment	After	$NJ + T + NJ_t + s_t$	$T + NJ_t + s_t$	$I_t - s_t$	$D - I_t$
		Before	NJ			
PA	Low wage employment	After	$PA + T + PA_t + I_t$	$T + PA_t + I_t$	$I_t - s_t$	$D - s_t$
		Before	PA			
	High wage employment	After	$PA + T + PA_t + s_t$	$T + PA_t + s_t$	$D - I_t$	$D - I_t - s_t$
		Before	PA			

DDD Example by Gruber

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES
ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. Treatment Individuals: Married Women, 20–40 Years Old:			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:		−0.062 (0.022)	
B. Control Group: Over 40 and Single Males 20–40:			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:		−0.008: (0.014)	
DDD:		−0.054 (0.026)	

DDD in Regression

$$Y_{ijt} = \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}$$

- The DDD estimate is the difference between the DD of interest and a placebo DD (which is supposed to be zero)
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias
- If the placebo DD is zero, then DD and DDD give the same results but DD is preferable because standard errors are smaller for DD than DDD
- But now you have multiple parallel trends assumption - both the control group trends are good counterfactuals, and within-state placebo trends for within-state treatment unit counterfactual trends

Implementing DDD

- Have to get the structure of the data correct because now you have (1) before and after, (2) treatment and control states, and (3) within state placebo
- I give an example in my Mixtape (p. 278) looking at abortion legalization's effect on longterm risky sexual behavior, including do file
- Let's review first the paper, then work through the exercise itself using data.

The Long-run Effect of Abortion on Sexually Transmitted Infections

Christopher Cornwell, *University of Georgia*, and Scott Cunningham,
Baylor University

Send correspondence to: Scott Cunningham, Department of Economics, Baylor University, One Bear Place #98003, Waco, TX 76798-8003, USA; Tel: 254-710-4753; Fax: 254-710-6142; E-mail: scott_cunningham@baylor.edu

There is a growing literature on the effects of abortion legalization on a range of fertility outcomes. The now-famous paper by Donohue and Levitt [2001. “The Impact of Legalized Abortion on Crime,” 116 *Quarterly Journal of Economics* 379–420], linking abortion to the decline in crime in the 1990s, has shifted the focus to non-fertility outcomes. We focus on STIs, specifically gonorrhea, exploiting the states that legalized abortion prior to *Roe v. Wade* as a quasi-experiment. Using data from the CDC,

Figure: Longrun effects of abortion legalization on Risky Sex

Motivation

- Legalization caused teen childbearing to fall by 12% (Levine 2004)
- Gruber, et al. (1999) showed that the marginal child would have been 60% more likely to live in a single-parent household, 50% more likely to live in poverty, and 45% more likely to be a recipient of public services
- Mechanism was believed to be non-random selection associated with high risk conditions

Emerging influence

- Donohue and Levitt (2001) linked abortion legalization to declining crime in the 1990s, one of several reasons given for his John Bates Clark award
- Freakonomics popularizes the sensational theory
- Other papers followed like Charles and Stephens (2006) who find that children exposed *in utero* to legalization were less likely to use illegal substances

Controversy

- Triple diff by Joyce finds no evidence for it when using an (arbitrary) cutoff of the median abortion rate within early repeal treatment states
- Foote and Goetz (2008) argue the abortion ratio was constructed incorrectly, and report a coding error leaving out state-year fixed effects; construction problem destroys results, state-year fixed effects somewhat attenuates
- Literature stops and theory is ignored

In defense of Steve Levitt

- I want to remind people though: we only know about the coding error bc Levitt posted his do files and gave them to anyone who asked (very easy to “lose do files”)
- Levitt had and has oodles of scientific integrity for his willingness to cooperate; not always the case

"If abortion lowers homicide rates by 20 – 30%, then it is likely to have affected an entire spectrum of outcomes associated with well-being: infant health, child development, schooling, earnings and marital status. Similarly, the policy implications are broader than abortion. Other interventions that affect fertility control and that lead to fewer unwanted births – contraception or sexual abstinence – have huge potential payoffs. In short, a causal relationship between legalized abortion and crime has such significant ramifications for social policy and at the same time is so controversial, that further assessment of the identifying assumptions and their robustness to alternative strategies is warranted." Ted Joyce in his triple diff paper

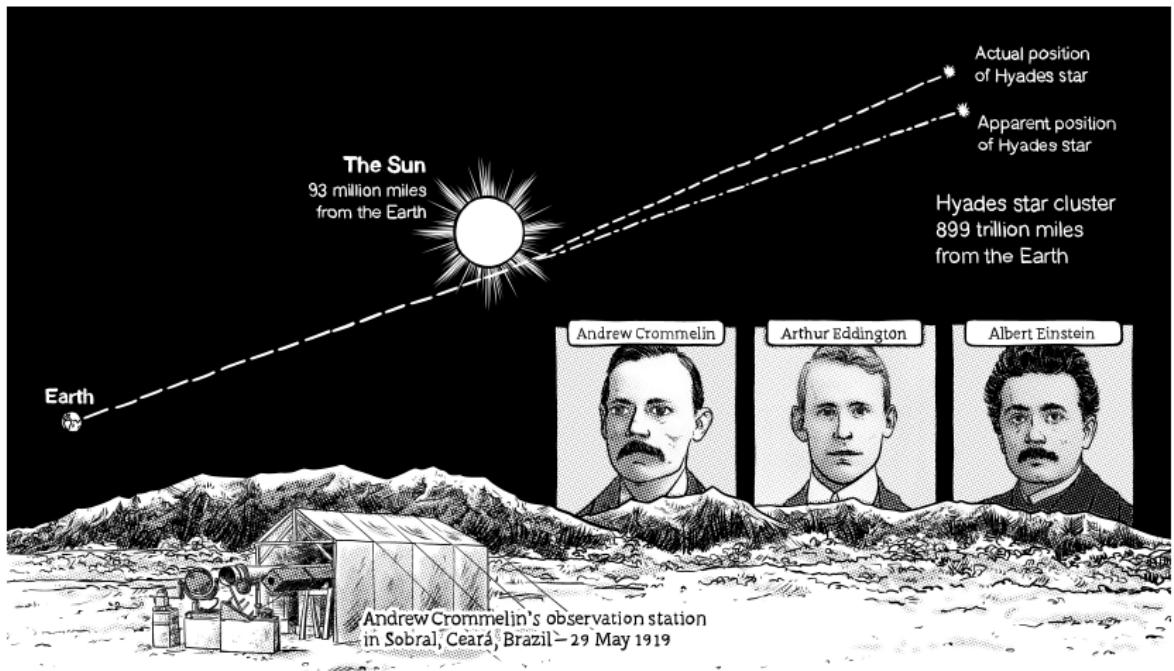


Figure: Light bending around the sun, predicted by Einstein, and confirmed in a natural experiment involving an eclipse. Artwork by Seth Hahne ©.

In defense of falsifiable predictions

- Theories which make falsifiable predictions (comparative statics) are *more* convincing of causal effects than simpler reduced form studies
- Great paper by Coleman on (2019) Snow's rhetoric in his 1849 essay and his 1855 book on cholera – mounts different data to make his argument, some of which is of this nature
- Those predictions are threefold:
 - Where we should find effects
 - Where we should not find effects
 - The kind of effects we should find
- If all three are met, an identified causal effect becomes epistemologically more credible

Falsifiable predictions contained in a diff-in-diff

		CDC Surveillance Data in Calendar Year																	
		1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000		
Age in calendar year		15	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	
16		69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	84	
17		68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	83	
18		67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	82	
19		66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	81	
20		65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	80	
21		64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	79	
22		63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	78	
23		62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	77	
24		61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	76	
25		60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	75	
26		59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	74	
27		58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	73	
28		57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	72	
29		56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	71	
Repeal (1)		0	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5	5	
No Repeal (2)		0	0	0	0	1	2	3	4	5	5	5	5	5	5	5	5	5	
Difference (3)		0	1	2	3	3	3	2	1	0	0	0	0	0	0	0	0	0	

Number of cohorts (age 15-19)
exposed, reforms in 71,74

Figure: Group-time differential exposure predicts a temporary parabolic ATT

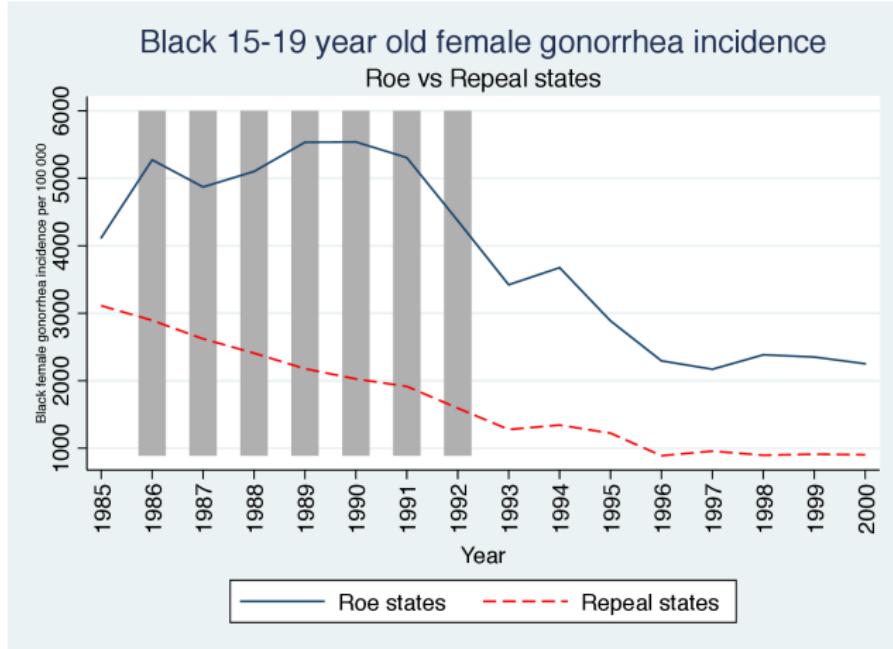


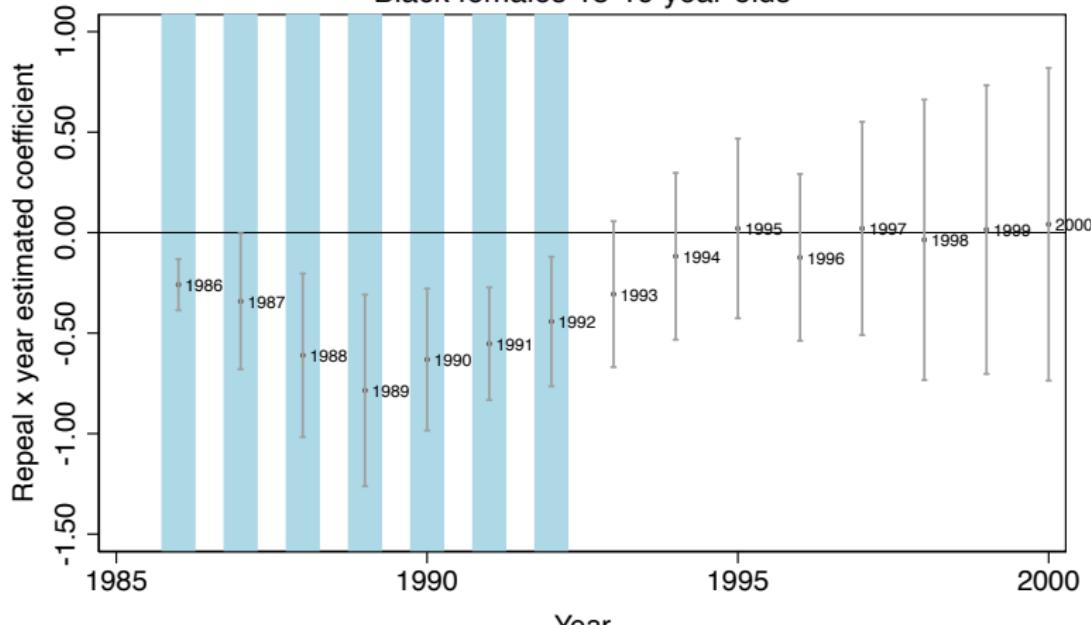
Figure: Raw data for repeal and Roe states.

Estimating equation

$$\begin{aligned} Y_{st} = & \beta_1 Repeals + \beta_2 DT_t + \beta_3 t Repeals \times DT_t + X_{st}\psi + \alpha_s DS_s \\ & + \gamma_1 t + \gamma_2 s \times t + \varepsilon_{st} \end{aligned}$$

Estimated effect of abortion legalization on gonorrhea

Black females 15-19 year-olds



Whisker plots are estimated coefficients of DD estimates

Figure: Differences in black female gonorrhea incidence between repeal and Roe cohorts.

Assuaging doubt

- Maybe spurious - something happened in those years, but what?
- Crack epidemic maybe? But we control for the crack index by Fryer, et al.
- Maybe something else - let's try a within-state control group (the older cohort)

DDD Equation

$$\begin{aligned} Y_{ast} = & \beta_1 Repeal_s + \beta_2 DT_t + \beta_3 DA + \beta_{4t} Repeal_s \cdot DT_t + \\ & + \beta_5 Repeal_s \cdot DA + \beta_{6t} DA \cdot DT_t + \beta_{7t} Repeal_s \cdot DA \cdot DT_t \\ & + X_{st}\xi + \alpha_{1s} DS_s + \alpha_{2s} DS_s \cdot DA + \gamma_1 t + \gamma_{2s} DS_s \cdot t + \gamma_3 DA \cdot t \\ & + \gamma_{4s} DS_s \cdot DA \cdot t + \epsilon_{ast} \end{aligned}$$

One will be dropped, but I want to focus your attention on the number of interactions needed to identify DDD parameters

Stacking Structure

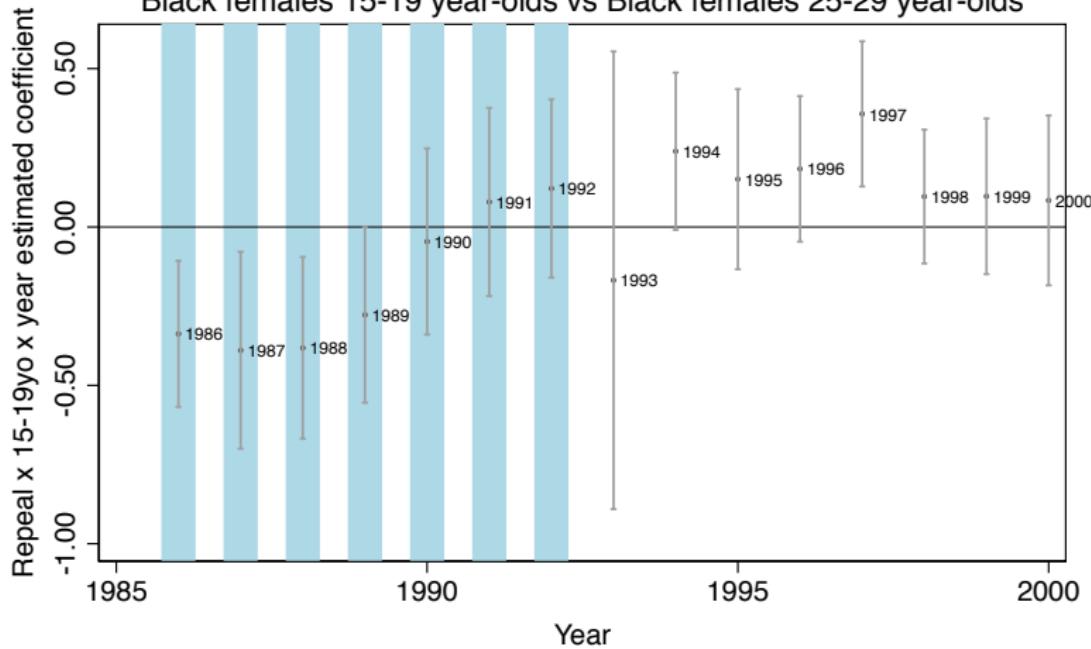
```
. list id wm wf bf bm age repeal year in 1/30
```

	id	wm15	wf15	bf15	bm15	age	repeal	year
1.	1	1	0	0	0	15	0	1985
2.	1	1	0	0	0	15	0	1986
3.	1	1	0	0	0	15	0	1987
4.	1	1	0	0	0	15	0	1988
5.	1	1	0	0	0	15	0	1989
6.	1	1	0	0	0	15	0	1990
7.	1	1	0	0	0	15	0	1991
8.	1	1	0	0	0	15	0	1992
9.	1	1	0	0	0	15	0	1993
10.	1	1	0	0	0	15	0	1994
11.	1	1	0	0	0	15	0	1995
12.	1	1	0	0	0	15	0	1996
13.	1	1	0	0	0	15	0	1997
14.	1	1	0	0	0	15	0	1998
15.	1	1	0	0	0	15	0	1999
16.	1	1	0	0	0	15	0	2000
17.	2	0	1	0	0	15	0	1985
18.	2	0	1	0	0	15	0	1986

DDD Results

Estimated effect of abortion legalization on gonorrhea

Black females 15-19 year-olds vs Black females 25-29 year-olds



Whisker plots are estimated coefficients of DDD coefficients

My original conclusions

- Model made narrow predictions of a *parabola* within a given window but only for the treatment cohort
- Amazingly we actually found that very shape in the DD – did we vindicate Gruber, et al. and Donohue and Levitt then?
- Also used older group as within-state controls in a DDD, and still found the parabola, though not as great a look as DD which is a bit of a red flag
- Paper also illustrates the usefulness of having a specific theoretical prediction. Limits the number of competing hypotheses (Popperian type of reasoning).
- But was I done? Look back at the table

Going beyond Cornwell and Cunningham (2013)

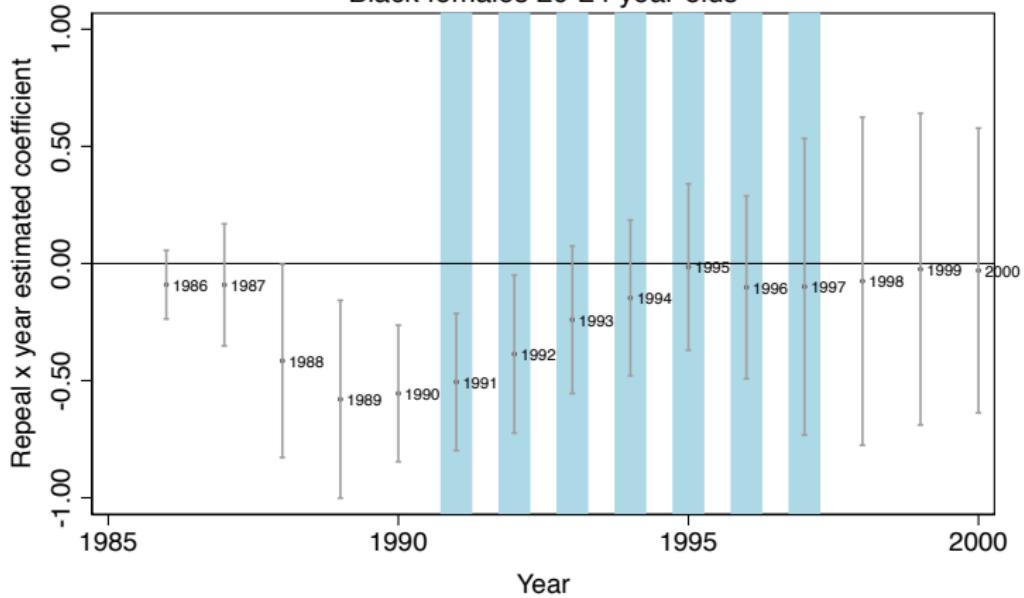
	CDC Surveillance Data in Calendar Year																
Age in calendar year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	
15	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	
16	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	
17	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	
18	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	
19	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	
20	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	
21	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	
22	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	
23	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	
24	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	
25	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	
26	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	
27	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	
28	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	
29	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	
Repeal (1)	0	0	0	0	0	0	1	2	3	4	5	5	5	5	5	5	
No Repeal (2)	0	0	0	0	0	0	0	0	0	1	2	3	4	5	5	5	
Difference (3)	0	0	0	0	0	0	1	2	3	3	3	2	1	0	0	0	

Number of cohorts (age 20-24) exposed, reforms in 71, 74

Figure: Second theoretical prediction - this time for 20-24 year olds

Estimated effect of abortion legalization on gonorrhea

Black females 20-24 year-olds



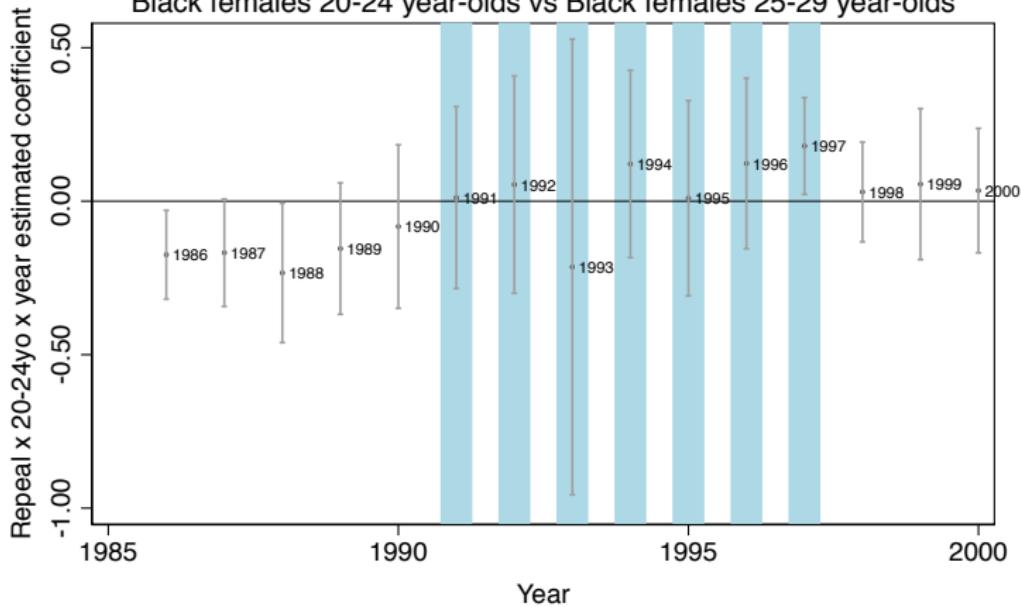
Whisker plots are estimated coefficients of DD estimates

Second prediction fails second DD model

- Ugh. *lo tov* (Hebrew to English: not good)
- Well, maybe DDD will look better?

Estimated effect of abortion legalization on gonorrhea

Black females 20-24 year-olds vs Black females 25-29 year-olds



Whisker plots are estimated coefficients of DDD coefficients

Second predictions fails DDD too

- Notice that when we exploited just one testable prediction, we found evidence
- But when we exploit all of the testable predictions, the results fall apart, suggesting original DD was spurious
- Imagine for a moment, though – what if we had seen the group-time ATT moving with the cohort as they aged?
- Other alternative is the repeal-Roe effects dissipate by early to late 20s, but what does Ockham's Razor say is the more credible explanation?
- Perhaps the Gruber, et al. (1999) and Donohue and Levitt (2001) hypothesis was always spurious

Stata replication

Let's replicate this using the abortion.do file. Pay close attention to the stacking of the data by group-state, not just state, and the exact way in which the interactions must therefore be constructed

Falsification test with alternative outcome

- The within-group control group (DDD) is a form of placebo analysis using the same *outcome*
- But there are also placebos using a *different outcome* – but you need a hypothesis of mechanisms to figure out what is in fact a *different outcome*
- Figure out what those are, and test them – finding no effect raises the epistemological credibility of the first result, interestingly
- Cheng and Hoekstra (2013) examine the effect of castle doctrine gun laws on non-gun related offenses like grand theft auto and find no evidence of an effect

Rational addiction as a placebo critique

Sometimes, an empirical literature may be criticized using nothing more than placebo analysis

"A majority of [our] respondents believe the literature is a success story that demonstrates the power of economic reasoning. At the same time, they also believe the empirical evidence is weak, and they disagree both on the type of evidence that would validate the theory and the policy implications. Taken together, this points to an interesting gap. On the one hand, most of the respondents claim that the theory has valuable real world implications. On the other hand, they do not believe the theory has received empirical support."

Placebo as critique of empirical rational addiction

- Auld and Grootendorst (2004) estimated standard “rational addiction” models (Becker and Murphy 1988) on data with milk, eggs, oranges and apples.
- They find these plausibly non-addictive goods are addictive, which casts doubt on the empirical rational addiction models.

Placebo as critique of peer effects

- Several studies found evidence for “peer effects” involving inter-peer transmission of smoking, alcohol use and happiness tendencies
- Christakis and Fowler (2007) found significant network effects on outcomes like obesity
- Cohen-Cole and Fletcher (2008) use similar models and data and find similar network “effects” for things that *aren't* contagious like acne, height and headaches
- Ockham's razor - given social interaction endogeneity (Manski 1993), homophily more likely explanation

Now on to some models focused on covariates

- We will discuss two papers now: Abadie (2005) and Sant'Anna and Zhao (2020)
- In some ways you can think of Abadie (2005) as the father of Callaway and Sant'Anna (which we discuss under the differential timing section) only here we don't use differential timing
- The Abadie (2005) is really used best for longitudinal data or repeated cross sections where treatment occurs at one point in time
- But like CS, it's used for modeling the differential selection based on what you think are covariates, which means you need to think carefully about what those might be

High level

Short and readable, though when it gets into theorems and proofs,
it's deep

"A good way to do econometrics is to look good for natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us. – Daniel McFadden

Why do this?

- No randomization. Remember, DD doesn't require randomization – it requires a version of parallel trends
- Treatment is selecting on observable covariates

DD method at its core

- Abadie (2005) proposed a method to estimate the ATT
- The method is a DD type estimator, but isn't using TWFE
- You need treatment and comparison group, before and after treatment
- But you also need conditional parallel trends (based on X)
- Kind of neat but it's a lot like Callaway and Sant'Anna, only not for differential timing interestingly

Why do this anyway?

- In a DD, we may need to control for X because treatment is only conditional on X
- But in TWFE, when you controlling for baseline X, it gets absorbed by the unit fixed effects
- And when you use time-varying controls, you can get even stranger weights than we had already seen from Bacon
- To get around this, he won't be proposing an OLS model with fixed effects, but he will be proposing a simpler difference in means in a DD framework by a specific form of weight called the propensity score which has been estimated with polynomial series

Not a critique, but an estimator

- Goodman-Bacon was a critique of TWFE, not a proposed estimator
- CS was a proposed estimator
- Abadie is a proposed estimator
- Let's look at the steps involved

Three step method

- ① Compute each unit's "after minus before" which is the DD part
- ② Then estimate a propensity score which you'll use to weight each unit
- ③ Finally, compare weighted changes in "after minus before" for treatment versus comparison groups

Inference will take into account step two, which is often the sticky part (see Abadie and Imbens matching paper which shows you can't use the bootstrap for matching, but you can for propensity scores)

Like CS, you can have heterogeneity too

Terms

- t is year of treatment which doesn't vary across units (so no differential timing)
- Y^1 and Y^0 are potential outcomes (counterfactual versus actual)
- D is 1 or 0 based on group and time
- b is the “baseline” which is similar to CS using g as the one year pre-treatment
- X are “baseline” covariates **only** – they do not vary over time, which means propensity scores are estimated off the b period **only**

Assumptions

Kind of common for this propensity score literature to only have two assumptions. But usually the first conditional independence. Now it is parallel trends because this is DD

- ① Conditional parallel trends

$$E[Y_t^0 - Y_b^0 | D = 1, X_b] - E[Y_t^0 - Y_t^0 | D = 0, X_b]$$

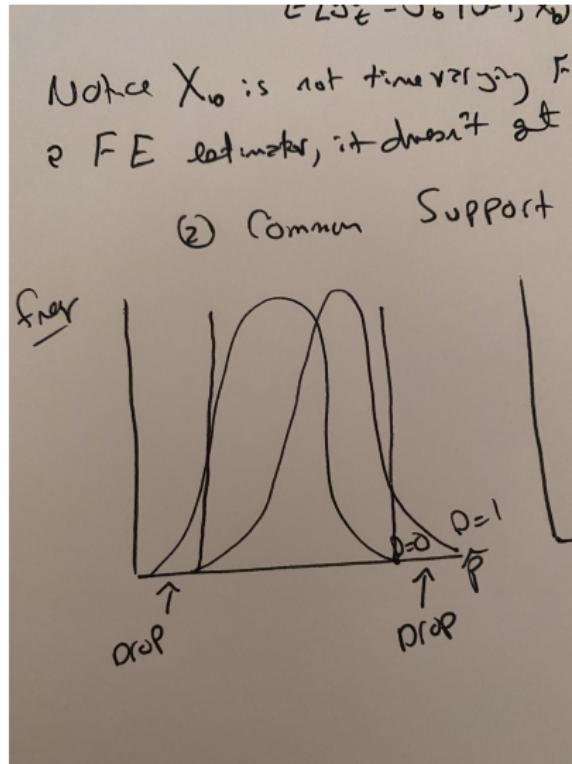
(Notice the b subscript. What is that you think?)

- ② Common support

$$Pr(D = 1) > 0 \text{ } \& \text{ } Pr(D = 1 | X) < 1$$

Let's see a picture of common support that I drew. Apologies it's horrible

Trimming the propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

$$ATT = E[Y_t^1 - Y_t^0 | D_t = 1] \quad (65)$$

Abadie's estimator

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right] \quad (66)$$

versus CS

$$ATT(g, t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right] \quad (67)$$

These are also using the “Horvitz” (non-normalized) weights from the inverse probability weighting literature

Propensity scores

- Usually there's almost no guidance that I've seen in how to estimate the propensity score except to say use logit or probit
- Dehejia and Wahba (2002) anyway
- Not so here – this is semi-parametric in the sense that you have to use a series of polynomials based on the X controls
- Weirdly, you can use OLS linear probability models (which I've never seen) or something called series logit estimation

Estimating propensity scores

It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions

$$\widehat{Pr}(X_b) = \widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots \varepsilon \quad (68)$$

$$\widehat{Pr}(X_b) = F(\widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots) \quad (69)$$

Stata

Stata command is called `absdid`

You need treatment (varname), X variables (can be a list), the order in which the variables occur (weird, but results change if the order changes), and the exact estimator (LPM or logit)

Why not try it yourselves using the LaLonde NSW job trainings program data?

https://github.com/scunning1975/mixtape/raw/master/nsw_mixtape.dta

When to use it

- LaLonde longitudinal data where you have a baseline and a follow-up
- Repeated cross-sections
- Controls will cause the estimates to vary based on the type of approximation you use (logit for instance vs LPM) and the order in which the polynomials are used

Doubly Robust

- DR literature can be found in the older matching literature (Hirano and Imbens 2001; etc.)
- They combine regression and weighting estimators into one specification and are consistent so long as:
 - The regression specification for the outcome is correctly specified
 - The propensity score specification is correctly specified
- DR is a class of estimators that possess this property
- You're basically controlling for X twice: with a linear regression, with a propensity score, to cover your bases

DR DD

- Sant'Anna and Zhao (2020) incorporate DR into DD
- Think of it as a way of incorporating X into our new DD models (I'll show you why)
- It's in the engine of Callaway and Sant'Anna (2020) so we badly need to understand it
- Dense paper with hairy notation; I'll do my best

Literature

- Pedro is excellent at bridging gaps while simultaneously moving the ball forward – this is a good example
- The outcome regression part of DR goes back to Heckman, et al. (1997) and I use this in section 5.3.2 (“Bias correction”) of the mixtape
- The propensity score part goes back to Abadie (2005) which we’ve discussed
- New work on machine learning fits into this

Organization

- Basic assumptions for DD with covariates
- TWFE assumptions for DD with covariates
- Estimation alternative to TWFE with covariates
- Efficiency and semiparametric bounds

Insurance

- We covered covariates with Abadie (2005); why again?
- Maybe you're unsure whether the propensity score was properly specified
- How about some insurance?
- Two strikes instead of one

ATT

- DD *always* estimates the ATT because it's only the treatment effect for the treatment group in the post-treatment period
- It is not the ATE, or the LATE

$$\delta = E[Y_{it}^1 - Y_{it}^0 | D_i = 1]$$

Basic assumptions of DD

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is hairy, and so I've chosen to focus on the panel data for this talk, but results are similar for repeated cross sections

Basic assumptions of DD

Assumption 2: Conditional parallel trends

If you were putting covariates into your DD regression, then you were assuming conditional parallel trends

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Basic assumptions of DD

Assumption 3: Common support or overlap

For some $e > 0$, the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on X is $\leq 1 - e$.

Intuition of assumption 3: Called overlap or common support.
Means there is at least a small fraction of the population that is treated and that for every value of the covariates X there is at least a small chance that the unit is not treated. It's called common support when it's a propensity score but it's just about the distribution of treatment and control across values of X .

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997
- Or we can use the propensity score approach of Abadie (2005)
- What about TWFE? Hold off on that question for a second until we look at the estimators based on Assumptions 1-3

Outcome regression

This is the Heckman, et al. (1997) approach where the outcome evolution is modeled with a regression

$$\hat{\delta}^{OR} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i)) \right]$$

where \bar{Y} is the sample average of Y among units in the treatment group at time t and $\hat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$. See my Section 5.3.2 for more about this.

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\hat{\delta}^{ipw} = \frac{1}{E_N[D]} E \left[\frac{D - \hat{p}(X)}{1 - \hat{p}(X)} (Y_1 - Y_0) \right]$$

where $\hat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

Caveat

- Outcome regression needs $\hat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\hat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified
- Well why don't we just use TWFE? I've never heard anyone complain about including covariates in TWFE and I've been doing it my entire adult life, so we're good right?
- Depends on if you want to assume three more things.
(Mixtape didn't know about this...)

TWFE

Here's the TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose the next three assumptions (let's first look at estimators based on .

Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

$$E[Y_1^1 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

Conditional parallel trends implies

$$E[Y_1^0 - Y_0^0 | D = 1, X] = E[Y_1^0 - Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] - E[Y_0^0 | D = 1, X] = E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0^0 | D = 1, X] + E[Y_1^0 | D = 0, X] - E[Y_0^0 | D = 0, X]$$

$$E[Y_1^0 | D = 1, X] = E[Y_0 | D = 1, X] + E[Y_1 | D = 0, X] - E[Y_0 | D = 0, X]$$

Last line from the switching equation. This gives us:

$$E[Y_1^0 | D = 1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Collecting terms

$$E[Y_1^1|D=1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$

$$E[Y_1^0|D=1, X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

$$E[Y_1^1|D=1, X] - E[Y_1^0|D=1, X]$$

$$= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X)$$

$$= \delta + (\theta_1 X - \theta_2 X)$$

By allowing for the possibility that $\theta_1 X \neq \theta_2 X$, we open up the possibility of bias from TWFE which is zero under three additional assumptions.

Assumption 4

The implications of that TWFE regression with assumptions 1-3 gave us those previous expressions which then require placing further restrictions on treatment effects and trends when estimating with TWFE.

TWFE Assumption 4: Homogenous treatment effects in X

$$E[Y_1^1 - Y_1^0 | X, D = 1] = E[Y_1^1 - Y_1^0 | D = 1]$$

This is because when you difference out those previous equations, you need θX to cancel to leave you with δ which implies homogeneity in X .

X-specific trends

TWFE places restrictions on trends for the two groups too. Take conditional expectations of our TWFE equation.

$$E[Y_1|D=1] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}$$

$$E[Y_0|D=1] = \alpha_1 + \alpha_3 + \theta X_{10}$$

$$E[Y_1|D=0] = \alpha_1 + \alpha_2 + \theta X_{01}$$

$$E[Y_0|D=0] = \alpha_1 + \theta X_{00}$$

X-specific trends

Now take DD:

$$\delta^{DD} = \left((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10}) \right) - \left((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00}) \right)$$

Eliminating terms, we get:

$$\delta^{DD} = \delta + (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})$$

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

For $D = 0, 1$, we need “no X-specific trends in both groups”:

$$E[Y_1 - Y_0 | D = d, X] = E[Y_1 - Y_0 | D = d]$$

Intuition: Sant'Anna and Zhao (2020) say in footnote 4 “[this] follows from analogous arguments” which is the previous slides’ manipulation of terms. Key is to remember these are time-varying covariates so they don’t cancel out within treatment category, so you need the trends in X to cancel out.

Without these six, in general TWFE will not identify ATT. Unclear how off it’ll be, but it will be biased is the point.

Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
 - Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
 - Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
 - TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Doubly robust combines them to give us insurance. That's the basic idea. Gives you two chances to be wrong

Next step

- Introduction to the three prior covariate DD models
- Assumptions – check
- Hints about combining OR and IPW
- Now we move into *estimation* phase
- Let's see what doubly robust estimator looks like
- As before, I'm going to only stick to the panel data expressions
bc all repeated cross-section does is add in some terms

Estimation

Some terms

$p(x)$: propensity score model

$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}$, where $\mu(X)$ is a model for

$m_{d,t} = E[Y_t | D = d, X = x]$

So that means $\mu_{1,\Delta}$ is just the treatment group's change in average Y for each $X = x$

We're off to see the (DR) wizard!

Population DR DD model for panel data

$$\delta^{dr} = E \left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E \left[\frac{p(X)(1-D)}{(1-p(X))} \right]} \right) (\Delta Y - \mu_{0,\Delta}(X)) \right]$$

Notice the propensity score modifying the control group (second term inside parentheses) *and* the $\mu(X)$ term modifying the long difference. This is the idea of the doubly robust – you only need one of these models to be correctly specified, not both.

Sidebar: This is also one of the options in the Callaway and Sant'Anna (2020) DD estimator. It lets you pick IPW, regression (OR) or DR. Pedro usually recommends DR because of its advantages.

Efficiency

- Last step is inference
- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DID estimator is also DR for inference
- Let's skip to Monte Carlos

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0.257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

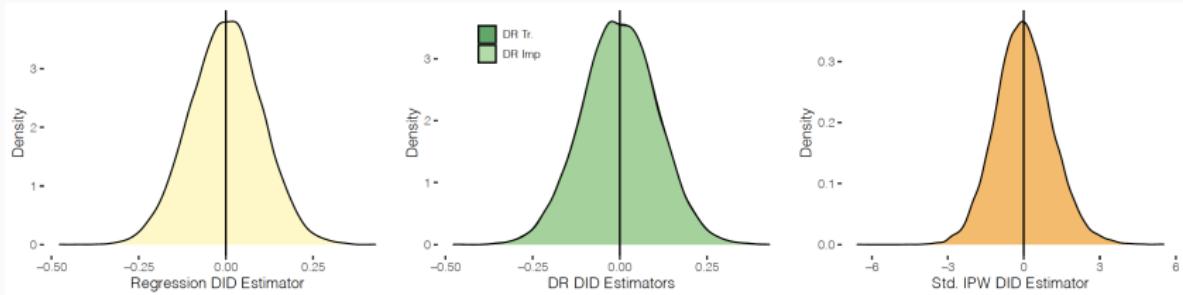
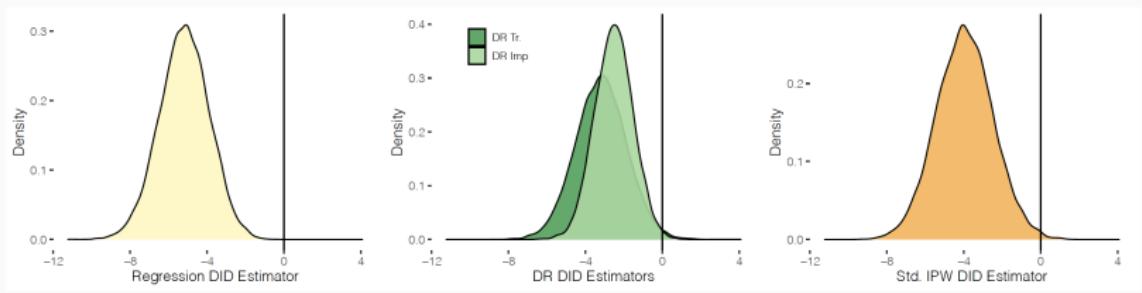


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



To a kid with a hammer, everything is a nail

- Use the right tool (oven) for the job (making lasagna), not the same tool (hammer) regardless of the job (making lasagna)
- One of the main things I learned from this paper was again biases in TWFE with covariates – Mixtape and MHE don't cover this
- This method only needed three assumptions not the six for TWFE
- Like everything Pedro does, there is code for this but it's only in R – DRDID
- But it's one of the main options in Callaway and Sant'anna under differential timing, and therefore it's crucial we understand this
- But you still have to have specified correctly either at least the outcome model or propensity score model

Hidden curriculum
Foundational causality stuff
Regression discontinuity designs
Instrumental variables
Twoway fixed effects estimator
Difference-in-differences
Comparative case studies
Matching and weighting
Concluding remarks

Two group case
Event study
Covariates
Differential timing
Revisiting event studies
Alternative DD estimators
Conclusion

I ❤️ federalism
(for the natural experiments)

Tweets 30.4K Following 5,933 Followers 11.8K Likes 80.5K Lists 1 Moments 0

Edit profile

Differential timing

- We've been considering situations where treatment occurs in one area for the most part – the two group case
- But the modal situation is when there is *differential timing* – groups treated at different time periods
- This happens in America usually because each area (state, municipality) will adopt a policy whenever they want to, which creates tendencies for roll out to occur
- Turns out, this is actually problematic *for TWFE*.
- Remember: $TWFE \neq DD$. Not all estimators are made the same.

Strict exogeneity

- Two main identification assumptions for TWFE
 - ① Strict exogeneity

$$E[\varepsilon_{it}|X_{i1}, X_{i2}, \dots, X_{iT}, c_i] = 0$$

- ② Full rank (regressors vary over time for at least some i)
- This is violated with differential timing and heterogeneity, but let's see why

Regression equation

We define the ATT as

$$\beta_{gp} = E[(Y_{gpit}^1) - E(Y_{gipt}^0) | g, p]$$

which varies by group and period (i.e., differential timing)

TWFE specification

$$Y_{gpit} = \beta_{gp} X_{gp} + \lambda_g + \lambda_p + \varepsilon_{gp}$$

Violation of strict exogeneity

$$E[Y_{gpit}|g, p, X_{gp}] = E[\beta_{gp}|X_{gp} = 1]X_{gp} + \lambda_g + \lambda_p + \mu$$

The first term is the overall ATT. The last term is the composite error term where μ equals $(\beta_{gp} - E[\beta_{gp}|X_{gp} = 1]X_{gp})$. It is not necessarily mean-zero condition on g, p and the group and time varying X .

TWFE will only identify this “overall average ATT” if all groups have the same ATT or only one treatment group (i.e., no differential timing).

This implies strict exogeneity is violated with heterogeneity and differential timing because the composite error term is correlated with treatment and group fixed effects

Differential timing

- In summary, differential timing was thought to be a simple extension of two group case, so people used TWFE to estimate ATT
- But it turns out TWFE is misspecified under differential timing with heterogeneity
- Event studies are also flawed for similar reasons as static parameter
- Here we will discuss recent work in econometrics
- Let's learn a paper that I will use for illustration

Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine



Cheng Cheng

Mark Hoekstra

Abstract

From 2000 to 2010, more than 20 states passed so-called "Castle Doctrine" or "stand your ground" laws. These laws expand the legal justification for the use of lethal force in self-defense, thereby lowering the expected cost of using lethal force and increasing the expected cost of committing violent crime. This paper exploits the within-state variation in self-defense law to examine their effect on homicides and violent crime. Results indicate the laws do not deter burglary, robbery, or aggravated assault. In contrast, they lead to a statistically significant 8 percent net increase in the number of reported murders and nonnegligent manslaughters.

Summary

- Cheng and Hoekstra (2013) are interested in whether expansions to “castle doctrine statutes” at the state level increase or decrease gun violence.
- Prior to these expansions, English common law principle required “duty to retreat” before using lethal force against an assailant except when the assailant is an intruder in the home
 - The home is one’s “castle” – hence, “castle doctrine”
 - When intruders threatened the victim in the home, the duty to retreat was waived and lethal force in self-defense was allowed

Castle doctrine law explained

- In 2005, Florida passed a law that expanded self-defense protections beyond the house
 - 2000 to 2010, 21 states explicitly put “castle doctrine” into statute, and (more importantly) extended it to places outside the home
 - In other words, 21 states removed the duty to retreat in specified circumstances
- Other changes:
 - Presumption of reasonable fear is added
 - Civil liability for those acting under the law is removed

Economic theory predicts more lethal homicides

- Workers supply legal or illegal labor and are therefore responsive to costs and benefits
- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- If people are rational, then lowering the price of lethal self-defense should increase lethal homicides

Economic theory also predicts less crime from deterrence

- Although deterrence is a theoretical possibility, note that the goal of the laws was to protect enhance victim rights, not deter crime
- Testable prediction with data and same design

Treatment passage

- Summary:

- 21 states passed laws removing “duty to retreat” in places outside the home
- 17 states removed “duty to retreat” in any place one had a legal right to be
- 13 states include a presumption of reasonable fear
- 18 states remove civil liability when force was justified under law

Cheng and Hoekstra's identification strategy

- Panel fixed effects estimation

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3(CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes “region-by-year fixed effects”

Data

- FBI Uniform Crime Reports Part 1 Offenses (2000-2010)
 - State-level crime rates, or “offenses per 100,000 population”
 - Falsification outcomes: motor vehicle theft and larceny
- Dataset on justifiable homicides by private citizens

Outcomes (in order)

- Deterrence and homicide outcomes:
 - ① Burglary: the unlawful entry of a structure to commit a felony or a theft
 - ② Robbery: the taking or attempting to take anything of value from the care, custody or control of a person or persons by force or threat of force or violence and/or putting the victim in fear
 - ③ Aggravated assault: unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury
- Homicide categories
 - ① Total homicides – murder plus non-negligent manslaughter (~14,000 per year)
 - ② Justifiable homicides by private citizens (~250/year)

Inference: Clustering

- Statistical inference: cluster standard errors at the state level
 - Are disturbances random draws from individually identical distribution?
 - It's likely that within a state, unobserved determinants of crime are serially correlated
 - They follow Bertrand, Duflo and Mullainathan (2004) and adjust for serial correlation in unobserved disturbances within states at the level of the treatment

Inference: Fisher's sharp null

- How likely is it that we estimate effects of this magnitude when using randomly chosen pre-treatment time periods and randomly assigning placebo treatments?
- Randomizes dates within-state for the pre-treatment period (<2000)
- Randomization inference and exact p-values

Region-by-year fixed effects

- Absent passing castle doctrine laws, outcomes in these 21 states would have changed similar to other states in their same region
 - Recall the “region-by-year fixed effects” in the X term
 - By including “region-by-year fixed effects”, they are arguing that unobserved changes in crime are running “parallel” to the treatment states within region over time
 - Need not hold across regions since the across region variation is not being used in this analysis due to the saturation of the model with “region-by-year fixed effects”

State specific time trends

- Alabama, et al. dummy interacted with TREND which equals 1 in 2000, 2 in 2001, ..., 11 in 2010
- Forces the identification to come from variation in outcomes around the state-specific linear trend
 - Outcomes must be large enough and different enough from a state-specific linear trend otherwise it is collinear with the state-trend
 - Same argument applies to any control though
 - Goodman-Bacon (2019) suggests group-trends are less taxing and satisfying than unit-specific trends

Control variables

- Controls (X matrix in earlier equation)
 - Full-time police employment per 100,000 state residents from the LEKOA data (FBI data)
 - Persons incarcerated in state prison per 100,000 residents
 - Shares of white/black men in 15-24 and 25-44 age groups
 - State per capita spending on public assistance
 - State per capita spending on public welfare

Parallel Leads

- Look at each set of treatment states against never-treated figure by figure (rare)
- Use a one-period lead in the regression model (not as common)
- I'm going to look at event study coefficients (most common)

Step one: Falsification test

- Policy-makers are not just randomly flipping coins when passing laws, but presumably do so because of things they observe on the ground
- Address concerns up front this isn't driven by spurious crime results
- Cheng and Hoekstra (2013) present falsification of larceny and motor vehicle theft first, then results

Step one (cont.)

- Results will be presented separately under six different specifications
 - Each new specification adds more controls
- Pop quiz: What should you expect to find on key variables of interest when conducting a falsification and why?

Answer

- No statistically significant association between the CDL passage and the placebos; preferably precise zeroes
- No association on the one-year lead either
- Basically, you should not find effects where there are no theoretical policy effects; gun laws shouldn't affect non-violent offenses

Step one (cont.)

- How do you interpret coefficients?
 - His model is “log outcomes” regressed onto a dummy variable (level), so these are semi-elasticities and approximate percentage changes – but you should transform them by taking the exponential of each coefficient and then differencing it from one to find the actual percentage change
 - Ex: CDL = -0.0137 (column 12, Table 3, “Log (larceny rate)” outcome.) $\text{Exp}(-0.0137) = 0.986$, and so $1-0.986 = 1.4$. Thus, CDL reduced larceny rates by 1.4 percent, which is not statistically significant.

Results – Falsification Exercise

Table 3: Placebo Tests

	OLS - Unweighted					
	7	8	9	10	11	12
Panel A: Larceny	Log (Larceny Rate)					
Castle Doctrine Law	0.00745 (0.0227)	0.00145 (0.0205)	-0.00188 (0.0210)	-0.00445 (0.0226)	-0.00361 (0.0201)	-0.0137 (0.0228)
One Year Before Adoption of Castle Doctrine Law				-0.0103 (0.0114)		
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft	Log (Motor Vehicle Theft Rate)					
Castle Doctrine Law	0.0767* (0.0413)	0.0138 (0.0444)	0.00814 (0.0407)	0.00775 (0.0462)	0.00977 (0.0391)	-0.00373 (0.0361)
One Year Before Adoption of Castle Doctrine Law				-0.00155 (0.0287)		
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft					Yes	
State-Specific Linear Time Trends						Yes

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics.

Step two: testing the deterrence hypothesis

- Having found no effect on their placebos, Cheng and Hoekstra (2013) examine the effect of CDL on three deterrence outcomes: burglary, robbery and aggravated assault
 - They will, again, have six specifications per outcome in the “weighted” regression, and then another five for the “unweighted” regression
- Pop quiz: What does deterrence look like?

Answer

- Negative signs on the CDL variable is consistent with deterrence – these crimes were “deterred”, in other words
- Based on early work by Becker (1968) and 1970s work by his student Isaac Ehrlich; higher probabilities of getting hurt in public may cause offenders to avoid violence in public altogether
- Bounds on the magnitudes from the standard errors are used to provide some confidence about the estimates as well

Results – Deterrence

	OLS - Weighted by State Population						OLS - Unweighted					
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Burglary												
Castle Doctrine Law							Log (Burglary Rate)					
	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237		0.0572**	0.00961	0.00663	0.00277	0.00683
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)		(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)
One Year Before Adoption of												
Castle Doctrine Law							-0.0201					
							(0.0139)					
Panel B: Robbery												
Castle Doctrine Law							Log (Robbery Rate)					
	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*		0.0448	0.0320	0.00839	0.00552	0.00874
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)		(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)
One Year Before Adoption of												
Castle Doctrine Law							-0.0156					
							(0.0167)					
Panel C: Aggravated Assault												
Castle Doctrine Law							Log (Aggravated Assault Rate)					
	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414		0.0555	0.0698	0.0343	0.0305	0.0341
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)		(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)
One Year Before Adoption of												
Castle Doctrine Law							-0.00343					
							(0.0161)					
Observations	550	550	550	550	550	550		550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes						Yes	
State-Specific Linear Time Trends						Yes						Yes

Conclusion

- “In short, these estimates provide strong evidence against the possibility that castle doctrine laws cause economically meaningful deterrence effects” (p. 17)
 - Translation: They can't find evidence of large deterrence effects
- “Thus, while castle doctrine law may well have benefits to those legally justified in protecting themselves in self-defense, there is no evidence that the law provides positive spillovers by deterring crime more generally” (p. 17)
 - They note in footnote 24 that they cannot measure the benefits to victims whose crimes were deterred, or the benefits from lower legal costs; their focus is limited to whether it deterred the crimes, not whether the net benefits from the laws were positive
 - Obviously, if there is no deterrence, though, then the net benefits are lower from CDL than they would be if they did deter

Step 3: Homicides

- The key finding in this study focuses on CDL and its effect on homicides and non-negligent manslaughter
- Pop quiz: what should the sign on CDL be here?

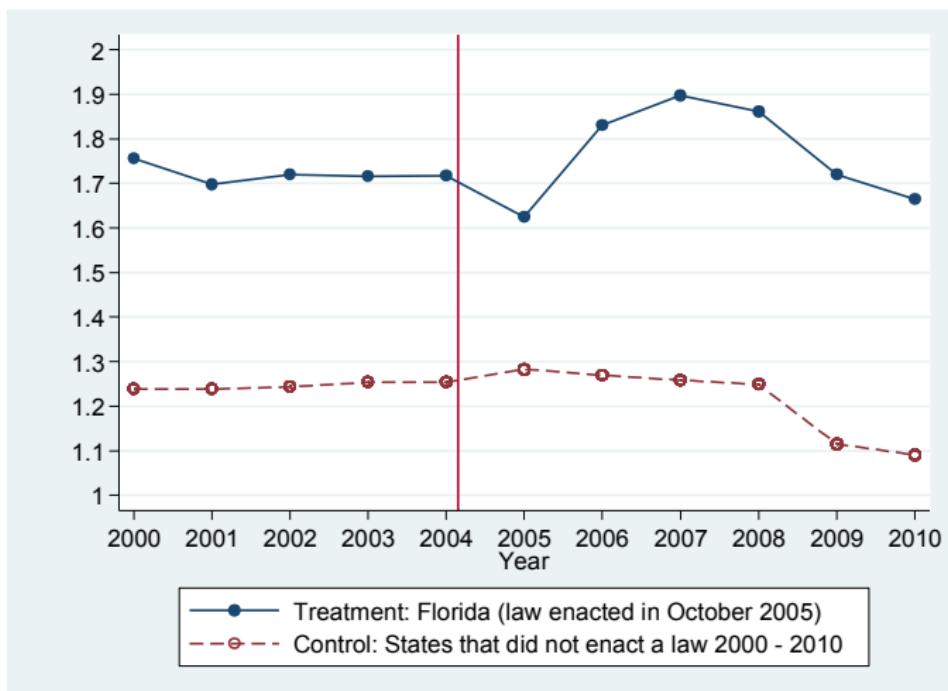
Answer

- Effects should be **positive**
- Cheng and Hoekstra want to show the raw data, but have differential timing
- Differential timing means you can't show pre-treatment raw data for the never-treated groups
- So they show it one by one – which isn't the most aesthetically pleasing way to do it, but which has the benefit of being transparent

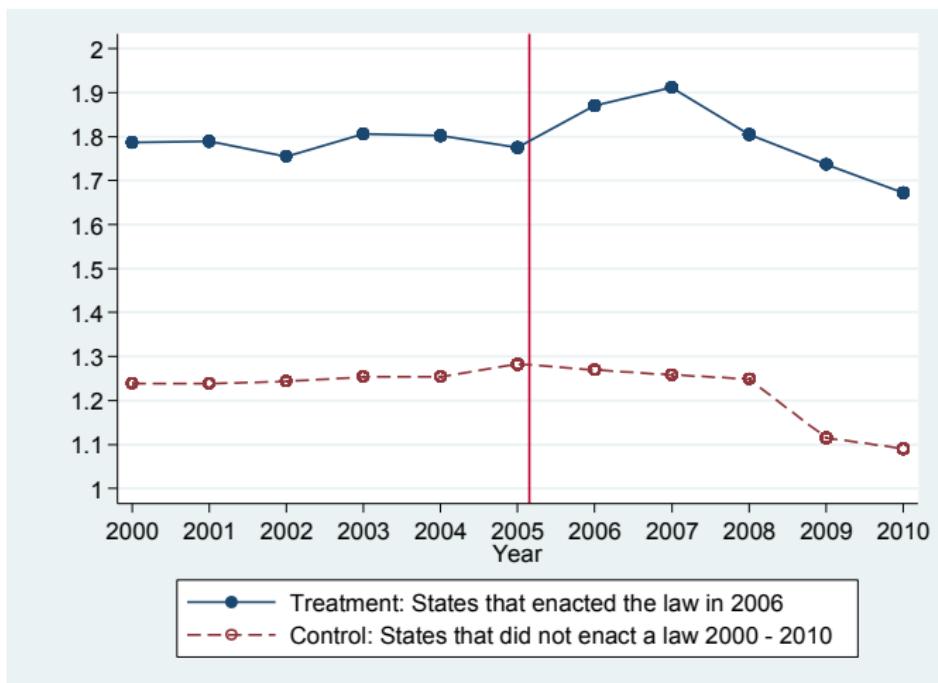
Parallel pre-treatment trends

- Keep your eyes on whether pre-treatment trends are parallel for treatment and control groups
 - Remember, though – he needs parallel trends within-region – these figures don't show that
 - But starting with pictures and raw data has value

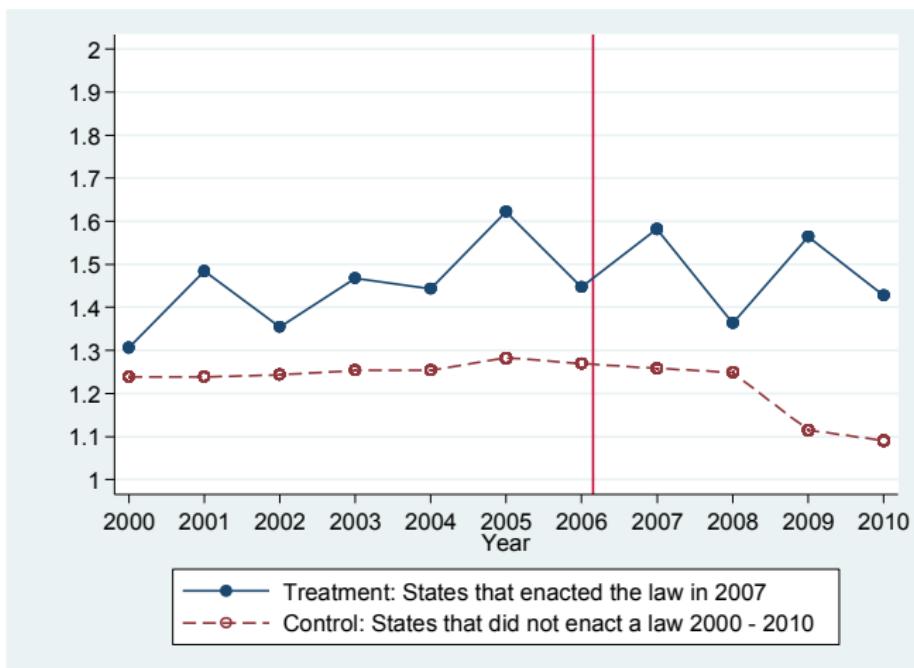
Log Homicide Rates – 2005 Adopter = Florida



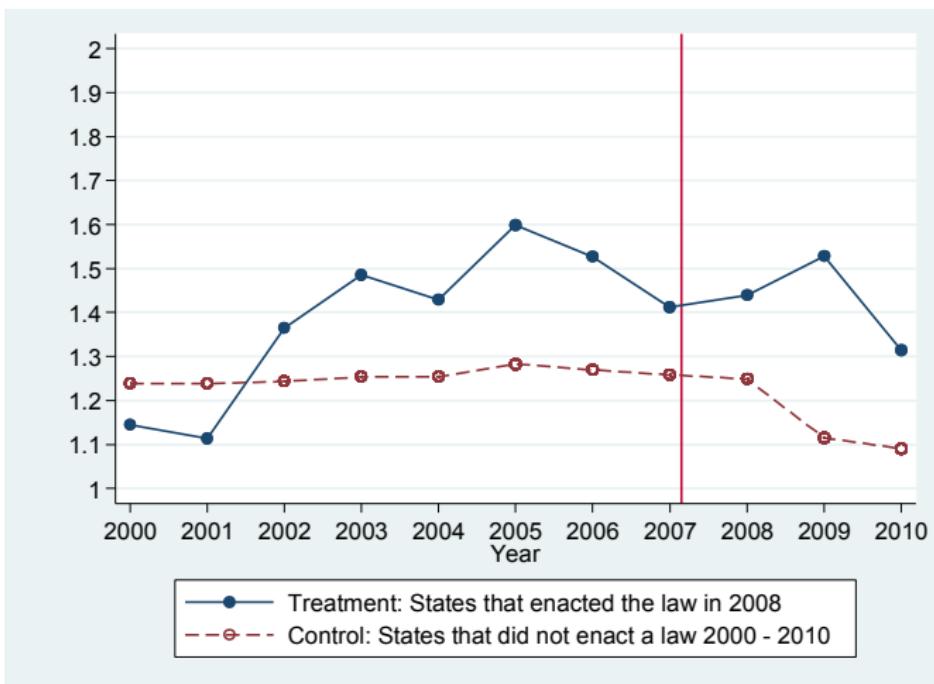
Log Homicide Rates – 2006 Adopter (13 states)



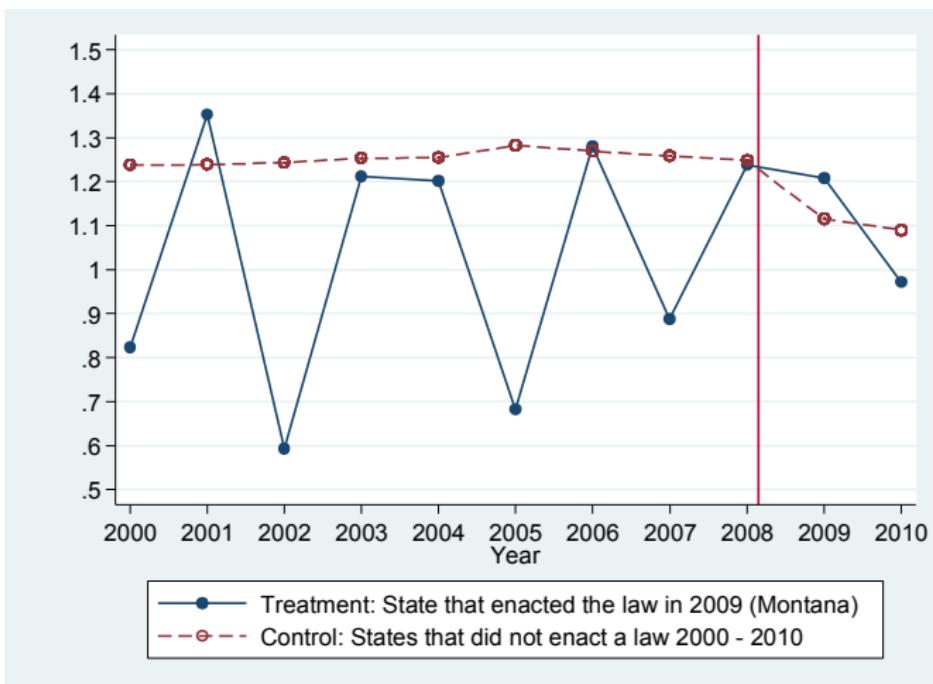
Log Homicide Rates – 2007 Adopter (4 states)



Log Homicide Rates – 2008 Adopter (2 states)



Log Homicide Rates – 2009 Adopter = Montana



Modeling

- He uses a class of estimators more appropriate for “counts” called “count models”, like the negative binomial estimated with maximum likelihood
- Results are robust to least squares and count models

Homicide – Negative Binomial; Murder – OLS

	1	2	3	4	5	6
<u>Panel C: Homicide (Negative Binomial - Unweighted)</u>						
Castle Doctrine Law	0.0565*	0.0734**	0.0879***	0.0783**	0.0937***	0.108***
	(0.0331)	(0.0305)	(0.0313)	(0.0355)	(0.0302)	(0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352		
				(0.0260)		
Observations	550	550	550	550	550	550
<u>Panel D: Log Murder Rate (OLS - Weighted)</u>						
Castle Doctrine Law	0.0906**	0.0955**	0.0916**	0.0884**	0.0981**	0.0813
	(0.0424)	(0.0389)	(0.0382)	(0.0404)	(0.0391)	(0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110		
				(0.0230)		
Observations	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Fisher sharp null

Move the 11-year panel back one year at a time (covering 1960-2009) and estimate 40 placebo “effects” of passing CDL 1 to 40 years earlier

Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	-0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

My replication using event study plots



Figure: Homicide event study plots using coefplot

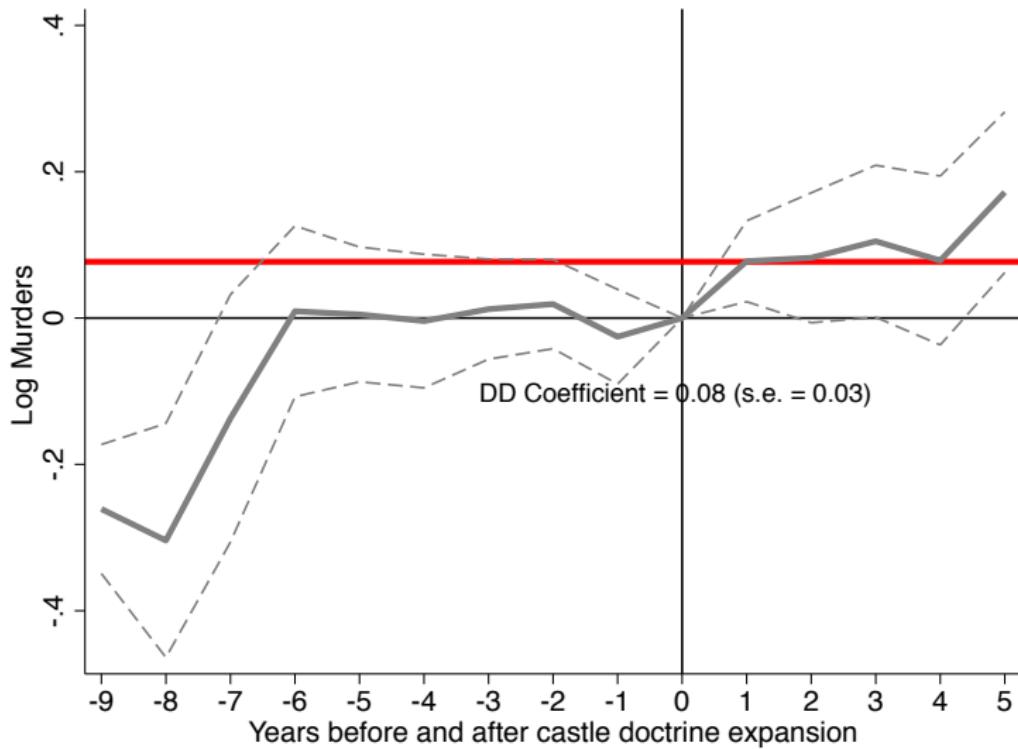


Figure: Homicide event study plots using two way

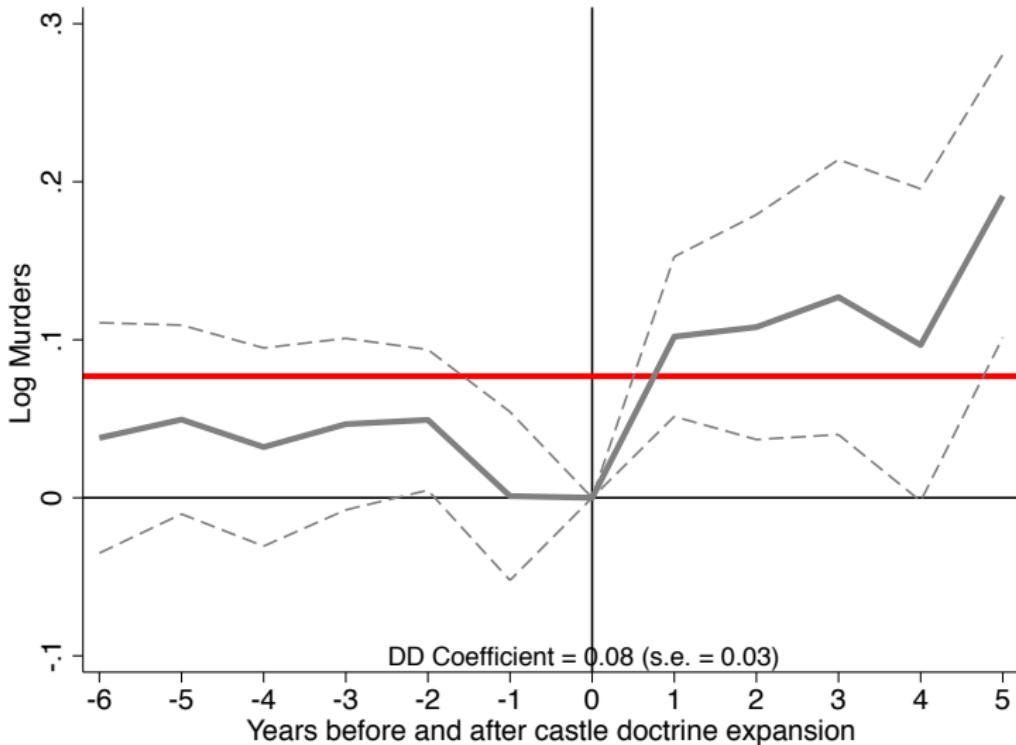


Figure: Homicide event study plots using two way and force early leads into one coefficient

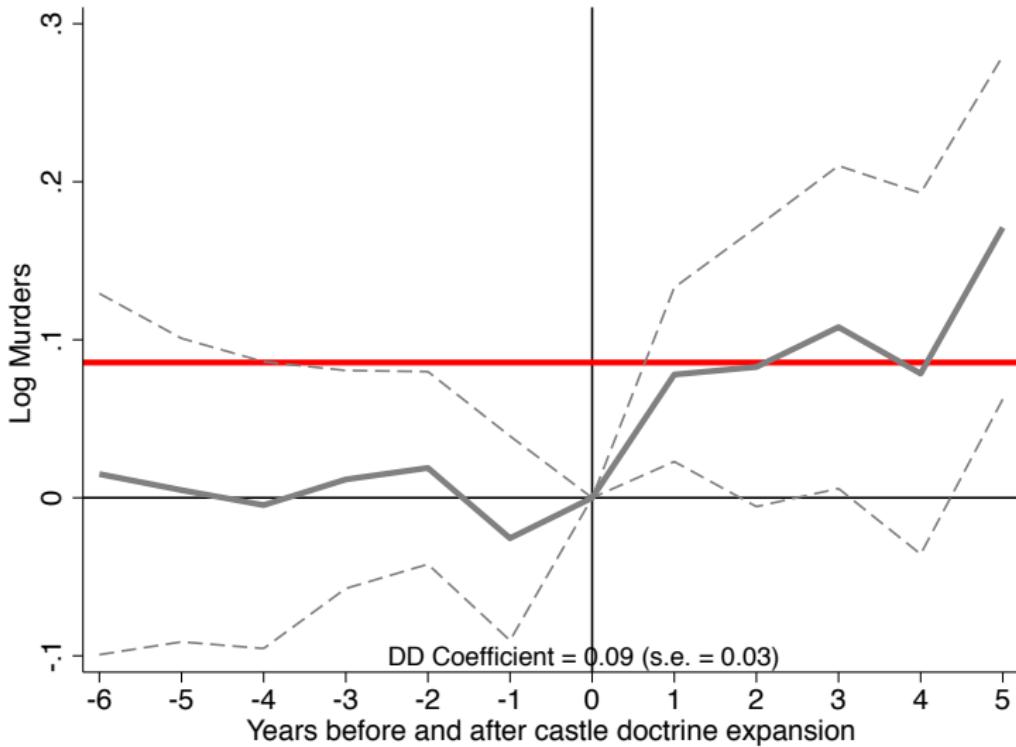


Figure: Homicide event study plots using two-way dropping imbalanced states

Interpretation

- No evidence that Castle Doctrine/Stand Your Ground Laws deter violent crimes such as burglary, robbery and aggravated assault
- These laws do lead to an 8% net increase in homicide rates, translating to around 600 additional homicides *per year* across the 21 adopting states
- Unlikely that all of the additional homicides were legally justified
- Incentives matter in some contexts (lethal force) but not others (deterrence)

Where to from here?

- Now that we've reviewed the two-way fixed effects with treatment that differed across time, how does this more general form of "differential timing" compare with the 2x2 DD that we reviewed?
- Complicated derivation, but simple interpretation - two-way fixed effects with differential timing estimates a weighted average of all 2x2
- Andrew Goodman-Bacon (2018; 2019) and Callaway and Sant'ann (2019)
- I will be making the argument that under certain *modal* situations, the two-way fixed effects model has major problems, even fatal ones, due to biases even when parallel trends plausibly holds

Difference-in-Differences with Variation in Treatment Timing

Andrew Goodman-Bacon

NBER Working Paper No. 25018

Issued in September 2018

NBER Program(s):Children, Development of the American Economy, Labor Studies, Public Economics

The canonical difference-in-differences (DD) model contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper derives an expression for this general DD estimator, and shows that it is a weighted average of all possible two-group/two-period DD estimators in the data. This result provides detailed guidance about how to use regression DD in practice. I define the DD estimand and show how it averages treatment effect heterogeneity and that it is biased when effects change over time. I propose a new balance test derived from a unified definition of common trends. I show how to decompose the difference between two specifications, and I apply it to models that drop untreated units, weight, disaggregate time fixed effects, control for unit-specific time trends, or exploit a third difference.



Reminder of 2x2 DD

To understand differential timing, we need to remind ourselves 2x2 form

$$\widehat{\delta}_{kU}^{2\times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Post to pre difference for treatment group compared to the post to pre difference for *never* treated

Different treatment dates by panel unit

$$y_{it} = \underbrace{\beta D_i + \tau Post_t + \delta(D_i \times Post_t) + X_{it}}_{2x2 DD} + \alpha_i + \alpha_t + \varepsilon_{it}$$
$$y_{it} = \underbrace{\delta D_{it} + X_{it}}_{\text{Two way FE}} + \alpha_i + \alpha_t + \epsilon_{it}$$

We know a lot about 2x2, but about the twoway fixed effects estimator when it comes to DD designs

Decomposition Preview

- Linear panel models estimate a treatment parameter that is a weighted average over all 2×2 in your sample
- The estimator is a weighted average of all potential $\delta^{2 \times 2}$ in which treated units act as both controls and treatment depending on the situation
- Weights are function of sample sizes of each “group” and the variance of the treatment dummies for the groups

Decomposition (cont.)

- Under the assumptions of variance weighted common trends (VWCT) and time invariant treatment effects, the estimator called the variance weighted ATT is a weighted average of all possible ATTs
- Under more restrictive assumptions it perfectly matches the ATT
- Time varying treatment effects generate a bias that needs to be accounted for

3 Group Example

- Suppose two treatment groups (k, l) and one untreated group (u)
- k, l define the groups based on when they receive treatment (differently in time) with k receiving it later than l
- Denote \bar{D}_k as the share of time each group spends in treatment status
- Denote $\hat{\delta}_{ab}^{2 \times 2, j}$ as the canonical 2×2 DD estimator for groups a and b where j is the treatment group
- So what are the possible 2×2 combinations?

How many 2x2?

- A lot!
- When there's three groups - a never treated (U), an early treated (k) and a late treated (l), there are four 2x2s
- But typically, we have more than 3 groups making the number of potential 2x2 even larger
- With K timing groups and one untreated group, there are K^2 distinct 2x2 DDs

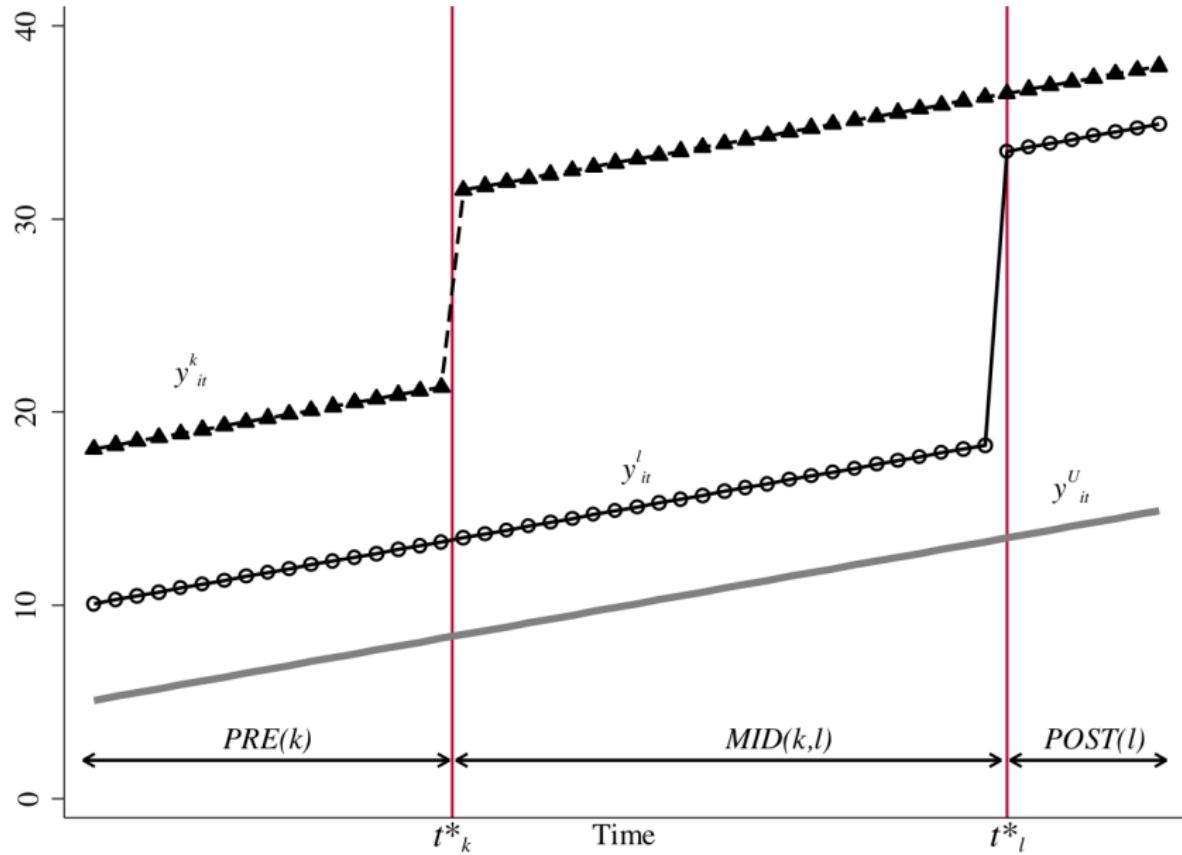
K^2 distinct DDs

Assume 3 timing groups (a, b and c) and one untreated group (U).
Then there should be 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

Simple example with 3 groups

- We'll stick with two groups, k and l, who will get the treatment at t_k^* and t_l^* , and the third group U will never get treated
- The earlier period before anyone is treated is “pre”, the period between k and l treatment is “mid”, and the period after l is treated is “post”



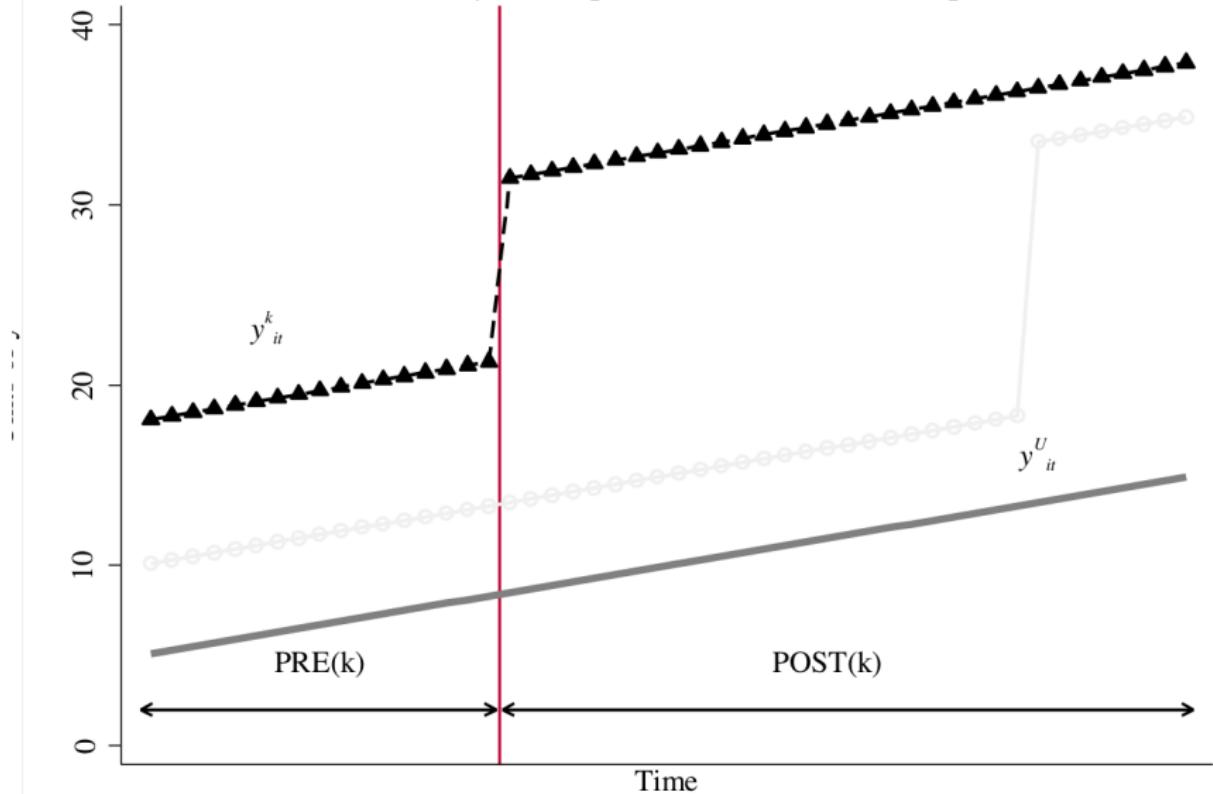
Three important 2x2 DDs

$$\begin{aligned}\hat{\delta}_{kU}^{2\times 2} &= \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right) \\ \hat{\delta}_{kl}^{2\times 2} &= \left(\bar{y}_k^{mid(k,l)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_l^{mid(k,l)} - \bar{y}_l^{pre(k)} \right) \\ \hat{\delta}_{lk}^{2\times 2} &= \left(\bar{y}_l^{post(l)} - \bar{y}_l^{mid(k,l)} \right) - \left(\bar{y}_k^{post(l)} - \bar{y}_k^{mid(k,l)} \right)\end{aligned}$$

where the first 2x2 is any timing group compared to untreated, the second is a group compared to yet-to-be-treated timing group, and the last is the eventually-treated compared to the already-treated controls.

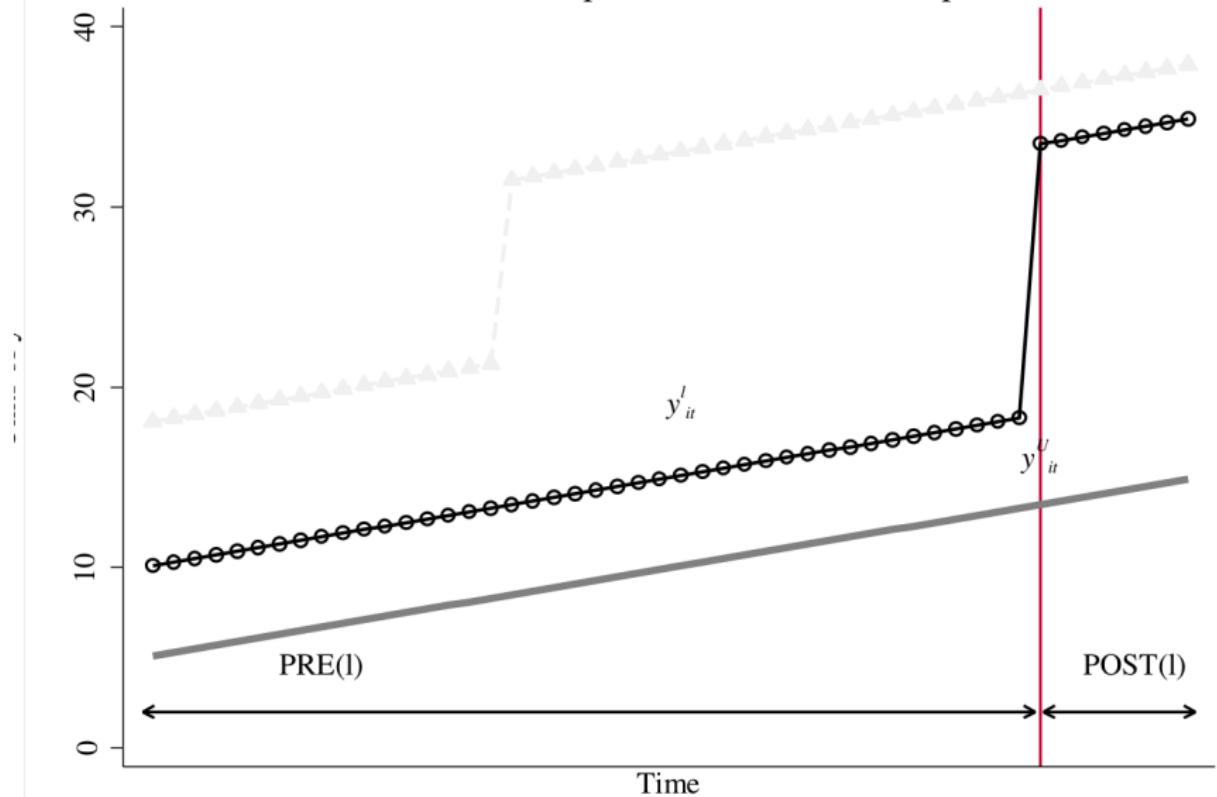
$$\widehat{\delta}_{kU}^{2 \times 2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

A. Early Group vs. Untreated Group

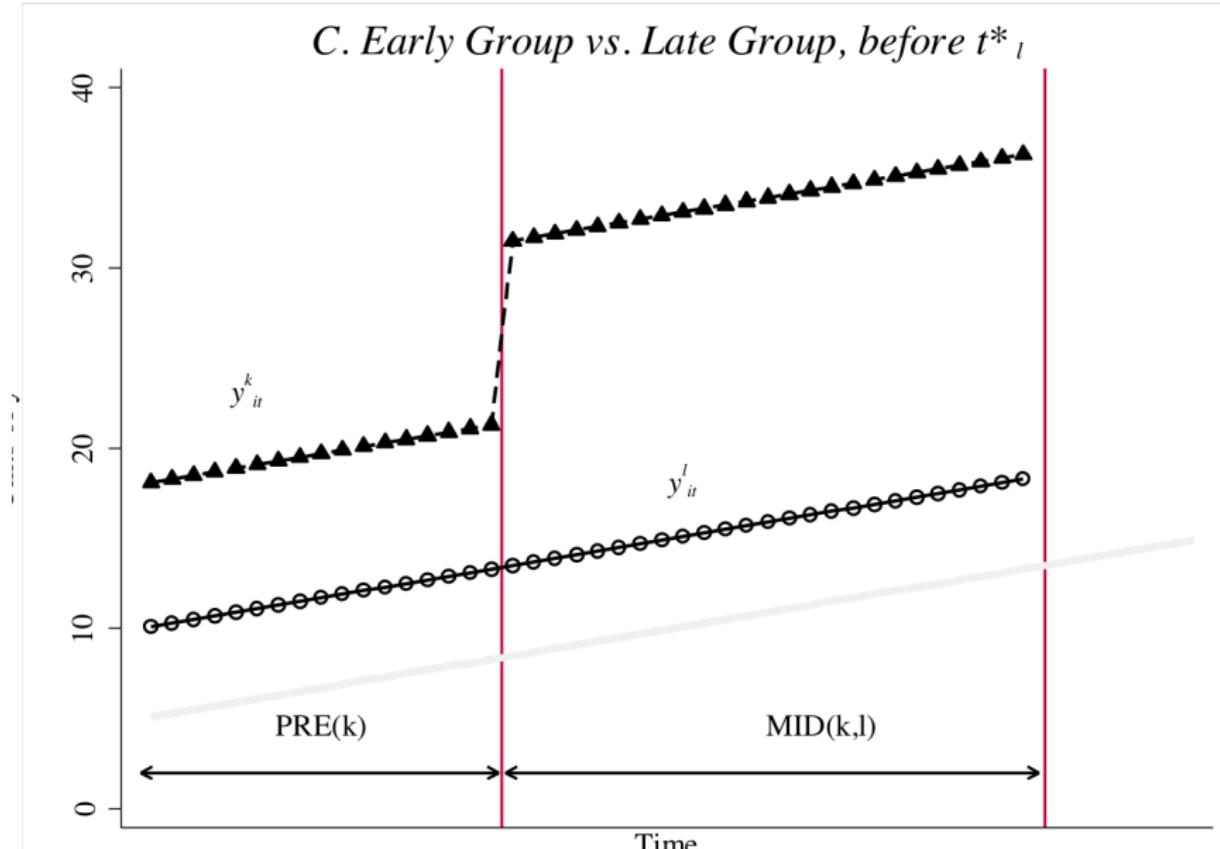


$$\widehat{\delta}_{IU}^{2 \times 2} = \left(\bar{y}_I^{post(I)} - \bar{y}_I^{pre(I)} \right) - \left(\bar{y}_U^{post(I)} - \bar{y}_U^{pre(I)} \right)$$

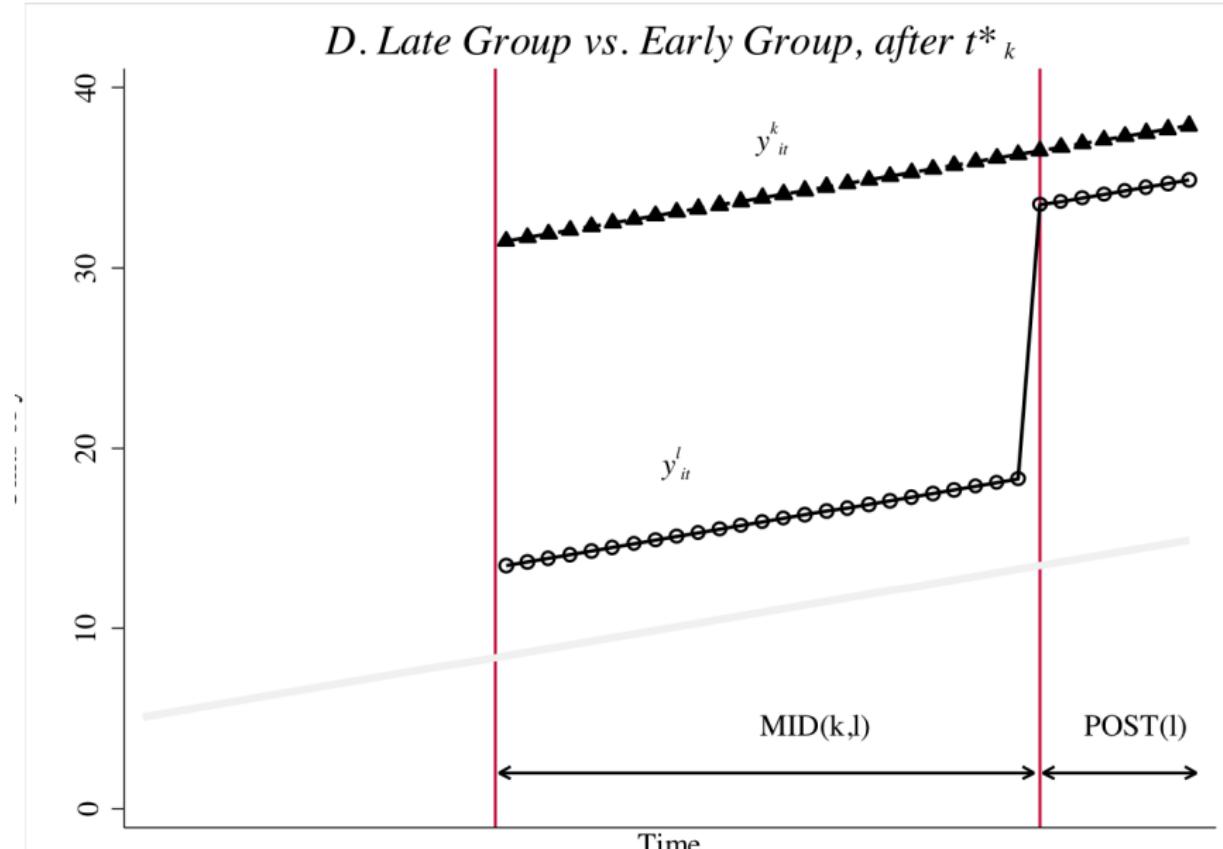
B. Late Group vs. Untreated Group



$$\delta_{kl}^{2 \times 2, k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{Pre(k,l)} \right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k,l)} \right)$$



$$\delta_{lk}^{2 \times 2,I} = \left(\bar{y}_I^{POST(k,l)} - \bar{y}_I^{MID(k,l)} \right) - \left(\bar{y}_k^{POST(k,l)} - \bar{y}_k^{MID(k,l)} \right)$$



Second, what makes up the DD estimator?

The least squares estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

$$\hat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \hat{\delta}_{kU}^{2 \times 2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \hat{\delta}_{kl}^{2 \times 2, k} + (1 - \mu_{kl}) \hat{\delta}_{lk}^{2 \times 2, l} \right]$$

where that first 2x2 is the k compared to U and the l compared to U (combined to make the equation shorter)

Third, the Weights

$$\begin{aligned}s_{ku} &= \frac{n_k n_u \bar{D}_k (1 - \bar{D}_k)}{\widehat{\text{Var}}(\tilde{D}_{it})} \\ s_{kl} &= \frac{n_k n_l (\bar{D}_k - \bar{D}_l) (1 - (\bar{D}_k - \bar{D}_l))}{\widehat{\text{Var}}(\tilde{D}_{it})} \\ \mu_{kl} &= \frac{1 - \bar{D}_k}{1 - (\bar{D}_k - \bar{D}_l)}\end{aligned}$$

where n refer to sample sizes, $\bar{D}_k(1 - \bar{D}_k)$
 $(\bar{D}_k - \bar{D}_l)(1 - (\bar{D}_k - \bar{D}_l))$ expressions refer to variance of
treatment, and the final equation is the same for two timing groups.

Weights discussion

- Two things pop out of these weights
 - “Group” variation matters more than unit-level variation. A group is if two states got treated in 1995. They are the 1995 group. More units in a group, the bigger that 2×2 is practically
 - Within-group *treatment* variance matters a lot.
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
 - ① $\bar{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
 - ② $\bar{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
 - ③ $\bar{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
- What's this mean? The weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the “treated on treated” weights? What’s this $\bar{D}_k - \bar{D}_I$ business about?
- Well, same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_I^* = 0.67$. Then $\bar{D}_k - \bar{D}_I = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

TWFE and centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- But when looking at treated to treated comparisons, when differences in timing have a spacing of around 1/2, those also weight up the respective 2s2s via variance weighting
- But there's no theoretical reason why should prefer this as it's just a weighting procedure being determined by how we drew the panel
- This is the first thing about TWFE that should give us pause, as not all estimators do this

Potential outcomes

- Previous just showed that DD was based on a weighted “adding up” of particular 2x2s. That tells us what DD is numerically. But that’s not the end
- Because the decomposition theorem expresses the DD coefficient in terms of sample averages, the movement to potential outcomes is easy.
- Now we express DD in terms of ATT which is essential for understanding identification and bias

Average treatment effect on the treatment group (ATT)

- Define the year-specific ATT as

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 | k, t = \tau]$$

- Now define it over a time window W (e.g., a post-treatment window)

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 | k, \tau \in W]$$

- Define differences in average potential outcomes over time as:

$$\Delta Y_k^h(W_1, W_0) = E[Y_{it}^h | k, W_1] - E[Y_{it}^h | k, W_0]$$

for $h = 0$ (i.e., Y^0) or $h = 1$ (i.e., Y^1)

Changing potential outcomes

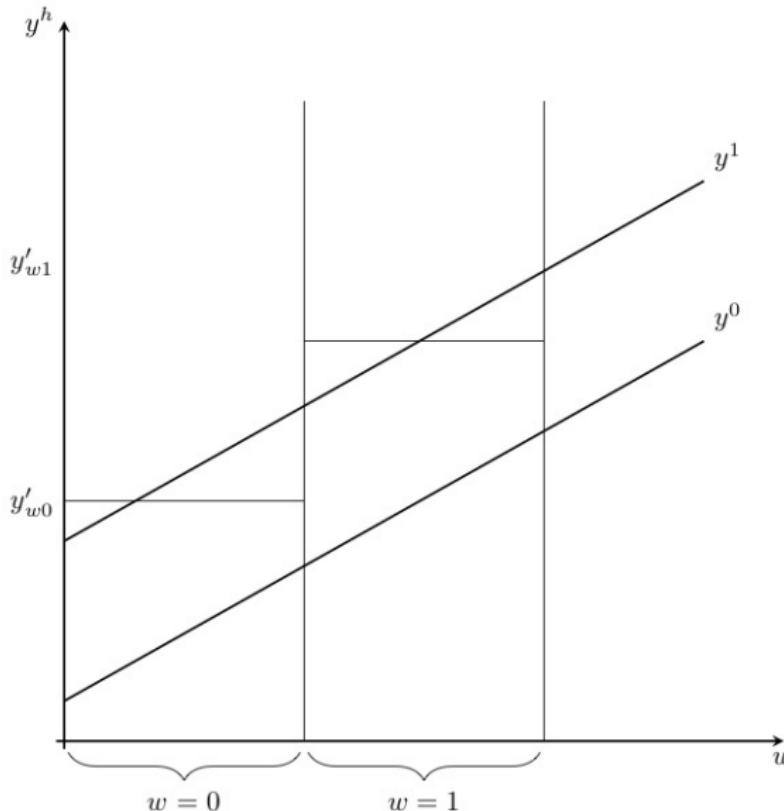


Figure: With trends, differences in mean potential outcomes is non-zero

From 2x2 to ATT

$$\begin{aligned}
 \widehat{\delta}_{kU}^{2x2} &= \left(E[Y_j|Post] - E[Y_j|Pre] \right) - \left(E[Y_u|Post] - E[Y_u|Pre] \right) \\
 &= \underbrace{\left(E[Y_j^1|Post] - E[Y_j^0|Pre] \right) - \left(E[Y_u^0|Post] - E[Y_u^0|Pre] \right)}_{\text{Switching equation}} \\
 &\quad + \underbrace{E[Y_j^0|Post] - E[Y_j^0|Post]}_{\text{Adding zero}} \\
 &= \underbrace{E[Y_j^1|Post] - E[Y_j^0|Post]}_{\text{ATT}} \\
 &\quad + \underbrace{\left[E[Y_j^0|Post] - E[Y_j^0|Pre] \right] - \left[E[Y_u^0|Post] - E[Y_u^0|Pre] \right]}_{\text{Non-parallel trends bias in 2x2 case}}
 \end{aligned}$$

Potential outcomes

$$\widehat{\delta}_{kU}^{2\times 2} = ATT_{Post,j} + \underbrace{\Delta Y_{Post,Pre,j}^0 - \Delta Y_{Post,Pre,U}^0}_{\text{Selection bias!}}$$

Hah! It's that another selection bias term, like when we decomposed the simple difference in outcomes! But here we see it's basis - non-parallel trends in potential outcomes themselves. Notice one of these is counterfactuals, but which one?

Two benign 2x2

$$\begin{aligned}\widehat{\delta}_{kU}^{2\times 2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\ \widehat{\delta}_{kl}^{2\times 2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)\end{aligned}$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated yet).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions like we did we get:

$$\widehat{\delta}_{lk}^{2\times 2} = ATT_{I, Post(I)} + \underbrace{\Delta Y_I^0(Post(I), MID) - \Delta Y_k^0(Post(I), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

Heterogeneity bias?

That old decomposition of the simple difference in outcomes rears its ugly head!

$$\begin{aligned}\widehat{\delta}_{kl}^{2 \times 2} &= ATT_{I, Post(I)} \\ &\quad + \Delta Y_I^0(Post(I), MID) - \Delta Y_k^0(Post(I), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

- The first part is the ATT we are looking for
- The selection bias which only zeroes out if Y^0 for k and I has the same parallel trends from mid to post period
- The heterogeneity bias (3) occurs if the ATT for k differs over time. If not, then it just zeroes out.

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2 \times 2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2 \times 2, k} + (1 - \mu_{kl}) \widehat{\delta}_{kl}^{2 \times 2, l} \right]$$

where we will make these substitutions

$$\begin{aligned}\widehat{\delta}_{kU}^{2 \times 2} &= ATT_k(Post) + \Delta Y_I^0(Post, Pre) - \Delta Y_U^0(Post, Pre) \\ \widehat{\delta}_{kl}^{2 \times 2, k} &= ATT_k(Mid) + \Delta Y_I^0(Mid, Pre) - \Delta Y_I^0(Mid, Pre) \\ \widehat{\delta}_{lk}^{2 \times 2, l} &= ATT_l(Post(l)) + \Delta Y_I^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

$$\begin{aligned} p \lim_{n \rightarrow \infty} \widehat{\delta}_{n \rightarrow \infty}^{DD} &= \delta^{DD} \\ &= VWATT + VWCT - \Delta ATT \end{aligned}$$

- Notice the number of assumptions needed even to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Let's look at each of these three parts more closely

Variance weighted ATT

$$\begin{aligned} VWATT &= \sum_{k \neq U} \sigma_{kU} ATT_k(Post(k)) \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} ATT_k(MID) + (1 - \mu_{kl}) ATT_l(POST(l)) \right] \end{aligned}$$

where σ is like s only population terms not samples.

- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant.
- But otherwise, it's basically weighting each of the individual sets of ATT we have been discussing, where weights depend on group size and variance

Variance weighted common trends

- VWCT can be understood as a variance weighted common trends component,
- This is the collection of selection biases we previously wrote out,
- But notice – identification requires *variance weighted* common trends to hold.
- You get this with identical trends, but you don't need identical trends anymore as the weights can make it hold without.
- Huge pain to write out, unfortunately.

Variance weighted common trends

$$\begin{aligned} VWCT &= \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0(Post(k), Pre) - \Delta Y_U^0(Post(k), Pre) \right] \\ &+ \sum_{k \neq U} \sum_{I > k} \sigma_{kI} \left[\mu_{kI} \{ \Delta Y_k^0(Mid, Pre(k)) - \Delta Y_I^0(Mid, Pre(k)) \} \right. \\ &\quad \left. + (1 - \mu_{kI}) \{ \Delta Y_I^0(Post(I), Mid) - \Delta Y_k^0(Post(I), Mid) \} \right] \end{aligned}$$

This is new. But while this is a lot to be equalling zero, it's ironically a *weaker* identifying assumption than we thought bc you don't need identical common trends since the weights can technically correct for unequal trends.

Heterogeneity bias

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \left[ATT_k(Post(l) - ATT_k(Mid)) \right]$$

Now, if the ATT is constant over time, then this difference is zero, but what if the ATT is not constant? Then TWFE is biased, and depending on the dynamics and the VWATT, may even flip signs

Case 1: ATT varies across units but not time

$$p \lim_{n \rightarrow \infty} \widehat{\delta}_{n \rightarrow \infty}^{DD} = VWATT + VWCT$$

because $\Delta ATT = 0$ here. Assume VWCT=0. Then the VWATT equals

$$\begin{aligned} VWATT &= \sum_{k \neq U} ATT_k \left[\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}(1 - \mu_{jk}) + \sum_{j=k+1}^K \sigma_{jk}\mu_{jk} \right] \\ &= \sum_{k \neq U} ATT_k w_k^T \end{aligned}$$

the VWATT weights together group-specific ATTs by a function of sample shares and treatment variance.

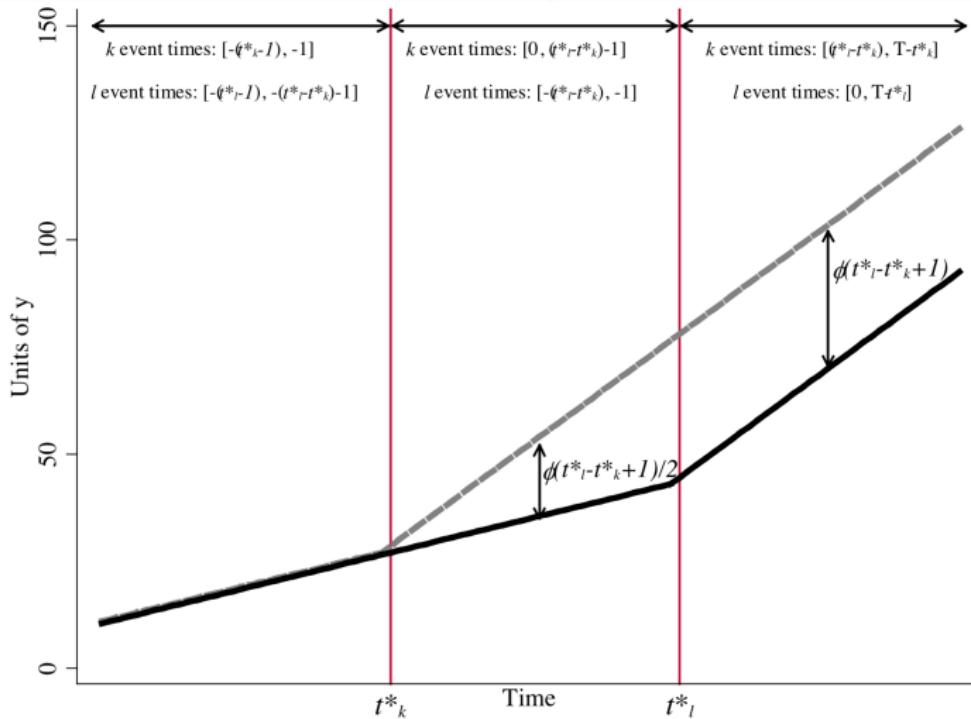
Case 1 cont.

- The processes that determine treatment timing are central to the interpretation of VWATT.
- Assume treatment rolls out first to units with the largest ATTs.
 - Then regression DD underestimates the sample-weighted ATT if t_1^* is early enough, or if there are a lot of post periods, so that \bar{D}_1 very small and $\bar{D}_k \approx 0.5$
 - Regression DD overestimates if t_1^* is late enough (or if there are a lot of pre periods) so that $\bar{D}_1 \approx 0.5$ and \bar{D}_k is small
- Goodman-Bacon (2018) suggests scattering the weights against each group's sample share. They may be close if there is little variation in treatment timing, if the untreated group is very large, or if some timing groups are very large

Case 2: Constant ATT across units, but heterogenous over time

- Time varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows
- Let's consider a case where the counterfactual outcomes are identical, but the treatment effect is a linear break in the trend. For instance, $Y_{it}^1 = Y_{it}^0 + \theta(t - t_1^* + 1)$ similar to Meer and West (2013)

Treatment effect is break in trend



Case 2 cont.

- The first 2x2 uses the later group as its control in the middle period. But in the late period, the later treated unit is using the earlier treated as its control
- But notice, this effect is biased because the control group is experiencing a trend in outcomes (heterogeneous treatment effects)
- This bias feeds through to the later 2x2 according to the size of the weight $(1 - \mu_{kl})$

Variance weighted common trends

- If treatment effects are constant over time, then we only need $VWCT = 0$ to identify VWATT. “Only”!
- The assumption itself is not testable because common trends is based on counterfactual Y^0 for the treatment groups in the post-treatment period, and we only have pre-treatment data
- But let's assume differential counterfactual trends Y_k^0 are linear throughout the panel. Then we can get a convenient approximation to the $VWCT$ on the next slide

Variance weighted common trends

$$\begin{aligned} VWCT &= \sum_{k \neq U} \Delta Y_k^0 \left[\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}(1 - 2\mu_{jk}) + \sum_{j=k+1}^K \sigma_{kj}(2\mu_{kj} - 1) \right] \\ &\quad - \Delta Y_U^0 \sum_{k \neq U} \sigma_{kU} \end{aligned}$$

Obviously, for this bias to be inconsequential, we need the sum of the two weighted counterfactual trends to be zero. You get this with identical trends, but those are not necessary due to the weights ability to shift non-identical trends so as to satisfy the zero condition.

Variance weighted common trends

The weight on each group's counterfactual trend equals the difference between the total weight it gets when it acts as a treatment group (w_k^T) minus the total weight it gets when it acts as a control (w_k^C).

$$\sum_k \Delta Y_k^0 [w_k^T - w_k^C] = 0$$

where w_k^T is the sum of all weights where group k is the treatment group

$$w_k^T = \sigma_{kU} + \sum_{k=1}^{K-1} \sigma_{jk}(1 - \mu_{jk}) + \sum_{j=k+1}^K \sigma_{kj}\mu_{kj}$$

and w_k^C is the sum of all weights where group k is the control group

$$w_k^C = \sum_{k=1}^{K-1} \sigma_{jk}\mu_{jk} + \sum_{j=k+1}^K \sigma_{jk}(1 - \mu_{jk})$$

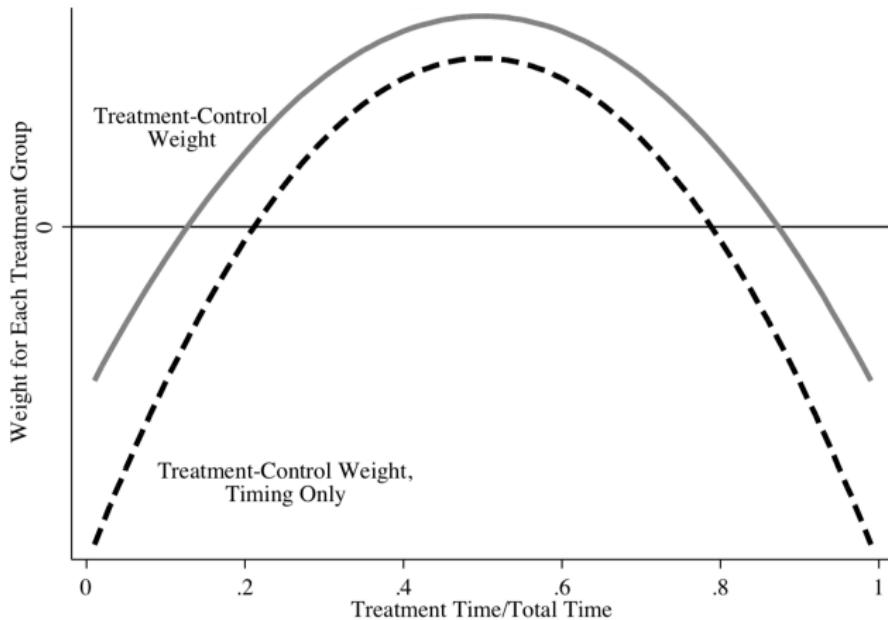
Variance weighted common trends

- The bias induced by each group will depend on whether it is a net treatment/control group
- A positive pre-trend for group j will bias the results upwards if j is a net treatment group ($w_j^T > w_j^C$) or down if its a net control group, and if they are equal, then the bias will be zero regardless of group pre-trend
- Units treated towards the ends of the panel get relatively more weight when they act as controls.
- Needless to say, the size of the bias from a given trend is larger for groups with more weight

Variance weighted common trends

- What this means is that while all units are acting as controls, treatment timing causes some units to be controls more often - hence why they become negative (e.g., $w_k^T - w_k^C < 0$ implies w_k^C has become relatively large)
- The earliest and/or latest units get more weight as controls than treatments
- Units treated in the middle of the panel have high treatment variance as we've noted repeatedly, and so get more weight when they act as the treatment group

Variance weighted common trend weights



Testing VWCT

The identifying assumption $\sum_k \Delta Y_k^0 [w_k^T - w_k^C] = 0$ shows us how to exactly weight averages of x_{it} and perform a single t -test that directly captures the identifying assumption.

- ① Generate a dummy for the effective treatment group

$$1[B_k] = w_k^T - w_k^C > 0$$

- ② Estimate

$$\bar{x}_k = \beta B_k + \varepsilon_k$$

weighted by $|w_k^T - w_k^C|$

The coefficient $\hat{\beta}$ equals covariate differences weighted by the actual identifying variation and its t -statistic tests the null of reweighted balance implied the VWCT equality

Software to check the 2x2s and weights

- Austin Nichols and Thomas Goldring have made available a package in Stata called `ddtiming.ado`
- This will estimate each individual 2x2 and the weights associated with a simple two-way fixed effects model
- Let's look at it. First download Cheng and Hoekstra data from earlier (`castle-doctrine-2000-2010.dta`)
- Now install `ddtiming.ado` and use the do file that I've supplied called `hoekstra-cheng.do`

Stata

```
. use castle-doctrine-2000-2010.dta, replace  
. areg l_murder post i.year, a(sid) robust
```

Dep var	Log homicide
Castle doctrine law	0.105 (0.032)

Recall the estimated ATT is 0.105

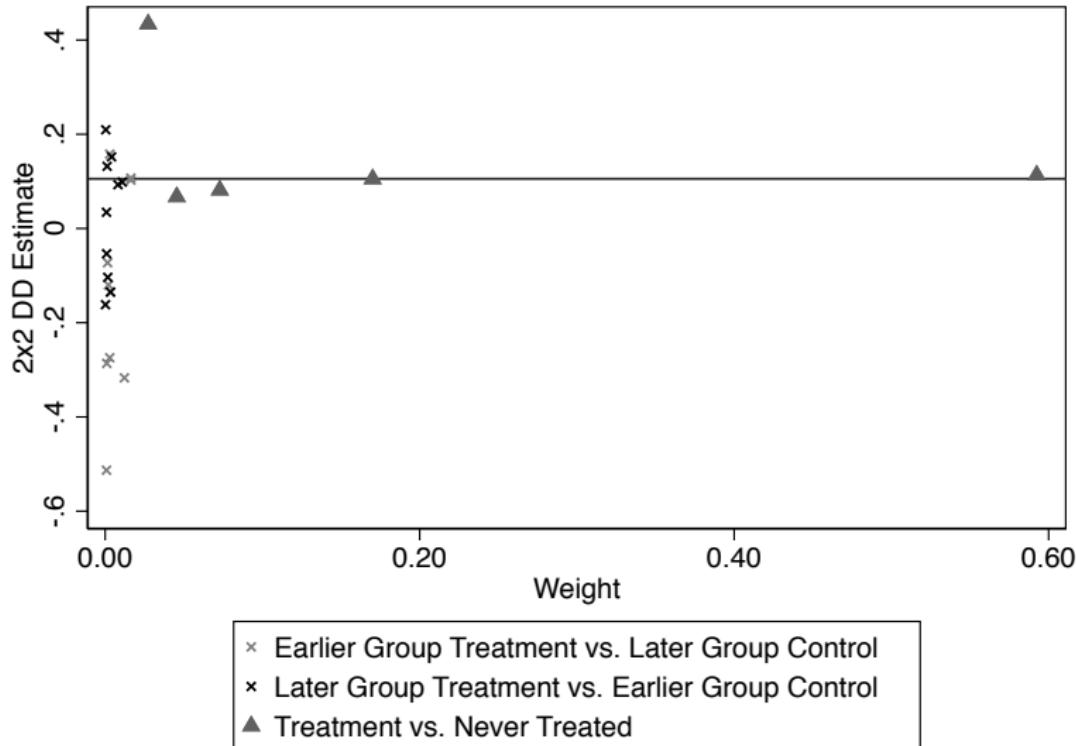
```
. ddtiming l_murder post, i(sid) t(year)
```

DD Comparison	Weight	Avg DD Est
Earlier T vs. Later C	0.060	-0.039
Later T vs. Earlier C	0.032	0.063
T vs. Never treated	0.908	0.116

```
. di (0.060*-0.039) + (0.032*0.063) + (0.908*0.116)  
. 0.105
```

Most of the 0.105 is coming from comparing treatment units to never treated units; the others cancel out

2×2 s and their corresponding weights



Biased DD with OLS

- Review baker.do
- So we see – with differential timing, and heterogeneous treatment effects over time, the TWFE bias can be gigantic because:

$$plim = VWATT + VWCT - \Delta ATT_{Ik}$$

- New papers are coming out focused on the issues that we are seeing with TWFE
- Callaway and Sant'anna (2019) is one of these (currently R&R at Journal of Econometrics)

Preliminary

Callaway and Sant'anna consider identification, estimation and inference procedures for ATE in DD models with

- ① multiple time periods
- ② variation in treatment timing (i.e., differential timing)
- ③ parallel trends only holds after conditioning on observables

Group-time ATE

Key concept: the ATE for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Their method will calculate an ATE per group/time which yields *many* individual ATE estimates
- Group-time ATE estimates are not determined by the estimation method one adopts (first difference or FE)
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Provides a way to aggregate over these to get a single ATE

Another contribution

- Typical econometrics paper: they propose estimators and provide asymptotically valid inference procedures for the causal parameter of interest
 - Uses a particularly kind of bootstrapping that is computationally convenient to obtain confidence intervals
- This is an extension of an older Abadie (2006) paper on semi-parametric DD with some subtle and substantive differences
- The estimator will look awfully similar to an inverse probability weighting estimator down to the use of propensity scores

Parallel trends assumption

- Parallel trends is *never* directly testable
- If you assume though that it holds in the pre-treatment period that therefore it holds in the counterfactual periods, then fine
- (IMO, this begs the question [as in assumes the conclusion]. Obviously if treatment is endogenous then parallel trends doesn't hold even if it did hold prior (see Kahn-Lang and Lang 2018))

Notation

- T periods going from $t = 1, \dots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t .
- C is also binary and indicates a control group unit equalling one if “never treated”
 - Recall the problem with OLS on using treatment units as controls
 - Callaway and Sant’anna seem to know this and working to specifically address it by essentially not using those units at all as controls
- Generalized propensity score:
 $p(\hat{X}) = Pr(G_g = 1|X, G_c + C = 1)$

Propensity scores

- They'll estimate a propensity score based on group covariates using probit or logit (but not OLS)
- That score will then be normalized (e.g., Hajek weight) which improves finite sample bias
- You may need to trim it on the [0.1,0.9] interval as is commonly suggested in other applications
- Essentially, units in control group will be weighted up if their propensity scores are high, and weighted down if low, making more apple-to-apples comparisons

Detour into IPW

Horvitz weights

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

Harjek weights

$$\widehat{\delta}_{ATT} = \left[\sum_{i=1}^N \frac{Y_i D_i}{\widehat{p}} \right] / \left[\sum_{i=1}^N \frac{D_i}{\widehat{p}} \right] - \left[\sum_{i=1}^N \frac{Y_i (1 - D_i)}{(1 - \widehat{p})} \right] / \left[\sum_{i=1}^N \frac{(1 - D_i)}{(1 - \widehat{p})} \right]$$

Parameter of interest

$$ATT(g, t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

Potential uses of this estimator

- ① Are treatment effects heterogenous by time of adoption?
- ② Does treatment effect change over time?
- ③ Are shortrun effects more pronounced than longrun effects?
- ④ Do treatment effect dynamics differ if people are first treated in a recession relative to expansion years?

Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Estimator

Theorem 1

$$ATT(g, t) = E \left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E \left[\frac{\hat{p}(X)C}{1-\hat{p}(X)} \right]} \right) (Y_t - Y_{g-1}) \right]$$

Which units will and will not be controls?

- Callaway and Sant'anna are keeping us from calculating DD's using TWFE, which is problematic in part bc you're implicitly calculating 2x2s by comparing later treated units to early treated units, which is a sin
- But what if you never have a true control group, or "never treated"?

Remarks about “staggered adoption” with universal coverage

Proof.

Remark 1: In some applications, eventually all units are treated, implying that C is never equal to one. In such cases one can consider the “not yet treated” ($D_t = 0$) as a control group instead of the “never treated?” ($C = 1$). □

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”

“We can aggregate the group-time treatment effects into fewer interpretable causal effect parameters, which makes interpretation easier, and also increases statistical power and reduces estimation uncertainty.” - Andrew Baker

Interesting Parameter 1

$$\frac{2}{T(T-1)} \sum_{g=2}^T \sum_{t=2}^T \mathbf{1}\{g \leq t\} ATT(g, t)$$

where T is number of pre-treatment years (Assumption 2 regarding conditional parallel trends). Let's look at an example.

Aggregating the first way

$$ATT(1986, 1986) = 10$$

$$ATT(1986, 1987) = 15$$

$$ATT(1986, 1988) = 20$$

Let data run from 1983 - 1988. Thus $T = 3$. ATT simple average is 15.

Interesting Parameter 2

$$\frac{1}{k} \sum_{g=2}^T \sum_{t=2}^T \mathbf{1}\{g \leq t\} ATT(g, t) P(G = g)$$

This is a weighted average of each $ATT(g, t)$ putting more weight on $ATT(g, t)$ with larger group sizes

Bootstrap inference

They propose a bootstrap procedure to conduct asymptotically valid inference which can adjust for autocorrelation and clustering

Coding example

See `baker.do` to illustrate basics of CS, and `castle_cs.R` to see CS in action

Comparing TWFE and CS for Cheng and Hoekstra

- Results are similar for the two estimators
- Event study plots seem similar as well
- Likely because of the large pool of never treated units which get a weight over 0.9 in TWFE
- Very little treatment effect heterogeneity – except for last period, group ATT is similar across all groups
- Be more concerned with most or all units are eventually treated by end of sample, as then you can't have the never treated comparisons

Stacking

- TWFE seems like it should identify ATT with staggered adoption since it does in simple 2x2
- Several papers show that this is not true with heterogeneity
- Several authors have shown that TWFE identifies some weighted average of group and time-specific ATT but the weights can be negative and non-interpretable

Alternatives – aggregation

- Alternatives we've examined so far are SA and CS
- Both estimate group-time ATT (or cohort-time ATT) – many parameters
- These can then be aggregated into whatever parameter you're interested in
- Easily implementable in available R or Stata software

Alternatives – stacking

- A separate stream went a different route
- Cengiz, et al. (2019) is a minimum wage study that used a “stacking” method
- Intuition is to transform the staggered adoption setting to a two-group two-period design (Gardner 2020)
- Done by creating many different datasets centering each treatment and control group on the same relative event time

Method

- For each treatment group, create a new dataset spanning a periods before and b periods after treatment adoption
- This dataset will consist of observations on the treatment group and the group of units that never receive the treatment
- So long as there are treated and untreated observations for each group group and period, you can do this
- Then stack these group-specific datasets and regress the outcomes onto treatment dummy and group-specific dummies and period dummies

Differential timing complicates plotting sample averages

- New Jersey treated in late 1992, New York in late 1993, Pennsylvania never treated
- Pre-treatment:
 - New Jersey: <1992
 - New York: <1993
 - Pennsylvania: undefined
- So how do we check parallel leads?

Early efforts at event studies

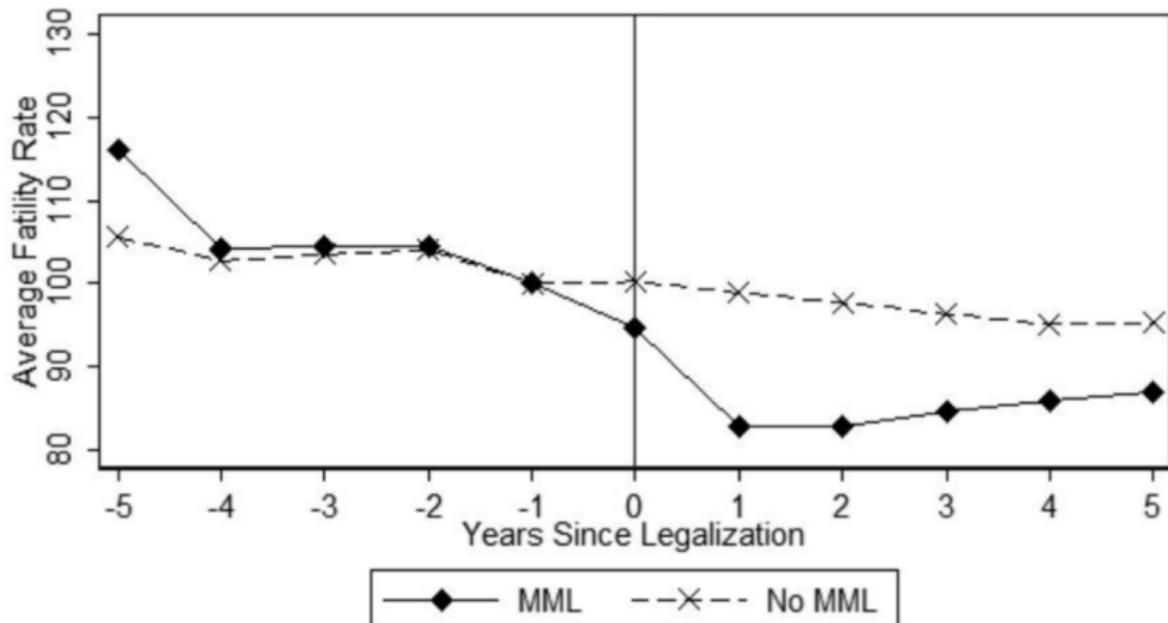


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Randomized control counties to receive arbitrary dates as treatment can be misleading

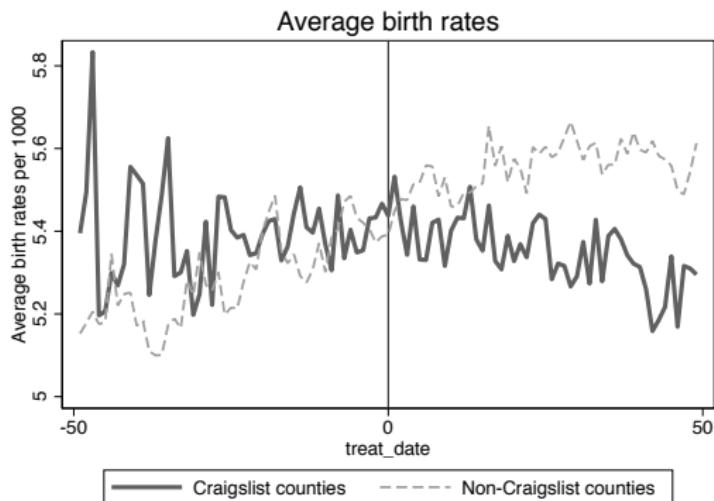
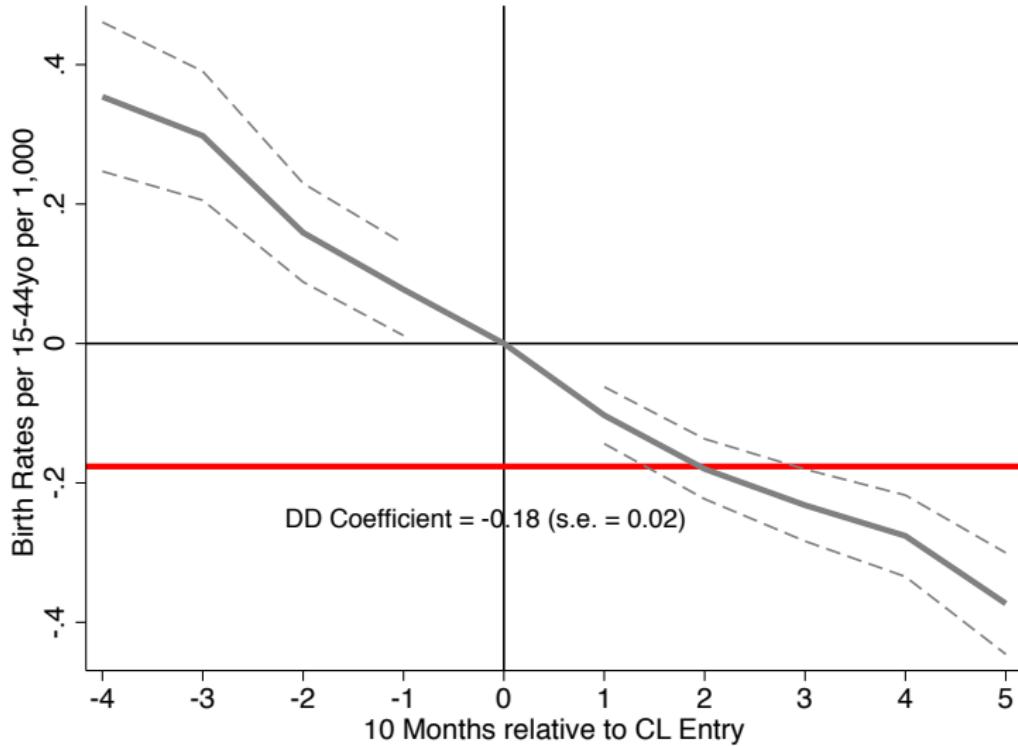


Figure: From one of my studies. Looks decent right?



Same data as a couple slides ago, leads don't look good

Revisiting the event study

- Recall our discussion of event studies estimated with TWFE under differential timing
- Now that we know about the biases of TWFE when estimating aggregate DD parameters, let's revisit event studies under differential timing
- Callaway and Sant'Anna (2020) propose alternative estimators for event studies that estimate group-time ATT in relative event time
- But now we will discuss Sun and Abraham (2020) [SA] which is like a blend of Goodman-Bacon's decomposition and Callaway and Sant'anna alternative estimator to TWFE

Summarizing

- Goodman-Bacon (2019) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

Summarizing (cont.)

- ① SA is a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
- ② They show that the population regression coefficient is “contaminated” by information from other leads and lags
- ③ SA presents an alternative estimator that is not so dissimilar to CS

Summarizing (cont.)

- Problems seem to occur with DD when we introduce treatment effect heterogeneity
- Under treatment effect heterogeneity, spurious non-zero positive lead coefficients even when there is no pretrend
- This problem is exacerbated by the TWFE related weights as under some scenarios, the weights sum to zero and “cancel out” the treatment effects from other periods
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Summarizing (cont.)

- Only decomposition of TWFE estimating dynamic leads and lags (Goodman-Bacon focused on a “static” specification)
- Contamination of coefficients on leads and lags by treatment effects depends on the magnitude of the weights on the true group-time ATT, or “cohort-specific ATT”
- Weights are a function of cohort composition
- Examining weights lets you gauge how treatment effect heterogeneity would interact with potential non-zero and non-convex weighting in population regression coefficients on the leads and lags

Difficult notation sadly

- When treatment occurs at the same time, we say they are part of the same cohort, e
- If we bin the data, then a lead or lag l will appear in the bin g so sometimes they use g instead of l or $l \in g$
- Building block is the “cohort-specific ATT” or $CATT_{e,l}$ – same thing as CS group-time ATT
- Estimate $CATT_{e,l}$ with population regression coefficient μ_l

Difficult notation (cont.)

- At each time t there are two possible treatment status $D_{i,t} \in \{0, 1\}$ over $T + 1$ time periods
- Path of treatment status scales exponentially with T and can take on 2^{T+1} possible values
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- If a group is never treated, the ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome:
$$Y_{i,t} - Y_{i,t}^{\infty}$$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as $l = \text{year} - \text{treatment date}$

Definition 1

Definition 1: The cohort-specific ATT / periods from initial treatment date e is:

$$CATT_{e,I} = E[Y_{i,e+I} - Y_{i,e+I}^{\infty} | E_i = e]$$

Identifying assumption 1

Assumption 1: Parallel trends in baseline outcomes:

$E[Y_{i,t}^\infty - Y^\infty + i, s | E_i = e]$ is the same for all $e \in \text{supp}(E_i)$ and for all s, t and is equal to $E[Y_{i,t}^\infty - Y_{i,s}^\infty]$

Interesting SA comment: Never-treated units are likely to differ from ever-treated units in many ways; think of a Roy model. What does it imply that they chose not to get treated? It may imply net negative treatment effects and that could mean they may not share the same evolution of baseline outcomes as the treatment groups. If you think they are unlikely to satisfy this assumption, then drop them. Almost like a synthetic control approach.

Assumption 2

Assumption 2: No anticipator behavior in pre-treatment periods: There is a set of pre-treatment periods such that $E[Y_{i,e+I}^e - Y_{i,e+I}^\infty | E_i = e] = 0$ for all possible leads.

Basically means that potential outcomes prior to treatment at baseline by on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

Assumption 3

Assumption 3: Treatment effect homogeneity: For each relative time period I , the $CATT_{e,I}$ doesn't depend on the cohort and is equal to $CATT_I$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn't rule out parallel trends

TWFE Regression

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

They say E_i is the initial time of a binary variable absorbing treatment for unit i . Fixed effects should be obvious. μ_g is the population regression coefficient on the leads and lags that we want to estimate. We estimate this using OLS and get $\widehat{\mu}_g$.

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

Specifying the leads and lags

How will we specify the $1\{t - E_i \in g\}$ term? SA considers a couple:

- ① Static specification:

$$Y_{i,t} = \alpha_i + \delta_t + \mu_g \sum_{l \geq 0} D_{i,t}^l + \varepsilon_{i,t}$$

- ② Dynamic specification:

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_l D_{i,t}^l + \sum_{l=0}^L \mu_l D_{i,t}^l + \varepsilon_{i,t}$$

Multicollinearity

Dynamic specification requires deciding which leads to drop. They recommend dropping two: $I = -1$ and some other one (they seem to favor $I = -4$). The reason is twofold. You drop one of them to avoid multicollinearity in the relative time indicators. You drop a second one because of the multicollinearity coming from the linear relationship between TWFE and the relative period indicators.

Trimming and binning

- First some terms: trimming and binning, I do both in the Mixtape when analyzing Cheng and Hoekstra (2013)
- Binning means placing all “distant” relative time indicators into a single one. Done because of the sparseness of units in such distant bins. So if there’s 3 distant leads and lags that aren’t balanced, combine them all into the last lead and lag
- Trimming means excluding any relative period for which you don’t have balance in relative time. This creates a balanced panel “in relative time”, but imbalanced panel length overall.
- They’ll analyze both and how they affect $\widehat{\mu}_g$ estimation using TWFE

Interpreting $\widehat{\mu}_g$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$\begin{aligned}\mu_g = & \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Good stuff}} \\ & + \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Bleh - Other included relative time}} \\ & + \underbrace{\sum_{l \in g^{excl}} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{More bleh - Excluded}}\end{aligned}$$

Superscript g associates the weight with coefficient μ_g . The weight associated with cohort e in relative period l is equal to the population regression coefficient on the $1\{t - E_i \in g\}$ from regression $D_{i,t}^l \times 1\{E_i = e\}$ on all bin indicators included in the regression and TWFE. Just the mechanics of double demeaning from TWFE

Weight ($w_{e,I}^g$) summation cheat sheet

- ① For relative periods of μ_g own $I \in g$, $\sum_{I \in g} \sum_e w_{e,I}^g = 1$
- ② For relative periods belonging to some other bin $I \in g'$ and $g' \neq g$, $\sum_{I \in g'} \sum_e w_{e,I}^g = 0$
- ③ For relative periods not included in G , $\sum_{I \in g^{excl}} \sum_e w_{e,I}^g = -1$

Estimating the weights

Regress $D_{i,t}^I \times 1\{E_i = e\}$ on:

- ① all bin indicators included in the main TWFE regression,
- ② $\{1\{t - E_i \in g\}\}_{g \in G}$ (i.e., leads and lags) and
- ③ the unit and time fixed effects

Interpretation of coefficients under parallel trends only

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bin g is a linear combination of $CATT_{e,I \in g}$ as well as $CATT_{d,I'}$ from other relative periods $I' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g = & \underbrace{\sum_{I \in g} \sum_e w_{e,I}^g CATT_{e,I}}_{\text{Desirable}} \\ & + \underbrace{\sum_{g' \neq g, g' \in G} \sum_{I' \in g'} \sum_e w_{e,I'}^g CATT_{e,I'}}_{\text{Undesirable - other specified bins}} \\ & + \underbrace{\sum_{I' \in g^{excl}} \sum_e w_{e,I'}^g CATT_{e,I'}}_{\text{Undesirable - excluded relative time indicators}}\end{aligned}$$

Comment on Proposition 2

The coefficient μ_g can be written as an average of $CATT_{e,I}$ from own periods but also $CATT_{e,I'}$ from other periods.

The weights are still functions of cohort comparisons, like in Proposition 1, which means μ_g can be written as non-convex averages of not only $CATT_{e,I}$ from own periods $I \in g$, but also $CATT_{e,I'}$ from other periods.

Means μ_g could in fact be the wrong sign to all $CATT_{e,I \in g}$.

Weights can help us gauge the severity of this problem.

When the weights have larger magnitude, treatment effect heterogeneity matters more as a particular $CATT_{e,I}$ can drive the overall estimates. But when weights are uniform, treatment effect heterogeneity matters less.

Interpretation under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$\begin{aligned}\mu_g = & \sum_{l' \in g, l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{l' \in g^{excl}, l' \geq 0} \sum_e w_{w,l'}^g CATT_{e,l'}\end{aligned}$$

Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods even though parallel trends hold in the pre period.

Proposition 4

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,I} = ATT_I$ is constant across e for a given I , and the population regression coefficient μ_g is equal to a linear combination of $ATT_{I \in g}$, as well as $ATT_{I' \notin g}$ from other relative periods

$$\begin{aligned}\mu_g &= \sum_{I \in g} w_I^g ATT_I \\ &+ \sum_{g' \neq g} \sum_{I' \in g'} w_{I'}^{g'} ATT_{I'} \\ &+ \sum_{I' \in g^{excl}} w_{I'}^{g'} ATT_{I'}\end{aligned}$$

Proposition 4 comment

The weight $w_I^g = \sum_e w_{e,I}^g$ sums over the weights $w_{e,I}^g$ from Proposition 1 and is equal to the population regression coefficient from the following auxiliary regression:

$$D'_{i,t} = \alpha_i + \lambda_t + \sum_{g \in G} w_I^g \cdot 1\{t - E_i \in g\} + u_{i,t}$$

which regresses $D'_{i,t}$ on all bin indicators and TWFE

On binning

- Many propose either binning or trimming to create “balanced” panels (in relative event time)
- But SA notes that binning in simulations creates uninterpretable weights (due to the binned $CATT_{e,I'}$ inclusion in μ_g), whereas trimming creates weights that are more reasonable
- This may be because trimming subtracts the corresponding $CATT_{e,I'}$ from μ regression coefficient

Intuition for contamination

- Stupid notation make Hulk smash!
- Let's do a simple toy example instead

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

Toy example

$$\begin{aligned}\mu_{-2} = & \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}} \\ & + \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}\end{aligned}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

Interaction-weighted estimator

- They propose an interacted weighted estimator (IW) as a consistent estimator for μ_g
- Estimator uses either never-treated as controls or “last cohort treated” if no never-treated (contra CS which uses “not yet treated”)
- No covariates bc this is a regression with fixed effects and time-varying covariates create own biases, although they note you can plug in CS for the DD calculation and recover *CATT* that way
- The interaction is a TWFE regression specification that interacts relative period indicators with cohort/group indicators, excluding indicators for never-treated cohorts

Interaction-weighted estimator

- **Step one:** Do this DD regression and hold on to $\hat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated.
The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

Interaction-weighted estimator

- **Step two:** Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

IW estimator

- **Step three:** Take a weighted average of estimates for $CATT_{e,I}$ from Step 1 with weight estimates from step 2

$$\hat{v}_g = \frac{1}{|g|} \sum_{I \in g} \sum_e \hat{\delta}_{e,I} \widehat{Pr}\{E_i = e | E_i \in [-I, T - I]\}$$

Consistency and Inference

- Under parallel trends and no anticipation, $\hat{\delta}_{e,I}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IV estimator is consistent for a weighted average of $CATT_{e,I}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

DD Estimator of CATT

Definition 2: DD estimator with pre-period s and control cohorts C estimates $CATT_{e,I}$ as:

$$\widehat{\delta}_{e,I} = \frac{E_N[(Y_{i,e+I} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+I} \times 1\{E_i \in C\})]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,I}$.

Software

Use staggered from

<https://github.com/jonathandroth/staggered> by Jon Roth
(Brown University). There is also a Stata wraparound using the
rcall package in Stata. See instructions on the URL above.

Conclusion of SA

- Bacon shows the TWFE coefficient on the static parameter is “contaminated” by other periods leads and lags
- Three strong assumptions needed for TWFE to be unbiased: parallel trends, no anticipation, and treatment homogeneity
- Three step interaction-weighted estimator is an alternative
- Doesn’t restrict to treatment profile homogeneity
- Callaway and Sant’Anna (2020) and Sun and Abraham (2020) use different controls, but under certain situations (no covariates, never treated) they are the same (“nested”)
- Software in R and Stata exist

Sharp DD

- In a “sharp” DD, a group gets treated in period 1, a control group does not
- Parallel trends allows you to identify ATT
- We discussed several methods
- But sometimes the lines between treatment and control groups get “fuzzy”

Fuzziness

- In a “fuzzy” DD design, there’s growth in treatment occurring naturally in the control group
 - They discuss an early 2000s Duflo paper where Indonesia pushed for more primary schooling
 - Used earlier cohorts as controls bc they were already past the age
 - But they saw growth in schools too
- In many applications, the “treatment rate” increase more in some groups than in others but there is no group that goes from fully untreated to fully treated
- But there is no group that also remains fully untreated

Earlier fuzzy estimators

- Popular estimator (10% of AERs from 2010-2012) divides DiD by the DiD of the treatment

$$Wald_{DiD} = \frac{\left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)}{\left(E[D_k|Post] - E[D_k|Pre] \right) - \left(E[D_U|Post] - E[D_U|Pre] \right)}$$

- It's Wald IV in that we scale the reduced form by the first stage but they call it Wald DiD
- de Chaisemartin and D'Haultfoeuille (2017) estimates the LATE for group's who go from untreated to treated

Personal takeaway

- Two main values of this paper that I found:
 - Situations where the control group is getting treated with unrelated policy shocks
 - Continuous treatments
- Code to do it is simple but in Stata

Most basic notation

For any random variable, R, we interpret as R_{dgt} as treatment status, treatment group, time

$$R_{101} \sim R|D = 1, G = 0, T = 1$$

Treatment status (D) is whether a unit is treated regardless of group; Group (G) is treatment or control *groups*; Time (T) is before or after

Cases under consideration

Case 1: Share of treated units in control don't change between periods

$$E[D_{01}] = E[D_{00}]$$

Wald_{DiD} identifies the LATE parameter for “switchers” (i.e., people whose treatment status changed between 0 and 1) if parallel trends holds and if the ATE of treated units at both dates is stable over time; proposes new estimators that don't depend on this

Stable ATE isn't required in a typical “sharp” DiD

Cases under consideration

Case 2: Share of treated units changes over time in control

$$E[D_{01}] > E[D_{00}]$$

Wald_{DiD} identifies the LATE of switchers under PT and stable ATE assumption and LATE of treatment and control group switchers are the same

Under certain assumptions, their alternative estimator will only be partially identified, and it depends on the size of the change of treated units in the control.

Concluding remarks on DD

- Chances are you are going to write more papers using DD than any other design
- Goodman-Bacon (2018, 2019) and Sun and Abraham (2020) is *worth your time* because their decompositions show sources of bias in TWFE under reasonable scenarios
- Callaway and Sant'anna (2020) is an extremely useful contribution to the DD toolbox for showing a way to estimate the group-time ATT using any variety of approaches, including regression
- Sun and Abraham (2020) also provides a way forward for event studies (as does CS)

What is synthetic control

- Synthetic control has been called the most important innovation in causal inference of the last 15 years (Athey and Imbens 2017)
- It's extremely useful for case studies, which is nice because that's often all we have
- Continues to also be methodologically a frontier for applied econometrics
- Consider this talk a starting point for you

What is a comparative case study

- Single treated unit – country, state, whatever
- Social scientists tackle such situations in two ways:
qualitatively and quantitatively
- In political science, probably others, you see as a stark dividing
line between camps
- Not so much in economics

Qualitative comparative case studies

- In qualitative comparative case studies, the goal is to reason *inductively* the causal effects of events or characteristics of a single unit on some outcome, oftentimes through logic and historical analysis.
 - May not answer the causal questions at all because there is rarely a counterfactual, or if so, it's ad hoc.
 - Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America (but he is regularly comparing the US to France)

Traditional quantitative comparative case studies

- Quantitative comparative case studies are often explicitly causal designs.
- Usually a natural experiment applied to a single aggregate unit (e.g., city, school, firm, state, country)
- Method compares the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same *ad hoc* aggregate control group (Card 1990; Card and Krueger 1994)

Pros and cons of traditional case study approaches

- Pros:
 - Policy interventions often take place at an aggregate level
 - Aggregate/macro data are often available
- Cons:
 - Selection of control group is *ad hoc*
 - Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest







Description of the Mariel Boatlift

- How do inflows of immigrants affect the wages and employment of natives in local labor markets?
- Card (1990) uses the Mariel Boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- The Mariel Boatlift increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston, Tampa-St. Petersburg)

Why these four?

Tables 3 and 4 present simple averages of wage rates and unemployment rates for whites, blacks, Cubans, and other Hispanics in the Miami labor market between 1979 and 1985. For comparative purposes, I have assembled similar data for whites, blacks, and Hispanics in four other cities: Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. These four cities were selected both because they had relatively large populations of blacks and Hispanics and because they exhibited a pattern of economic growth similar to that in Miami over the late 1970s and early 1980s. A comparison of employment growth rates (based on establishment-level data) suggests that economic conditions were very similar in Miami and the average of the four comparison cities between 1976 and 1984.

Card's main results

Differences-in-differences estimates of the effect of immigration on unemployment^a

Group	Year		
	1979 (1)	1981 (2)	1981–1979 (3)
Whites			
(1) Miami	5.1 (1.1)	3.9 (0.9)	– 1.2 (1.4)
(2) Comparison cities	4.4 (0.3)	4.3 (0.3)	– 0.1 (0.4)
(3) Difference Miami-comparison	0.7 (1.1)	– 0.4 (0.95)	– 1.1 (1.5)
Blacks			
(4) Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5) Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6) Difference Miami-comparison	– 2.0 (1.9)	– 3.0 (2.0)	– 1.0 (2.8)

^a Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Can this ever lead to subjective biases?

- Card found that the Mariel boatlift reduced unemployment *compared to the four cities he chose*
- Is there anything principled we could do that doesn't give the researcher so much control over control group?
- Enter synthetic control (Abadie and Gardeazabal 2003; Abadie, Diamond and Hainmueller 2010)

Synthetic Control

- First appears in Abadie and Gardeazabal (2003) in a study of a terrorist attack in Spain (Basque) on GDP
- Revisited again in a 2011 JASA with Diamond and Hainmueller, two political scientists who were PhD students at Harvard (more proofs and inference)
- A combination of comparison units often does a better job reproducing the characteristics of a treated unit than single comparison unit alone

Researcher's objectives

- Our goal here is to reproduce the counterfactual of a treated unit by finding the combination of untreated units that best resembles the treated unit *before* the intervention in terms of the values of k relevant covariates (predictors of the outcome of interest)
- Method selects *weighted average of all potential comparison units* that best resembles the characteristics of the treated unit(s) - called the “synthetic control”

Synthetic control method: advantages

- Precludes extrapolation (unlike regression) because counterfactual forms a convex hull
- Does not require access to post-treatment outcomes in the “design” phase of the study - no peeking
- Makes explicit the contribution of each comparison unit to the counterfactual
- Formalizing the way comparison units are chosen has direct implications for inference

Synthetic control method: disadvantages

- ① Subjective researcher bias kicked down to the model selection stage
- ② Significant diversity at the moment as to how to principally select models - from machine learning to modifications - as well as estimation and software

Furman and Pinto (2018) recommend showing a few different results in their “cherry picking” JPAM

Synthetic control method: estimation

Suppose that we observe $J + 1$ units in periods $1, 2, \dots, T$

- Unit “one” is exposed to the intervention of interest (that is, “treated”) during periods $T_0 + 1, \dots, T$
- The remaining J are an untreated reservoir of potential controls (a “donor pool”)

Potential outcomes notation

- Let Y_{it}^0 be the outcome that would be observed for unit i at time t in the absence of the intervention
- Let Y_{it}^1 be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T .

Dynamic ATT

Treatment effect parameter is defined as dynamic ATT where

$$\begin{aligned}\delta_{1t} &= Y_{1t}^1 - Y_{1t}^0 \\ &= Y_{1t} - Y_{1t}^0\end{aligned}$$

for each post-treatment period, $t > T_0$ and Y_{1t} is the outcome for unit one at time t . We will estimate Y_{1t}^0 using the J units in the donor pool

Estimating optimal weights

- Let $W = (w_2, \dots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \dots, J + 1$ and $w_2 + \dots + w_{J+1} = 1$. Each value of W represents a potential synthetic control
- Let X_1 be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^* = (w_2^*, \dots, w_{J+1}^*)'$ is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints

Optimal weights differ by another weighting matrix

Abadie, et al. consider

$$||X_1 - X_0 W|| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where X_{jm} is the value of the m -th covariates for unit j and V is some $(k \times k)$ symmetric and positive semidefinite matrix

More on the V matrix

Typically, V is diagonal with main diagonal v_1, \dots, v_k . Then, the synthetic control weights w_2^*, \dots, w_{J+1}^* minimize:

$$\sum_{m=1}^k v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

where v_m is a weight that reflects the relative importance that we assign to the m -th variable when we measure the discrepancy between the treated unit and the synthetic controls

Choice of V is critical

- The synthetic control $W^*(V^*)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment
- Therefore, the V^* weights directly shape W^*

Estimating the V matrix

Choice of v_1, \dots, v_k can be based on

- Assess the predictive power of the covariates using regression
- Subjectively assess the predictive power of each of the covariates, or calibration inspecting how different values for v_1, \dots, v_k affect the discrepancies between the treated unit and the synthetic control
- Minimize mean square prediction error (MSPE) for the pre-treatment period (default):

$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^J w_j^*(V^*) Y_{jt} \right)^2$$

Cross validation

- Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
- For any given V , calculate $W^*(V)$ in the training period.
- Minimize the MSPE of $W^*(V)$ in the validation period

Suppose Y^0 is given by a factor model

What about unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors?

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it}$$

where α_t is an unknown common factor with constant factor loadings across units, and λ_t is a vector of unobserved common factors

With some manipulation

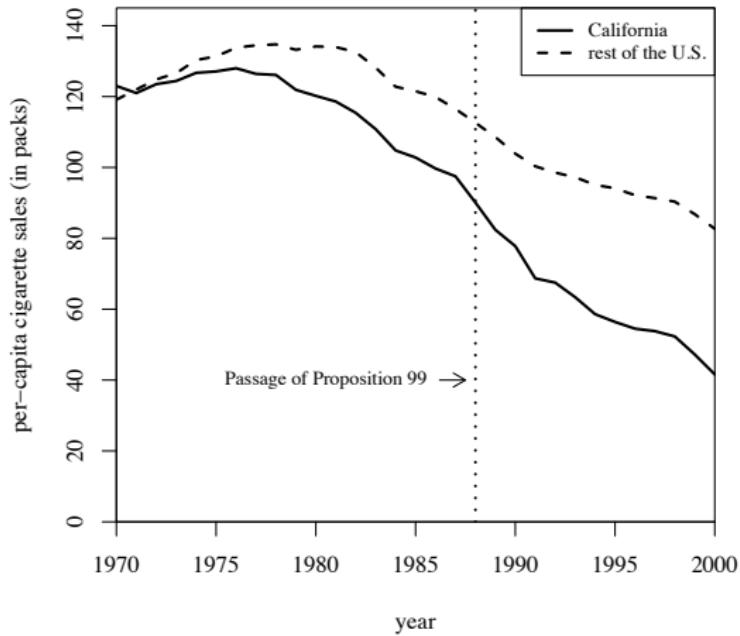
$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda'_n \lambda_n \right)^{-1} \lambda'_s (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

- If $\sum_{t=1}^{T_0} \lambda'_t \lambda_t$ is nonsingular, then RHS will be close to zero if number of preintervention periods is “large” relative to size of transitory shocks
- Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time
- Proof in Appendix B of ADH (2011)

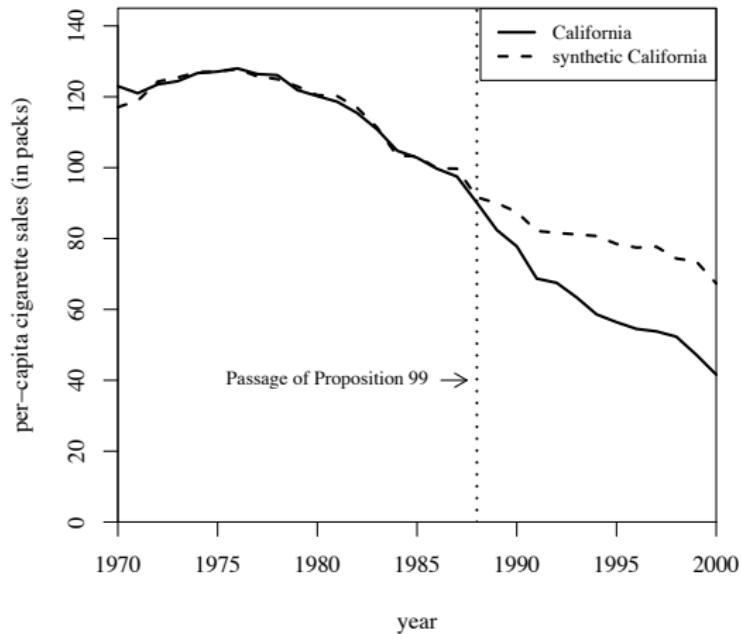
Example: California's Proposition 99

- In 1988, California first passed comprehensive tobacco control legislation:
 - increased cigarette tax by 25 cents/pack
 - earmarked tax revenues to health and anti-smoking budgets
 - funded anti-smoking media campaigns
 - spurred clean-air ordinances throughout the state
 - produced more than \$100 million per year in anti-tobacco projects
- Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)

Cigarette Consumption: CA and the Rest of the US



Cigarette Consumption: CA and synthetic CA

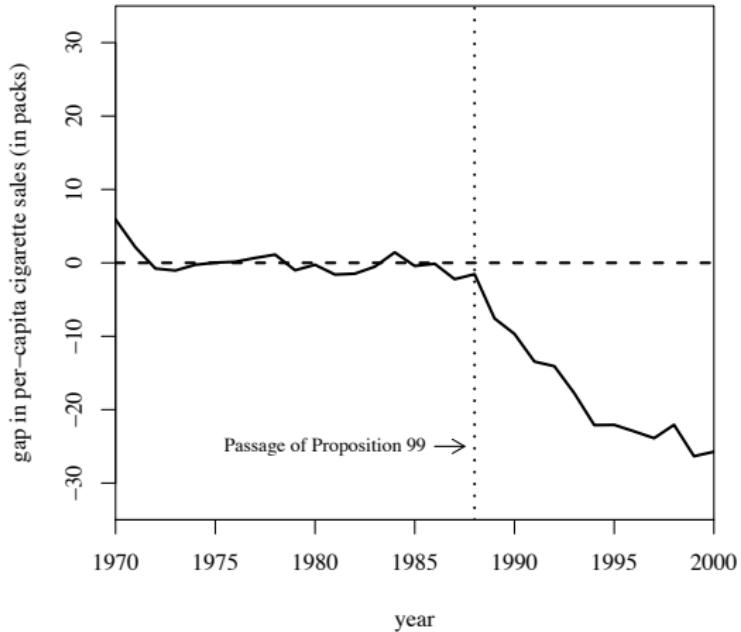


Predictor Means: Actual vs. Synthetic California

Variables	Real	California Synthetic	Average of 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 between CA and synthetic CA



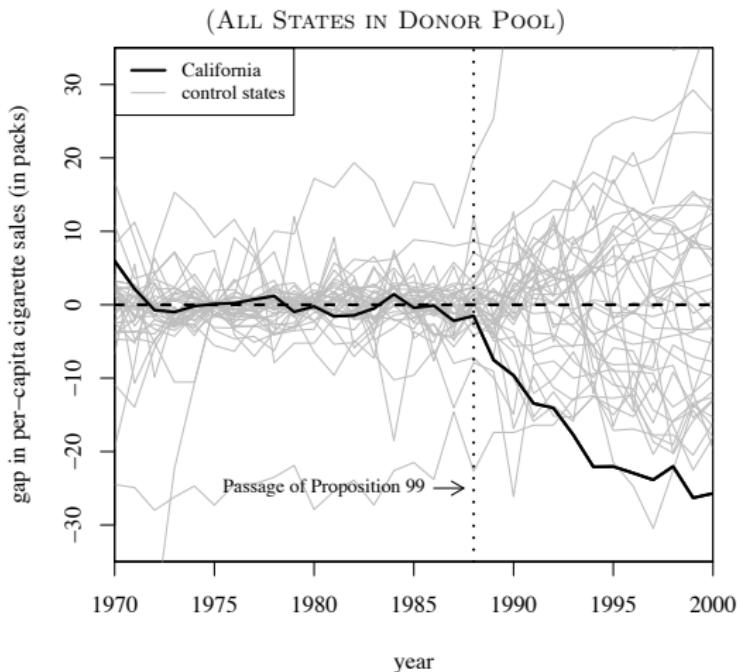
Inference

- To assess significance, we calculate exact p-values under Fisher's sharp null using a test statistic equal to after to before ratio of RMSPE
- Exact p-value method
 - Iteratively apply the synthetic method to each country/state in the donor pool and obtain a distribution of placebo effects
 - Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example the post-Prop. 99 RMSPE is:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^T \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

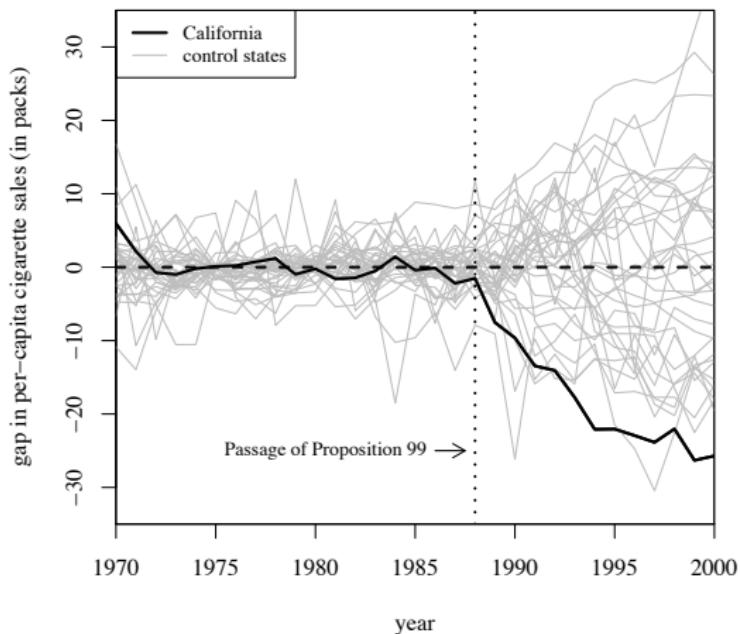
and the exact p-value is the treatment unit rank divided by J

Smoking Gap for CA and 38 control states



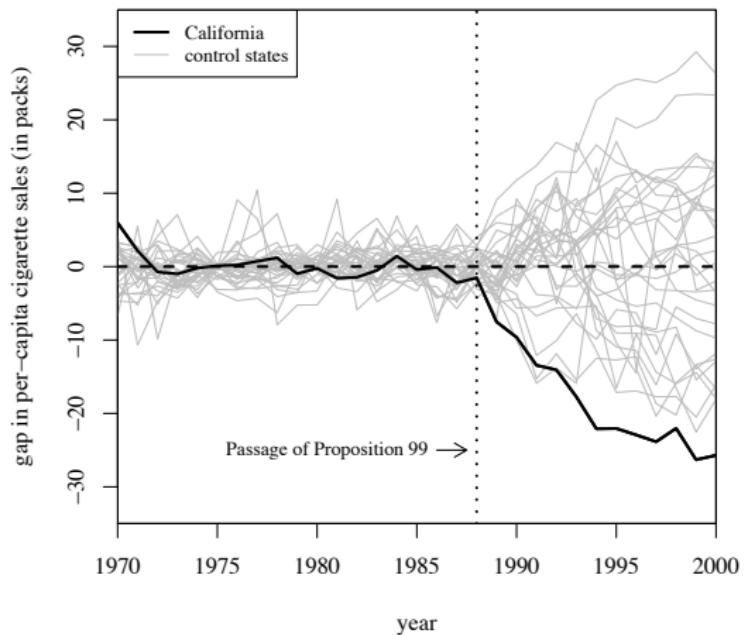
Smoking Gap for CA and 34 control states

(PRE-PROP. 99 MSPE \leq 20 TIMES PRE-PROP. 99 MSPE FOR CA)



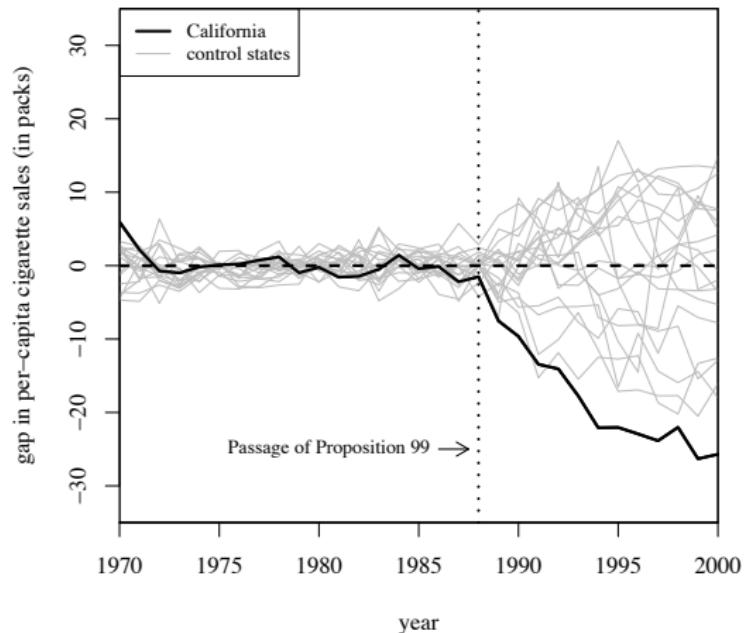
Smoking Gap for CA and 29 control states

(PRE-PROP. 99 MSPE \leq 5 TIMES PRE-PROP. 99 MSPE FOR CA)

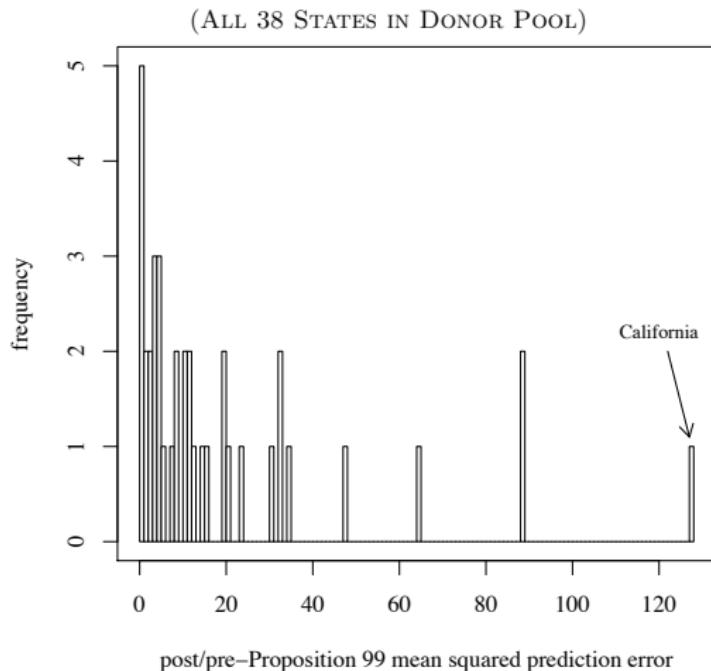


Smoking Gap for CA and 19 control states

(PRE-PROP. 99 MSPE \leq 2 TIMES PRE-PROP. 99 MSPE FOR CA)



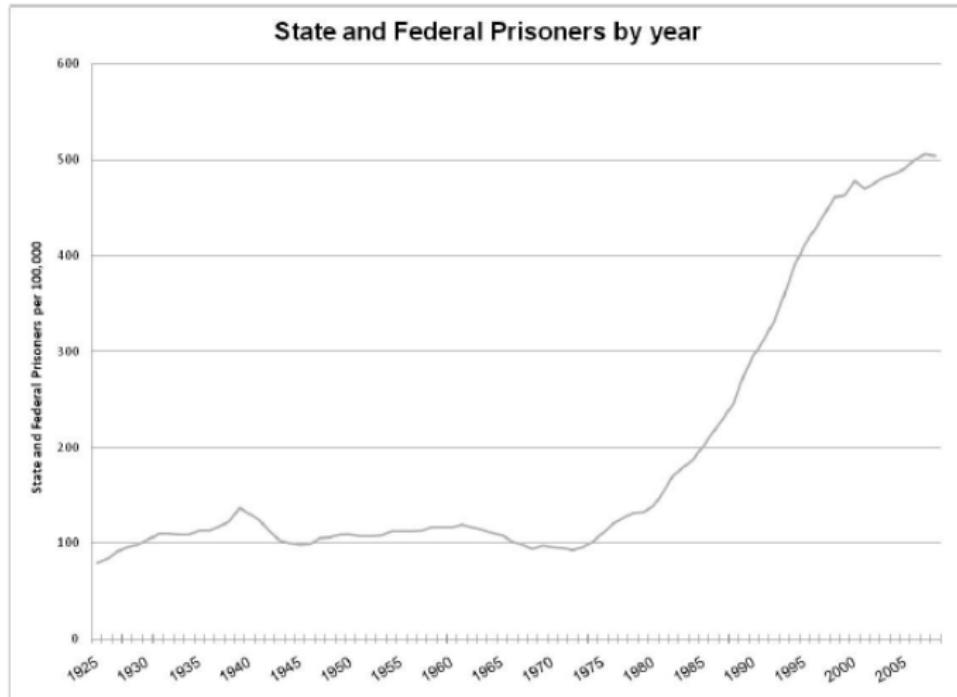
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Facts

- The US has the highest prison population of any OECD country in the world
- 2.3 million are currently incarcerated in US federal and state prisons and county jails
- Another 4.75 million are on parole
- From the early 1970s to the present, incarceration and prison admission rates quintupled in size

Figure 1
History of the imprisonment rate, 1925 - 2008



Source: www.albany.edu/sourcebook/tost_6.html

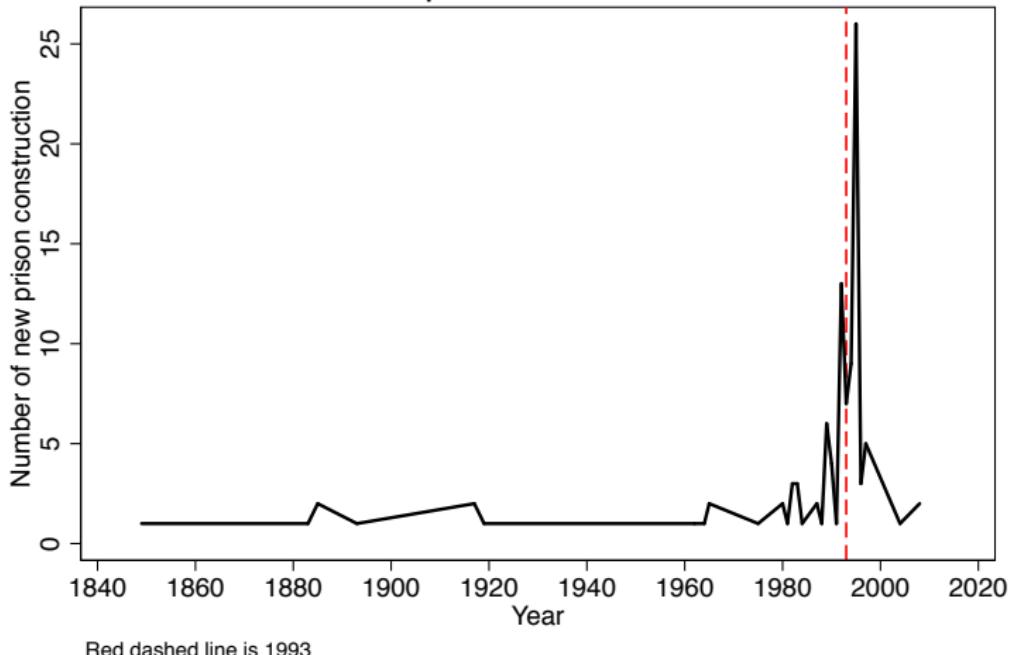
Prison constraints

- Prisons are and have been at capacity for a long time.
- Requires managing flows through
 - Prison construction
 - Overcrowding
 - Paroles

Texas prison boom

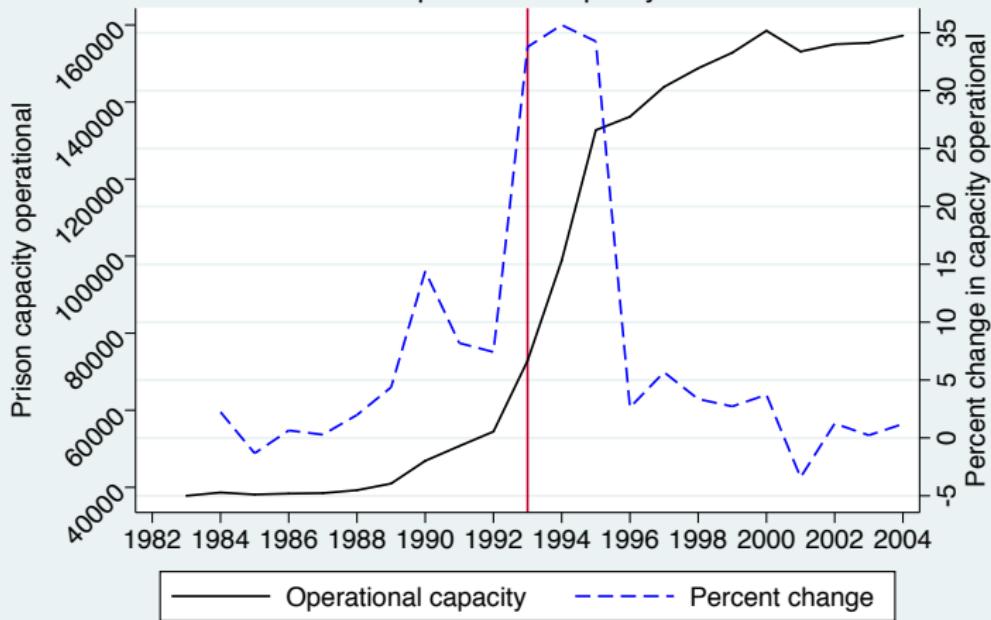
- Ruiz v. Estelle 1980
 - Class action lawsuit against TX Dept of Corrections (Estelle, warden).
 - TDC lost. Lengthy period of appeals and legal decrees.
 - Lengthy period of time relying on paroles to manage flows
- Governor Ann Richards (D) 1991-1995
 - Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
 - Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
 - Building of new prisons (private and public)

New prison construction

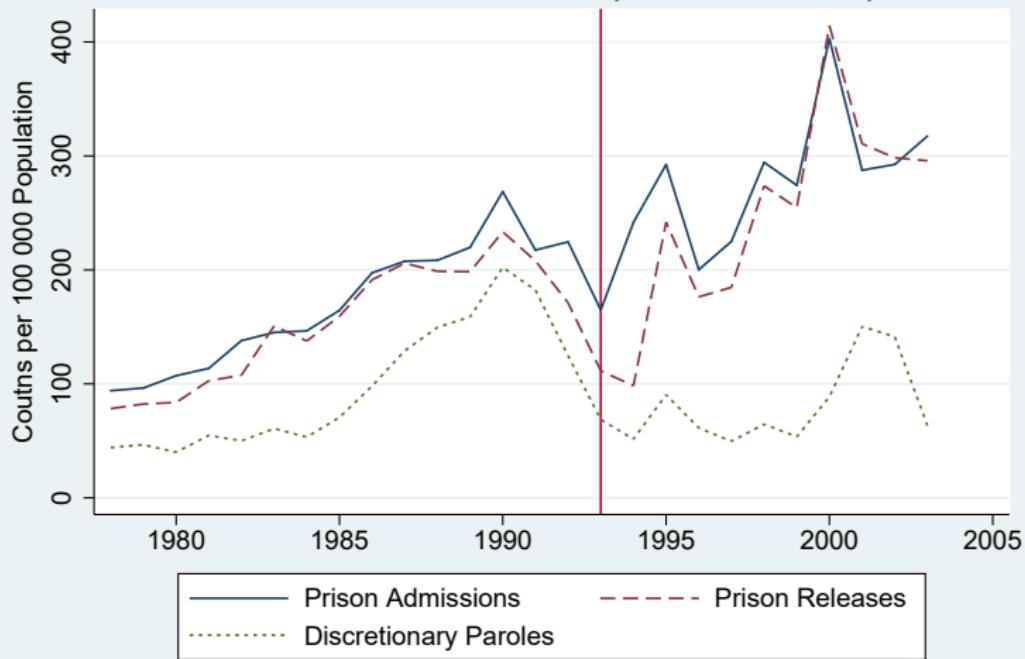


Texas prison growth

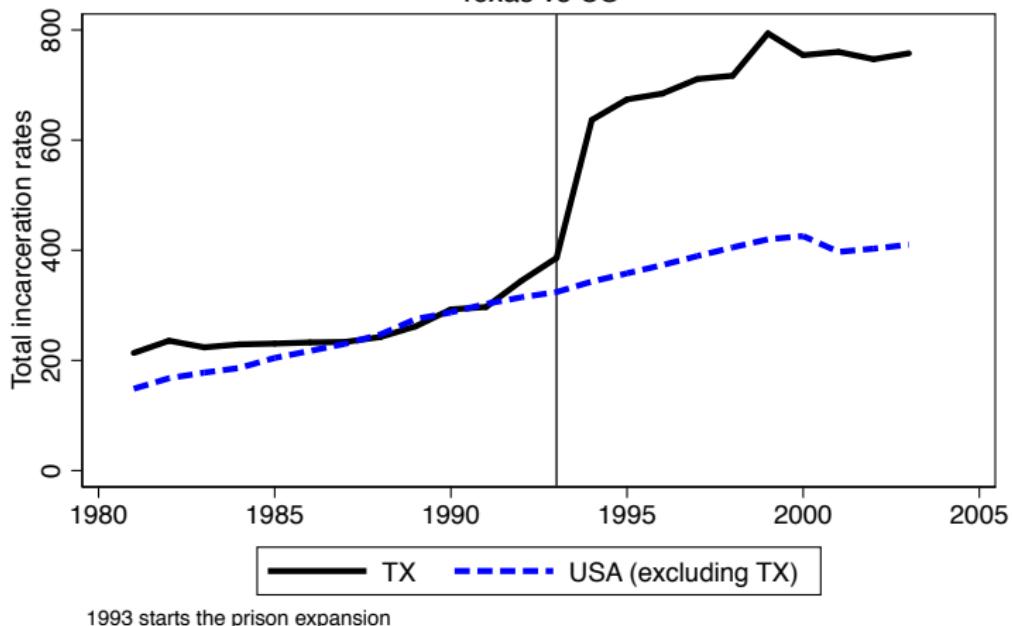
Operational capacity



Texas Prison Flows Measures per 100 000 Population



Total incarceration per 100 000 Texas vs US



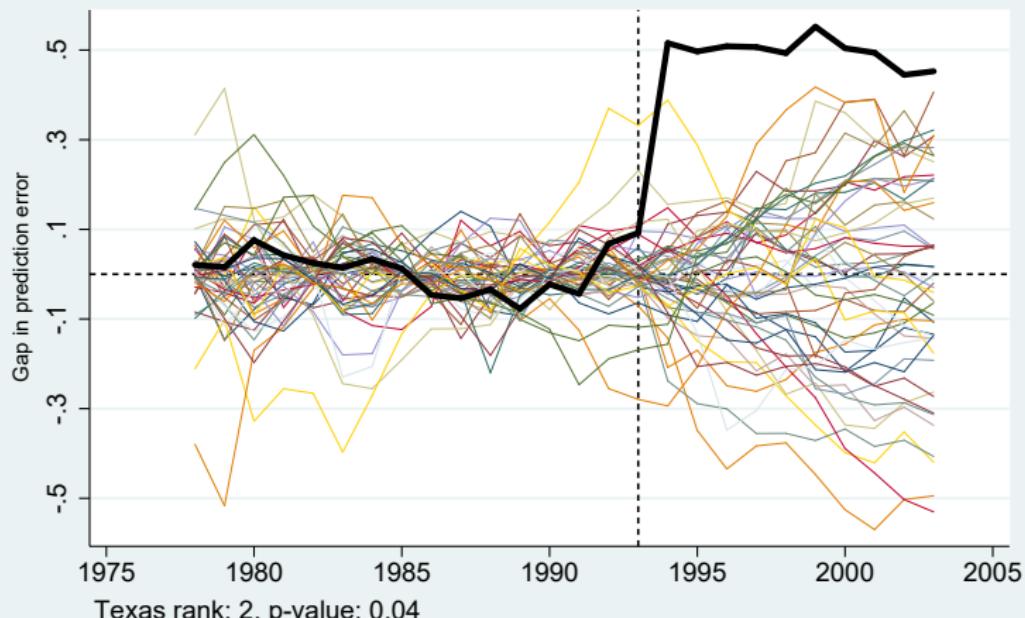
1993 starts the prison expansion

Data

- National Prisoner Statistics - prison measures, including race and gender-specific incarceration
- Current Population Survey - controls
- SEER - population

Incarcerated persons per 100,000

1993 Treatment



Parameters of interest

- We've discussed several important causal effects we may care about:
 - Average Treatment Effects – most external validity
 - Average Treatment on the Treated – impact on just the treated group
 - Local Average Treatment Effects – impact only on the subpopulation who complies
- RDD and IV can identify the LATE
- DD and synth can identify ATT
- What can estimate the ATE?

History of non-experimental matching

- A set of techniques for estimating ATE emerged in the 20th century from statistics and epidemiology
- Largely driven by smoking's connection to lung cancer
- Weighting methods emerged to estimate ATE from non-experimental data
- Interestingly, these methods would evolve into the contemporary suite of estimators I call matching but it includes weighting too

What is matching?

Define the ATE:

$$E[Y^1] - E[Y^0]$$

and its sample analog:

$$\frac{1}{N} \sum_i [Y_i^1 - Y_i^0]$$

What is matching?

Estimator:

$$\hat{\delta}^{ATT} = \frac{N_T}{N} \hat{\delta}^{ATT} + \frac{N_C}{N} \hat{\delta}^{ATU}$$

where $\hat{\delta}^{ATT} = \frac{1}{N_T} \sum_i \left[Y_i^1 - \hat{Y}_i^0 \right]$ and equivalent version for ATU.

What is \hat{Y}_i^0 for the treatment group? It's an **estimated** counterfactual. And this is its estimated value:

$$\hat{Y}_i^0 = \frac{1}{Pr(D)} \sum_{j \in \{D_j=0\}} w_{ij} Y_j^0$$

What is matching?

- We estimate counterfactuals as **weighted** averages over outcomes from the other group (e.g., control)
- Basically, no matter how fancy we get, matching is really always this form
- Estimators differ in how the weights are calculated
- Part of our goal is **balance** covariates between treatment and control group

Smoking thought experiment

- Split a large enough population to gain enough power to detect causal effects into treatment and control
- Treatment spends their lives smoking a pack of day; control abstains
- Compare lung cancer rates between the two groups
- Low realism: can't really expect people to comply with a longterm experiment like this
- But important, so how did scientists proceed?
- Through weighting based methods

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

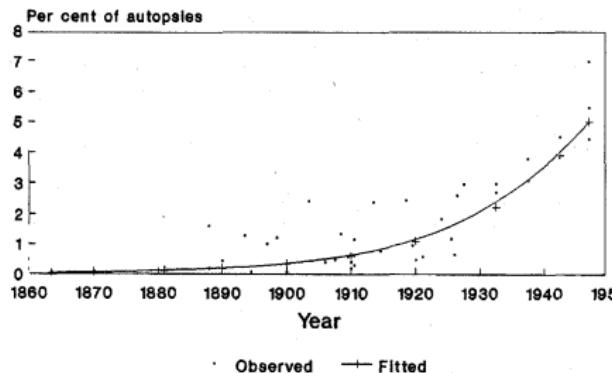


Figure 4
Smoking and Lung Cancer Case-control Studies

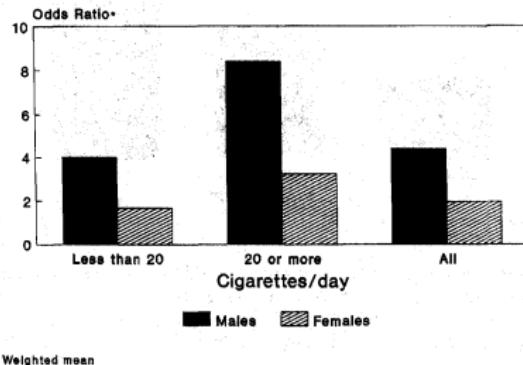
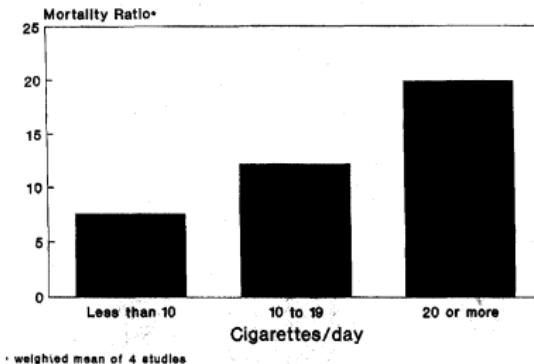
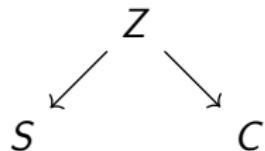


Figure 5
Smoking and Lung cancer Cohort Studies in Males



Does Smoking Cause Cancer?

Smoking, S , causes lung cancer, C ($S \rightarrow C$) versus spurious correlation due to the Z confounder:



Legitimate criticism at the time, but incorrect in hindsight – hindsight is 20/20

Nature of the criticism

Other criticisms came from giants like Joseph Berkson, Jerzy Neyman and Ronald Fisher

- ① Correlation b/w smoking and lung cancer was spurious due to biased selection of subjects
- ② Complaints about functional forms using “risk ratios” and “odds ratios”
- ③ Implausible magnitudes
- ④ Killer critique: *no experimental evidence* to incriminate smoking as a cause of lung cancer

Fisher's confounding theory

- Fisher, equally famous as a geneticist, argued from logic, statistics and genetic evidence for a hypothetical confounding genome, Z , and therefore smokers and non-smokers were not exchangeable (violation of independence assumption)
- Other studies showed that cigarette smokers and non-smokers were different on observables – more extraverted than non-smokers and pipe smokers, differed in age, differed in income, differed in education, etc.
- (FWIW Fisher was a chain smoking pipe smoker, he died of cancer, and he was a paid expert witness for the tobacco industry)

Broken clocks are sometimes right

- Always easy to criticize someone when we look back with more information
- Evidence for the *causal* link was shallow:
"the [epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions." Robert Hooke (1983)
- Scientists shifted to fixing their broken clocks for good

Observable selection bias

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

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

Are cigars more dangerous than cigarettes? Are cigarettes safe?

Non-smokers and smokers differ in mortality but also age

Table: Mean ages, years (Cochran 1968)

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

- Older people die at a higher rate, but for reasons other than just smoking cigars
- Could cigar smokers have higher observed death rates because they're older on average?
- How can we check this?

Force groups to have the same age distribution

- Covariates are *not balanced* – their mean values differ for treatment and control group.
- Subclassification (also called stratification) compares mortality rates across the different smoking groups *within* each age group
- Compare within covariate strata and then combine differences to neutralize observed confounders
- Weight the data so that covariates are balanced, then compare mortality across treatment and control
- Which weights?

Weighting the data

- Divide the smoking group samples into age groups
- For each of the smoking group samples, calculate the mortality rates for the age group
- Construct probability weights for each age group as the proportion of the sample with a given age
- Compute the weighted averages of the age groups mortality rates for each smoking group using the probability weights
- This will oddly enough balance the observed covariates between treatment and control

Simple weighting example

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

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

Simple weighting example

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

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

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

Simple weighting example

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

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

Simple weighting example

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

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

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

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

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

Observable Covariate

We are going to use matching methods to adjust for observable differences between these three groups

Definition: Predetermined Covariates

Variable X is a covariate if for each individual i , the value of X_i does not depend on treatment status.

- Does not imply X and treatment status are independent
- Sometimes covariates do not change over time, but doesn't have to be
- Beware of colliders or "bad controls"

Adjustment for Observables

- Simple weighting methods (e.g., subclassification)
- Exact matching methods (e.g., nearest neighbors)
- Approximate matching methods (e.g., propensity scores)

Assumptions, data and statistics

- We need three things, and really only three things, to estimate a causal effect
- **Assumptions:** what must we assume is true so that our models work with data?
- **Data:** what data with what covariates and outcomes do we need for this project?
- **Statistical models:** sometimes called “estimators” which crank data into estimates which equal the causal effect?
- Many moving parts with many players

Assumption I: Independence

- Randomized treatment assignment guarantees “independence”
$$(Y^0, Y^1) \perp\!\!\!\perp D$$

- Independence allows to estimate accurate causal effects through simple methods like differences in averages

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

Violations of independence

- Problem with smoking and cancer was smoking wasn't *randomly assigned*
- Since it wasn't randomly assigned, smoking was not "independent" of potential outcomes
- When a treatment is "dependent" on potential outcomes, it means people smoke because they expect something is better when they smoke (Y^1) than when they don't (Y^0)
- So what? Means you can't compare naively – you have to adjust for whatever is needed to recover "conditional independence"

Identification under conditional independence

Identification assumptions:

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

Implications of independence

- Assumption 1 lets you plug Y for Y^j

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

- Assumption 2 lets you weight

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

Independence breaks down

- Independence is violated, though, if the treatment was assigned *because* we expected things to improve or not
- That's because the causal effect is $\delta = Y^1 - Y^0$
- If you take an action because you think $\delta > 0$, then you are admitting independence doesn't hold
- Firms don't ordinarily flip coins to set prices – they set prices based on profit maximization

Conditional independence

- But, if there is *some* random price setting conditional on observable factors (which only employees, managers and executives could possibly know about), then we may be able to defend “conditional independence”
- Conditional independence means that once we adjust for covariates, all remaining variation in treatment assignment had nothing to do with profit maximization or potential outcomes more generally

Implications of assumptions

Conditional independence lets us do two things:

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

Can we defend any conditional independence?

- Maybe prices are conditionally independent of Y^0 but not Y^1
- Technically a weaker assumption, but also means we can't estimate ATE
- But we can estimate ATT

Partial conditional independence

- ① $Y^0 \perp\!\!\!\perp D|X$
- ② $Pr(D = 1|X) < 1$ (with $Pr(D = 1) > 0$)

We can then estimate the ATT.

Two parameter estimates

Weighted averages under either assumption:

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

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

Subclassification by Age ($K = 2$)

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

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

Subclassification by Age ($K = 2$)

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

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

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

Subclassification by Age ($K = 2$)

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

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

Subclassification by Age ($K = 2$)

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

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

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

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

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

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

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

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

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

Not identified!

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

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

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

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

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

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

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

Curse of Dimensionality

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

Curse of Dimensionality

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

Nearest Neighbor Matching

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

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

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

Matching

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

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

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

Matching

- We can also use matching to estimate δ_{ATE} . In that case, we match in both directions:
 - ① If observation i is treated, we impute Y_i^0 using the control matches, $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
 - ② If observation i is control, we impute Y_i^1 using the treatment matches, $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
- The estimator is:

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N (2D_i - 1) \left[Y_i - \left(\frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} \right) \right]$$

Matching example with single covariate

unit <i>i</i>	Potential Outcome			D_I	X_i	
	under Treatment	Y_i^1	under Control	Y_i^0		
1		6		?	1	3
2		1		?	1	1
3		0		?	1	10
4				0	0	2
5				9	0	3
6				1	0	-2
7				1	0	-4

Question: What is $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Matching example with single covariate

unit <i>i</i>	Potential Outcome		D_I	X_i
	under Treatment	under Control		
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Match and plug in!

Matching example with single covariate

unit <i>i</i>	Potential Outcome			
	under Treatment	under Control	D_I	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

$$\widehat{\delta}_{ATT} = \frac{1}{3} \cdot (6 - 9) + \frac{1}{3} \cdot (1 - 0) + \frac{1}{3} \cdot (0 - 9) = -3.7$$

A Training Example

Trainees			Non-Trainees		
unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900
2	34	10200	2	50	31000
3	29	14400	3	30	21000
4	25	20800	4	27	9300
5	29	6100	5	54	41100
6	23	28600	6	48	29800
7	33	21900	7	39	42000
8	27	28800	8	28	8800
9	31	20300	9	24	25500
10	26	28100	10	33	15500
11	25	9400	11	26	400
12	27	14300	12	31	26600
13	29	12500	13	26	16500
14	24	19700	14	34	24200
15	25	10100	15	25	23300
16	43	10700	16	24	9700
17	28	11500	17	29	6200
18	27	10700	18	35	30200
19	28	16300	19	32	17800
Average:		28.5	16426	20	23
			21	32	25900
			Average:		20724

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:		20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:		20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

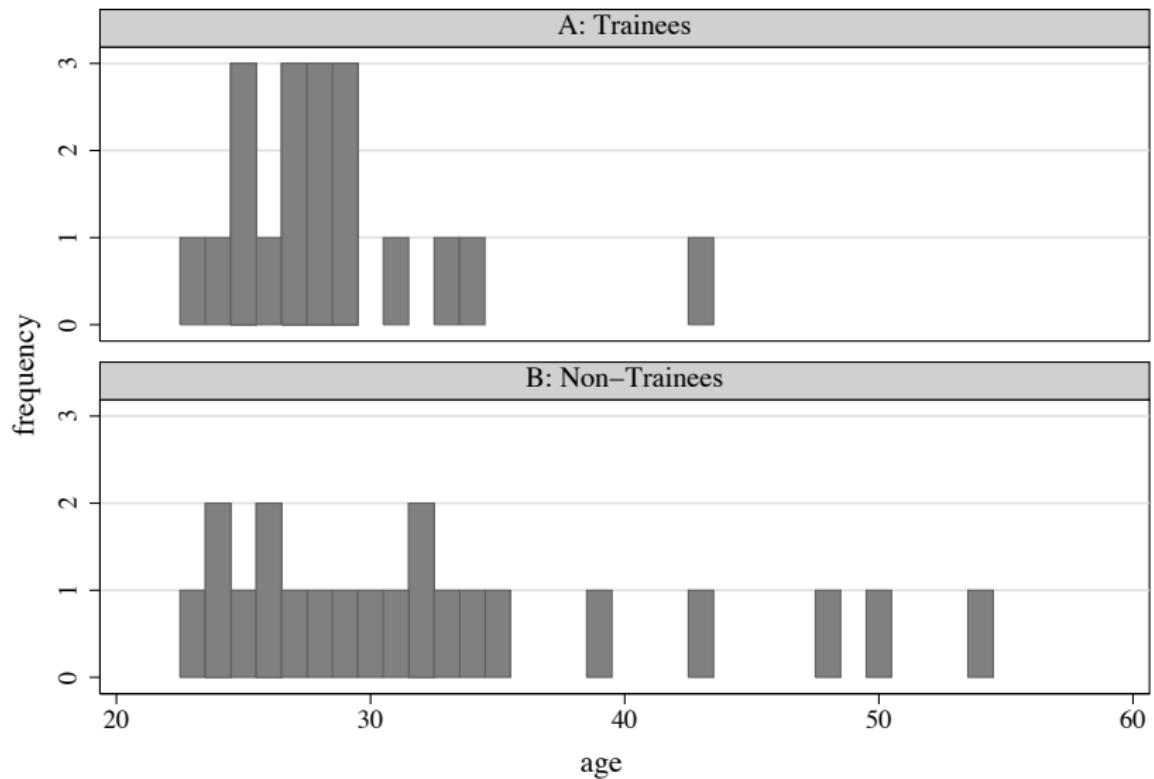
A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

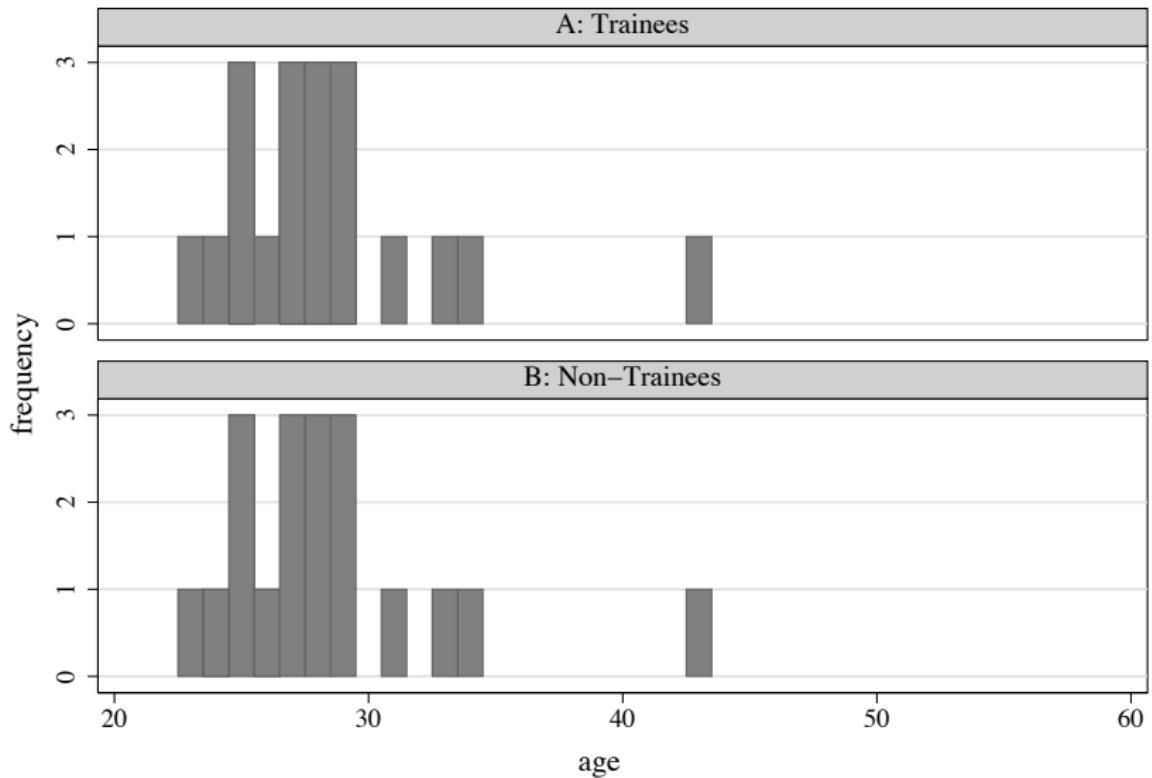
A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:	28.5	13982
			21	32	25900			
			Average:		20724			

Age Distribution: Before Matching



Age Distribution: After Matching



Training Effect Estimates

Difference in average earnings between trainees and non-trainees

- Before matching

$$16426 - 20724 = -4298$$

- After matching:

$$16426 - 13982 = 2444$$

Alternative distance metric: Euclidean distance

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

Definition: Euclidean distance

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

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

Normalized Euclidean distance

Definition: Normalized Euclidean distance

A commonly used distance is the normalized Euclidean distance:

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

where

$$\hat{V}^{-1} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_k^2 \end{pmatrix}$$

- Notice that the normalized Euclidean distance is equal to:

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})}{\hat{\sigma}_n^2}}$$

- Thus, if there are changes in the scale of X_{ni} , these changes also affect $\hat{\sigma}_n^2$, and the normalized Euclidean distance does not change

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

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

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

Arbitrary weights

Or, you could just create your own arbitrary weights

$$||X_i - X_j|| = \sqrt{\sum_{n=1}^k \omega_n \cdot (X_{ni} - X_{nj})^2}$$

(with all $\omega_n \geq 0$) so that we assign large ω_n 's to those covariates that we want to match particularly well.

Matching and the Curse of Dimensionality

Dimensionality creates headaches for us in matching.

- **Bad news:** Matching discrepancies $\|X_i - X_{j(i)}\|$ tend to increase with k , the dimension of X
- **Good news:** Matching discrepancies converge to zero . . .
- **Bad news:** . . . but they converge very slow if k is large
- **Good news:** Mathematically, it can be shown that $\|X_i - X_{j(i)}\|$ converges to zero at the same rate as $\frac{1}{N^{\frac{1}{k}}}$
- **Bad news:** It's hard to find good matches when X has a large dimension: you need many observations if k is big.

Deriving the matching bias

Derive the matching bias by first writing out the sample ATT estimate:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where each i and $j(i)$ units are matched, $X_i \approx X_{j(i)}$ and $D_{j(i)} = 0$. Define potential outcomes and switching eq.

$$\begin{aligned}\mu^0(x) &= E[Y|X=x, D=0] = E[Y^0|X=x], \\ \mu^1(x) &= E[Y|X=x, D=1] = E[Y^1|X=x], \\ Y_i &= \mu^{D_i}(X_i) + \varepsilon_i\end{aligned}$$

Substitute and distribute terms

$$\begin{aligned}\hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Deriving the matching bias

Difference between sample estimate and population parameter is:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Algebraic manipulation and simplification:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_i) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\mu^0(X_i) - \mu^0(X_{j(i)})) .\end{aligned}$$

Deriving the matching bias

Apply the Central Limit Theorem and the difference,

$$\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT}),$$

converges to a Normal distribution with zero mean. But, however,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$ may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$ may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Bias is often an issue when we match in many dimensions

Solutions to matching bias problem

The bias of the matching estimator is caused by large matching discrepancies $\|X_i - X_{j(i)}\|$. The curse of dimensionality virtually guarantees this. However:

- ① But the matching discrepancies are observed. We can always check in the data how well we're matching the covariates.
- ② For $\widehat{\delta}_{ATT}$ we can always make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making N_C large).
- ③ If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques
- ④ Partial solution: propensity score methods (coming soon...)

Matching with bias correction

- Each treated observation contributes

$$\mu^0(X_i) - \mu^0(X_{j(i)})$$

to the bias.

- Bias-corrected (BC) matching:

$$\widehat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[(Y_i - Y_{j(i)}) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})) \right]$$

where $\widehat{\mu^0}(X)$ is an estimate of $E[Y|X = x, D = 0]$. For example using OLS.

- Under some conditions, the bias correction eliminates the bias of the matching estimator without affecting the variance.

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

Bias adjustment in matched data

unit	Potential Outcome		D_i	X_i
	under Treatment	under Control		
i	Y_i^1	Y_i^0		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{(10 - 8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4 - 1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3} \\ &\quad + \frac{(10 - 9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33\end{aligned}$$

Matching bias: Implications for practice

Bias arises because of the effect of large matching discrepancies on $\mu^0(X_i) - \mu^0(X_{j(i)})$. To minimize matching discrepancies:

- ① Use a small M (e.g., $M = 1$). Larger values of M produce large matching discrepancies.
- ② Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement.
- ③ Try to match covariates with a large effect on $\mu^0(\cdot)$ particularly well.

Large sample distribution for matching estimators

- Matching estimators have a Normal distribution in large samples (provided the bias is small):

$$\sqrt{N_T}(\hat{\delta}_{ATT} - \delta_{ATT}) \xrightarrow{d} N(0, \sigma_{ATT}^2)$$

- For matching without replacement, the “usual” variance estimator:

$$\hat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \hat{\delta}_{ATT} \right)^2,$$

is valid.

Large sample distribution for matching estimators

- For matching with replacement:

$$\begin{aligned}\widehat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2 \\ &\quad + \frac{1}{N_T} \sum_{D_i=0} \left(\frac{K_i(K_i-1)}{M^2} \right) \widehat{\text{var}}(\varepsilon | X_i, D_i = 0)\end{aligned}$$

where K_i is the number of times observation i is used as a match.

- $\widehat{\text{var}}(Y_i | X_i, D_i = 0)$ can be estimated also by matching. For example, take two observations with $D_i = D_j = 0$ and $X_i \approx X_j$, then

$$\widehat{\text{var}}(Y_i | X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of $\widehat{\text{var}}(\varepsilon_i | X_i, D_i = 0)$)

- The bootstrap doesn't work!

Avoiding dimensionality problems

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

Least squares

- OLS is best linear predictor and approximation to the conditional expectation function
- But if probability of treatment is nonlinear, this conditional mean may be less informative
- Propensity scores relax the linearity assumption and have other advantages

The Idea behind propensity scores

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

Identifying assumptions

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

Identifying assumption I: Conditional independence

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

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

Identifying assumption II: Common support

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

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

Formal Definition

Definition of Propensity score

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

Two Necessary Identification Assumptions:

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

Steps

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

Estimating the propensity score

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

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

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

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

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

Identification

- A group of unit's average treatment effect may depend on some characteristic, X

$$\begin{aligned} E[\delta_i(X_i)] &= E[Y_i^1 - Y_i^0 | X_i = x] \\ &= E[Y_i^1 | X_i = x] - E[Y_i^0 | X_i = x] \end{aligned}$$

- CIA allow us to substitute

$$E[Y_i | D_i = 1, X_i = x] = E[Y_i^1 | D_i = 1, X_i = x]$$

and similar for other term Y^0 using a switching equation

- Common support allows us to estimate both terms

Propensity score theorem

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

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

Proof

- Before diving into the proof, first recognize that

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

because

$$\begin{aligned} E[D | Y^0, Y^1, \rho(X)] &= 1 \times Pr(D = 1 | Y^0, Y^1, \rho(X)) \\ &\quad + 0 \times Pr(D = 0 | Y^0, Y^1, \rho(X)) \end{aligned}$$

and the second term cancels out.

Proof.

Assume $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (CIA). Then:

$$\begin{aligned} Pr(D = 1|Y^1, Y^0, \rho(X)) &= \underbrace{E[D|Y^1, Y^0, \rho(X)]}_{\text{See previous slide}} \\ &= \underbrace{E[E[D|Y^1, Y^0, \rho(X), X]|Y^1, Y^0, \rho(X)]}_{\text{by LIE}} \\ &= \underbrace{E[E[D|Y^1, Y^0, X]|Y^1, Y^0, \rho(X)]}_{\text{Given } X, \text{ we know } \rho(X)} \\ &= \underbrace{E[E[D|X]|Y^1, Y^0, \rho(X)]}_{\text{by CIA}} \\ &= \underbrace{E[\rho(X)|Y^1, Y^0, \rho(X)]}_{\text{propensity score definition}} \\ &= \rho(X) \end{aligned}$$



Similar proof

We also can show that the probability of treatment conditional on the propensity score is the propensity score using a similar argument:

$$\begin{aligned} \Pr(D = 1 | \rho(X)) &= \underbrace{E[D | \rho(X)]}_{\text{Previous slide}} \\ &= \underbrace{E[E[D | X] | \rho(X)]}_{\text{LIE}} \\ &= \underbrace{E[p(X) | \rho(X)]}_{\text{definition}} \\ &= \rho(X) \end{aligned}$$

and $\Pr(D = 1 | Y^1, Y^0, \rho(X)) = \Pr(D = 1 | \rho(X))$ by CIA

Unbiased estimation of the ATE

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

Corollary: Estimating the ATE

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

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

Balancing property

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

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

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

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

Balancing property

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

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

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

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

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

Propensity score theorem

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

Checking the common support assumption

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

Inverse probability weighting

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

Inverse probability weighting

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

Inverse Probability Weighting

Proposition

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

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

IPW Proof

Proof.

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

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



Weighting on the propensity score

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

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

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

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

Weighting on the propensity score

Standard errors can be constructed a few different ways:

- We need to adjust the standard errors for first-step estimation of $\rho(X)$
- Parameteric first step: Newey and McFadden (1994)
- Non-parametric first step: Newey (1994)
- Or bootstrap the entire two-step procedure (Adudumilli 2018 and Bodory et al. 2020)

Implementation with software

- I like estimating with IPW manually because I like being reminded how simple a procedure it is
- But Stata's `-teffects-` and R's `-ipw-` do it too, and `-teffects-` uses the Hajek normalization weights which will produce identical estimates to my program
- My programs don't do the inference, but I think that would be fun and easy to do using the bootstrap
- Let's look at it real quickly now with an example from LaLonde's 1986 paper on the NSW job trainings program (which I'll discuss again soon)

Double robust estimators

- Lots of papers: Robins and Rotnizky (1995) originally, Hirano and Imbens (2001), etc.
- Basic idea is you are going to control for covariates twice: through regression and then through the propensity score
- We say that estimators combining regression with IPW are double robust so long as
 - The regression for the outcome is properly specified, or
 - The propensity score is properly specified
- Hence the name “double robust”. We give ourselves two chances to get it right (either/or not both/and)

Estimation of outcome model

$$y_i = \alpha_0 + X_i\beta + \tilde{\alpha}_1 D_i + \theta_0 \frac{D_i}{\widehat{\rho(X_i)}} + \theta_1 \frac{1 - D_i}{1 - \widehat{\rho(X_i)}} + \tilde{\varepsilon}_i$$

Propensity score matching

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

Standard matching strategy

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

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

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

Imputing the counterfactuals

A parameter of interest:

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

We estimate it as follows

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

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

Matching methods

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

Stratification

- Stratification is used to force covariate balance by finding strata where there is no difference in mean covariate values.
- You then use those strata to calculate within differences in means and sum over properly weighted strata. See Becker and Ichino (2002)
- Stratification is a brute force method for imposing balance by grouping the data and testing for differences in covariate means
- It's actually kind of similar to coarsened exact matching, only using the propensity score for the "stratification" not the covariates

Stratification: Achieving Balance

The algorithm is brute force covariate balancing

- ① Sort the data by propensity score and divide into groups of observations with similar propensity scores (e.g., percentiles)
- ② Within each group, test (using a t-test) whether the means of the covariates (X) are equal between treatment and control
- ③ If so, then stop. If not, it means the covariates aren't balanced *within that group*. Divide the group in half and repeat
- ④ If a particular covariate is unbalanced for multiple groups, modify the initial logit or probit equation by including higher order terms and/or interactions with that covariate and repeat

Historically this could be done with `-pscore2.ado-` or manually oneself if they felt so inclined, but it was dropped with `-teffects-`

Nearest Neighbor

Pretty similar to covariate matching. Formula is

$$ATT^{NN} = \frac{1}{N_T} \sum_{i: W_i=1} \left[Y_i - \sum_{j \in C(i)_M} w_{ij} Y_j \right]$$

- N_T is the number of Treatment units i
- w_{ij} is equal to $\frac{1}{N_C}$ if j is a control unit and zero otherwise; N_C is number of control units j
- And unit j is chosen as a control for i if it's propensity score is nearest to that of i

NN Matching: Bias vs. Variance

- But how far away on the propensity score will you use? Herein lies the different types of matching proposed
 - Matching just one nearest neighbor minimizes bias at the cost of larger variance
 - Matching using additional nearest neighbors increases the bias but decreases the variance
- Matching with or without replacement
 - with replacement keeps bias low at the cost of larger variance
 - without replacement keeps variance low but at the cost of potential bias

Distance between treatment and control units

- What was historically done was limiting “distance” through various *ad hoc* choices
- Imagine these choices as creating like a lasso (like the cowboy rope)
- Anything within the lasso could be used for the imputation; anything outside the lasso could not
- There were two common ways – caliper matching and radius matching.

Caliper matching

- Caliper matching is a variation on NN matching that tries to build brakes into the algorithm as to avoid “bad neighbors”
- It does this by imposing a tolerable maximum distance (e.g., 0.2 units in the propensity score away from a treatment unit i 's propensity score)
- Note – this is a one-to-one imputation, and if there doesn't exist anybody in the control group unit j within that “caliper”, then treatment unit i is discarded
- Means we aren't estimating the ATE anymore once we start dropping units
- It's difficult to know what this caliper should be *ex ante*, hence why I said it is somewhat *ad hoc*

Radius matching

- Each treatment unit i is matched with the control group units whose propensity score are in a predefined neighborhood of the propensity score of the treatment unit.
- All the control units with $\hat{\rho}_j$ falling within a radius r from $\hat{\rho}_i$ are matched to the treatment unit i – this is what distinguishes it from calipers, and makes it more similar to covariate matching (Abadie and Imbens 2006, 2008)
- The smaller the radius, the better the quality of the matches, but the higher the possibility some treatment units are not matched because the neighborhood does not contain control group units j

Software

- I think you can use -teffects, psmatch- to get at these two nearest neighbor approaches by setting the number of matches
- You can use -pscore2- for stratification, but I think the standard errors are wrong, so you may need to just do it manually using bootstrapping or variance approximation, and that may be a pain to program up
- Not sure of the R command, but I know it's out there

King and Nielsen (2019)

- There is a King and Nielsen (2019) critique of these methods that is popularly known but not popularly studied
- King and Nielsen (2019) is not a critique of the propensity score, because it does not apply to stratification, regression adjustment, or inverse probability weighting
- It only applies to nearest neighbor and is related to forced balance through trimming and a myriad of other common choices made by the researcher

“[The] more balanced the data, or the more balance it becomes by [trimming] some of the observations through matching, the more likely propensity score matching will degrade inferences.” – King and Nielsen (2019)

Examples of propensity score matching

- Workhorse example of propensity score matching is the Job Trainings Program (NSW)
- First studied by LaLonde (1986) evaluating multiple econometric models for program evaluation
 - All the standard estimators failed to estimate the known ATE when replacing experimental controls with non-experimental controls – even difference-in-differences
- Dehejia and Wahba (1999; 2002) use LaLonde's data with propensity score matching and found better results
- Critiques by Petra Todd, Jeff Smith and others followed which I won't review here for sake of time

Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
 - **Treatment group:** received all the benefits of NSW program
 - **Control group:** left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

NSW Program

- Treatment group members were:
 - guaranteed a job for 9-18 months depending on the target group and site
 - divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
 - paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

NSW Program

- Other details about the NSW program:
 - Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
 - Post-treatment: after their term expired, they were forced to find regular employment
 - Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

NSW Data

- NSW data collection:

- MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
- Conducted up to 4 post-baseline interviews
- Different sample sizes from study to study can be confusing, but has simple explanations

NSW Data

- Estimation:
 - NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:
$$\frac{1}{N_t} \sum_{D_i=1} Y_i - \frac{1}{N_c} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$
in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e., $(Y^0, Y^1) \perp D$.
- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful – $\approx \$900$ (LaLonde 1986) to $\$1,800$ (Dehejia and Wahba 2002) depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of \$900

LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
 - His estimates were way off and usually the wrong sign
 - Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences:		Unrestricted Difference in Earnings Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	(\$455 (704))	(\$697 (760))	(\$509 (967))	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences:		Unrestricted Difference in Earnings Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

covariate	All		CPS	NSW		
	mean	(s.d.)	Controls	Trainees	t-stat	diff
			$N_c = 15,992$	$N_t = 297$		
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Dehejia and Wahba (1999)

- Dehejia and Wahba (DW) update LaLonde's original study using propensity score matching
 - ① Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". Journal of the American Statistical Association, vol. 94(448): 1053-1062 (pdf)
- Can propensity score matching improve over the estimators that LaLonde examined?

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
NSW/Lalonde:^a									
Treated	297	24.63 (.32)	10.38 (.09)	.80 (.02)	.09 (.01)	.73 (.02)	.17 (.02)	3,066 (236)	
Control	425	24.45 (.32)	10.19 (.08)	.80 (.02)	.11 (.02)	.81 (.02)	.16 (.02)	3,026 (252)	
RE74 subset:^b									
Treated	185	25.81 (.35)	10.35 (.10)	.84 (.02)	.059 (.01)	.71 (.02)	.19 (.02)	2,096 (237)	1,532 (156)
Control	260	25.05 (.34)	10.09 (.08)	.83 (.02)	.1 (.02)	.83 (.02)	.15 (.02)	2,107 (276)	1,267 (151)
Comparison groups:^c									
PSID-1	2,490	34.85 [.78]	12.11 [.23]	.25 [.03]	.032 [.01]	.31 [.04]	.87 [.03]	19,429 [991]	19,063 [1,002]
PSID-2	253	36.10 [1.00]	10.77 [.27]	.39 [.04]	.067 [.02]	.49 [.05]	.74 [.04]	11,027 [853]	7,569 [695]
PSID-3	128	38.25 [1.17]	10.30 [.29]	.45 [.05]	.18 [.03]	.51 [.05]	.70 [.05]	5,566 [686]	2,611 [499]
CPS-1	15,992	33.22 [.81]	12.02 [.21]	.07 [.02]	.07 [.02]	.29 [.03]	.71 [.03]	14,016 [705]	13,650 [682]
CPS-2	2,369	28.25 [.87]	11.24 [.19]	.11 [.02]	.08 [.02]	.45 [.04]	.46 [.04]	8,728 [667]	7,397 [600]
CPS-3	429	28.03 [.87]	10.23 [.23]	.21 [.03]	.14 [.03]	.60 [.04]	.51 [.04]	5,619 [552]	2,467 [288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE74 = earnings in calendar year 19x.

^a NSW sample as constructed by Lalonde (1986).

^b The subset of the Lalonde sample for which RE74 is available.

^c Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

PSID-1 and CPS-1 are identical to those used by Lalonde. CPS2-3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

Table 2. Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparison Groups From the PSID and the CPS^a

A. Lalonde's original sample				B. RE74 subsample (results do not use RE74)				C. RE74 subsample (results use RE74)				
Comparison group	NSW		Unrestricted differences in earnings less comparison group earnings growth 1978		NSW		Unrestricted differences in earnings less comparison group earnings growth 1978		NSW		Unrestricted differences in earnings less comparison group earnings growth 1978	
	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e
NSW	886 (472)	798 (472)	879 (467)	802 (468)	820 (468)	1,794 (633)	1,672 (637)	1,750 (632)	1,631 (637)	1,612 (639)	1,794 (633)	1,688 (636)
PSID-1	-15,578 (913)	-8,057 (990)	-2,380 (680)	-2,119 (746)	-1,844 (762)	-15,205 (1155)	-7,741 (1175)	-582 (841)	-265 (881)	186 (901)	-15,205 (1155)	-879 (931)
PSID-2	-4,020 (781)	-3,482 (935)	-1,364 (729)	-1,694 (878)	-1,876 (885)	-3,647 (960)	-2,810 (1082)	721 (886)	298 (1004)	111 (1032)	-3,647 (960)	94 (1042)
PSID-3	697 (760)	-508 (967)	629 (757)	-552 (967)	-576 (968)	1,070 (900)	35 (1101)	1,370 (897)	243 (1101)	298 (1105)	1,070 (900)	821 (1100)
CPS-1	-8,870 (562)	-4,416 (577)	-1,543 (426)	-1,102 (450)	-987 (452)	-8,498 (712)	-4,417 (714)	-78 (537)	525 (557)	709 (560)	-8,498 (712)	-8 (572)
CPS-2	-4,195 (533)	-2,341 (620)	-1,649 (459)	-1,129 (551)	-1,149 (551)	-3,822 (671)	-2,206 (746)	-263 (574)	371 (662)	305 (666)	-3,822 (671)	615 (672)
CPS-3	-1,008 (539)	-1 (681)	-1,204 (532)	-263 (677)	-234 (675)	-635 (657)	375 (821)	-91 (641)	844 (808)	875 (810)	-635 (657)	1,270 (798)

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID-1–3, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly recreate his subset. Column (5) differs because the obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

^b Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

^c The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample (\$886) and for the RE74 sample (\$1,794).

^d The exogenous variables used in the regressions-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

^e Regresses RE78 on a treatment indicator and RE75.

^f The same as (d), but controls for the additional variables listed under (c).

^g Controls for all pretreatment covariates.

Proposition 2

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

- Conditional on the propensity score, the covariates are independent of the treatment, suggesting that the distribution of covariate values should be the same for both treatment and control groups
- This can be checked as we have data on all three once we've estimated the propensity score

Trimming the data

- Common terms are “trimming” or “pruning”
- Drop units which do not overlap in terms of estimated propensity score
- Sometimes as a rule of thumb, just keep units on the $[0.1, 0.9]$ interval

Common support

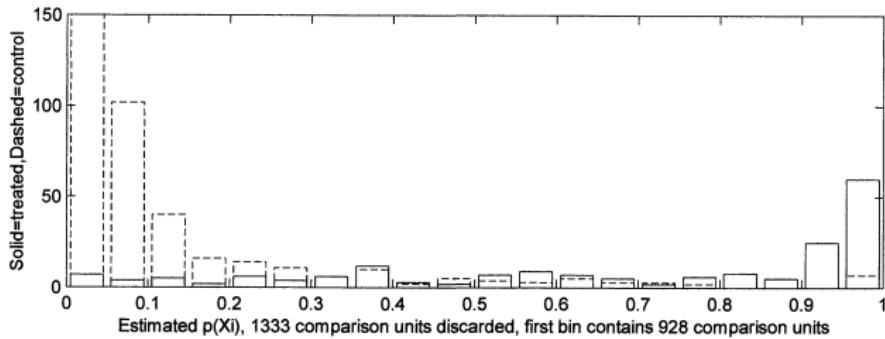


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8-.85, .85-.9, and .9-.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

Overlap

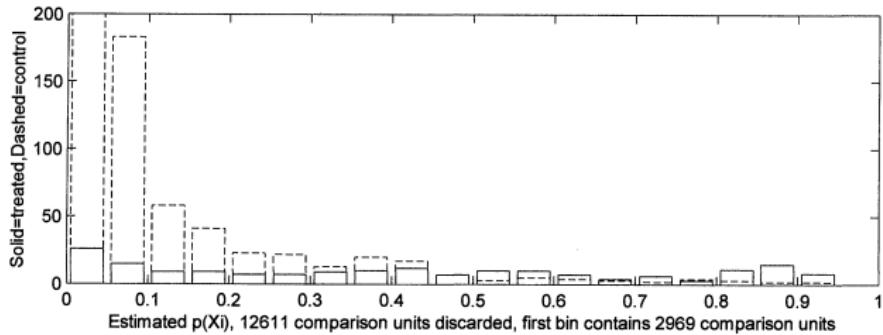


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45-.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Stratifying on the score			Matching on the score		
	(1) Unadjusted	(2) Adjusted ^a	Quadratic in score ^b (3)	(4) Unadjusted	(5) Adjusted	(6) Observations ^c	(7) Unadjusted	(8) Adjusted ^d
NSW	1,794 (633)	1,672 (638)						
PSID-1 ^e	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 ^f	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 ^f	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 ^g	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 ^g	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 ^g	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

^a Least squares regression: RE78 on a constant, a treatment indicator, age, age², education, no degree, black, Hispanic, RE74, RE75.

^b Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

^c Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

^d Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)].

Propensity scores are estimated using the logistic model, with specifications as follows:

^e PSID-1: Prob ($T_i = 1$) = F(age, age², education, education², married, no degree, black, Hispanic, RE74, RE75, RE74², RE75², u74*black).

^f PSID-2 and PSID-3: Prob ($T_i = 1$) = F(age, age², education, education², no degree, married, black, Hispanic, RE74, RE74², RE75², u74, u75).

^g CPS-1, CPS-2, and CPS-3: Prob ($T_i = 1$) = F(age, age², education, education², no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age³).

Covariate balance

Table 4. Sample Means of Characteristics for Matched Control Samples

<i>Matched samples</i>	<i>No. of observations</i>	<i>Age</i>	<i>Education</i>	<i>Black</i>	<i>Hispanic</i>	<i>No degree</i>	<i>Married</i>	<i>RE74 (U.S. \$)</i>	<i>RE75 (U.S. \$)</i>
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.

MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

Estimation in Stata

- I have written up code that will implement IPW on the DW data
- It's nonparametric, so it doesn't use any packages
- But you are welcome to try some packages, particularly the -teffects- command

Kernel matching

- Alternatively we can perform propensity score matching with a kernel-based method.
- Notice on the next slide that the estimate of the ATT switches sign relative to that produced by the NN matching algorithm

Stata syntax

```
psmatch2 treated , pscore(score) outcome(re78)
    kernel k(normal) bw(0.01)
pstest2 age black hispanic married educ nodegree
    re78 , sum graph
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	5976.35202	21553.9209	-15577.5689	913.328457	-17.06
	ATT	5976.35202	6882.18396	-905.831935	2151.26377	-0.42

Note: S.E. does not take into account that the propensity score is estimated.

	psmatch2:
psmatch2:	Common
Treatment	support
assignment	On support Total
	+-----+-----+
Untreated	2,490 2,490
Treated	297 297
	+-----+-----+
Total	2,787 2,787

Variable	Sample	Mean		%reduct		t-test	
		Treated	Control	%bias	bias	t	p> t
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	24.626	24.572	0.6	99.5	0.09	0.926
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.80135	.81763	-3.9	97.0	-0.50	0.614
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09428	.08306	4.6	81.8	0.48	0.631
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.16835	.1439	6.8	96.5	0.82	0.413
education	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.38	10.238	5.6	91.8	0.81	0.415
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.73064	.72101	2.1	97.7	0.26	0.793
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3066.1	3905.8	-8.2	94.8	-1.99	0.047

Matchings vs. Propensity score

Table 2. Experimental and nonexperimental estimates for the NSW data

	$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Panel A:										
Experimental estimates										
Covariate matching	1.22	(0.84)	1.99	(0.74)	1.75	(0.74)	2.20	(0.70)	1.79	(0.67)
Bias-adjusted cov matching	1.16	(0.84)	1.84	(0.74)	1.54	(0.75)	1.74	(0.71)	1.72	(0.68)
Pscore matching	1.43	(0.81)	1.95	(0.69)	1.85	(0.69)	1.85	(0.68)	1.79	(0.67)
Bias-adjusted pscore matching	1.22	(0.81)	1.89	(0.71)	1.78	(0.70)	1.67	(0.69)	1.72	(0.68)
Regression estimates										
Mean difference	1.79	(0.67)								
Linear	1.72	(0.68)								
Quadratic	2.27	(0.80)								
Weighting on pscore	1.79	(0.67)								
Weighting and linear regression	1.69	(0.66)								
Panel B:										
Nonexperimental estimates										
Simple matching	2.07	(1.13)	1.62	(0.91)	0.47	(0.85)	-0.11	(0.75)	-15.20	(0.61)
Bias-adjusted matching	2.42	(1.13)	2.51	(0.90)	2.48	(0.83)	2.26	(0.71)	0.84	(0.63)
Pscore matching	2.32	(1.21)	2.06	(1.01)	0.79	(1.25)	-0.18	(0.92)	-1.55	(0.80)
Bias-adjusted pscore matching	3.10	(1.21)	2.61	(1.03)	2.37	(1.28)	2.32	(0.94)	2.00	(0.84)
Regression estimates										
Mean difference	-15.20	(0.66)								
Linear	0.84	(0.88)								
Quadratic	3.26	(1.04)								
Weighting on pscore	1.77	(0.67)								
Weighting and linear regression	1.65	(0.66)								

NOTE: The outcome is earnings in 1978 in thousands of dollars.

Subsequent studies

- Heckman et al. (1996, 1998) used experimental data from the US National Job Training Partnership Act (JTPA)
- They conclude that in order for matching estimators to have low bias, it is important that the data include a rich set of variables related to program participation and labor market outcomes, that the nonexperimental comparison group be drawn from the same local labor markets as the participants and the dependent variable (typically earnings) be measured in the same way for participants and nonparticipants
- All three of these conditions fail to hold in DW (1999, 2002) according to Smith and Todd (2005)

Smith and Todd

- Difference-in-differences with propensity scores tended to work well in Smith and Todd (2005)
- But hard to make this a rule, because it's hard to know ex ante if we've specified the propensity score correctly (i.e., have CIA)
- It is vital you know your data, if you're going to use these methods, which means understanding at a deep level the way in which selection (i.e., treatment assignment) works in your data

Beating a dead horse

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

Coarsened exact matching

- There are two kinds of matching as we've said
 - ① *Exact matching* matches a treated unit to all of the control units with the same covariate value. Sometimes this is impossible (e.g., continuous covariate).
 - ② *Approximate matching* specifies a metric to find control units that are close to the treated unit. Requires a distance metric, such as Euclidean, Mahalanobis, or the propensity score. All of which can be implemented in Stata's teffects.
- Iacus, King and Porro (2011) propose another version of matching they call coarsened exact matching (CEM). Some big picture ideas

Checking imbalance

- Iacus, King and Porro (2008) say that in practice approximate matching requires setting the matching solution beforehand, then checking for imbalance after.
- Start over, repeat, until the user is exhausted by checking for imbalance.

CEM Algorithm

- ① Begin with covariates X . Make a copy called X^*
- ② Coarsen X^* according to user-defined cutpoints or CEM's automatic binning algorithm
 - Schooling → less than high school, high school, some college, college, post college
- ③ Create one stratum per unique observation of X^* and place each observation in a stratum
- ④ Assign these strata to the original data, X , and drop any observation whose stratum doesn't contain at least one treated and control unit

You then add weights for stratum size and analyze without matching.

Tradeoffs

- Larger bins mean more coarsening. This results in fewer strata.
- Fewer strata result in more diverse observations within the same strata and thus higher imbalance
- CEM prunes both treatment and control group units, which changes the parameter of interest. Be transparent about this as you're not estimating the ATE or the ATT when you start pruning

Benefits

- The key benefit of CEM is that it is in a class of matching methods called *monotonic imbalance bounding*
- MIB methods bound the maximum imbalance in some feature of the empirical distributions by an ex ante decision by the user
- In CEM, this ex ante choice is the coarsening decision
- By choosing the coarsening beforehand, users can control the amount of imbalance in the matching solution
- It's also wicked fast.

Imbalance

- There are several ways of measuring imbalance, but here we focus on the $\mathcal{L}_1(f, g)$ measure which is

$$\mathcal{L}_1(f, g) = \frac{1}{2} \sum_{I_1 \dots I_k} |f_{I_1 \dots I_k} - g_{I_1 \dots I_k}|$$

where the f and g record the relative frequencies for the treatment and control group units.

- Perfect global imbalance is indicated by $\mathcal{L}_1 = 0$. Larger values indicate larger imbalance between the groups, with a maximum of $\mathcal{L}_1 = 1$.

Stata

- Download `cem` from Stata: `ssc install cem, replace`
- You will automatically compute the global imbalance measure, as well as several unidimensional measures of imbalance, when using `cem`
- I got a $\mathcal{L}_1 = 0.55$. What does it mean?
 - By itself, it's meaningless. It's a reference point between matching solutions.
 - Once we have a matching solution, we will compare its \mathcal{L}_1 to 0.55 and gauge the increase in balance due to the matching solution from that difference.
 - Thus \mathcal{L}_1 works for imbalance as R^2 works for model fit: the absolute values mean less than comparisons between matching solutions.

More Stata

- Because `cem` bounds the imbalance *ex ante*, the most important information in the Stata output is the number of observations matched.
- You can also choose the coarsening as opposed to relying on the algorithm's automated binning.
- Once you have estimated the strata, you regress the outcome onto the treatment and then weight the regression by `cem_weights`. For instance,

```
regress re78 treat [iweight=cem_weight]
```

- For more on this, see Blackwell, et al. Stata journal article from 2009.

The credibility revolution won

- People like Heckman, Rubin, Ashenfelter, LaLonde, Angrist, Krueger, Card, Imbens, Athey, Duflo, Abadie and many others built on their backs a movement of sorts
- The movement tried to shift economists and other social scientists away from naive empirical methods that couldn't hope to estimate behavioral causal parameters towards things that might
- It's in your best interest to study empirical methods, papers that use them, how they communicate their findings and the econometricians so that you can be ready when the opportunity arises

Make the stone stoney again

- A man walks up the mountain barefoot til he can't feel his feet again – Victor Schlosskey said art is there to make “the stone feel like a stone again”. I want research to feel like research again for you
- Research is a quest for honest answers to good faith questions that people care about
- Most of all, research is truly fun for those who find such things fun. It's a form of self-expression and creativity for many of us
- And it is fun to understand the answers you get and why those answers are reliable which requires checklists, workflows, clearly defined assumptions and proper tools for the job
- It is not fun to get a bad answer to a poorly defined question that you're not confident about

A Priori Knowledge is Necessary for Identification

- Think hard about these questions:
 - Can you write down a DAG or otherwise model the data generating process?
 - What parameter do you think is interesting (e.g., ATT, LATE)
 - What are the assumptions needed for identifying that parameter?
- Pick estimators based on these questions, not the other way around
- Less so, pick data based on these questions, not the other way around (usually)

What's a good research design?

- A good research design is one you are excited to tell people about – that's basically what characterizes *all* research designs, whether propensity score matching or regression discontinuity designs
- Don't get enamored by statistical modeling that obscures the identification problem from plain sight.
- Always understand what assumptions you *must* make, be clear which parameters you are and are not identifying
- Good research designs help you believe and not be afraid of your answers

What's the reason for your work?

- Causal identification is a necessary but not a sufficient condition for publishing well these days bc the credibility revolution won
- Must also be an “interesting” question - admittedly subjective
- If it must be interesting, then the best thing you can do for yourself is choose a topic that you care about
- Publishing is simply too difficult to be working on something you find trivial

Free disposal advice

- My colleague said “a good study and a bad study take the same amount of time” – don’t work on stuff just to work on it
- Finding projects with upside, in a set of potential projects, is a good idea
- The sooner you can cut bait on a bad project and move on, the better – beware the sunk cost fallacy
- For my personality, questions are practically existential quests for the meaning of life, but not everyone needs extreme incentives
- So know yourself, work to your strengths, figure out things that downplay your weaknesses, believe in yourself, find your sponsors and mentors, seek help