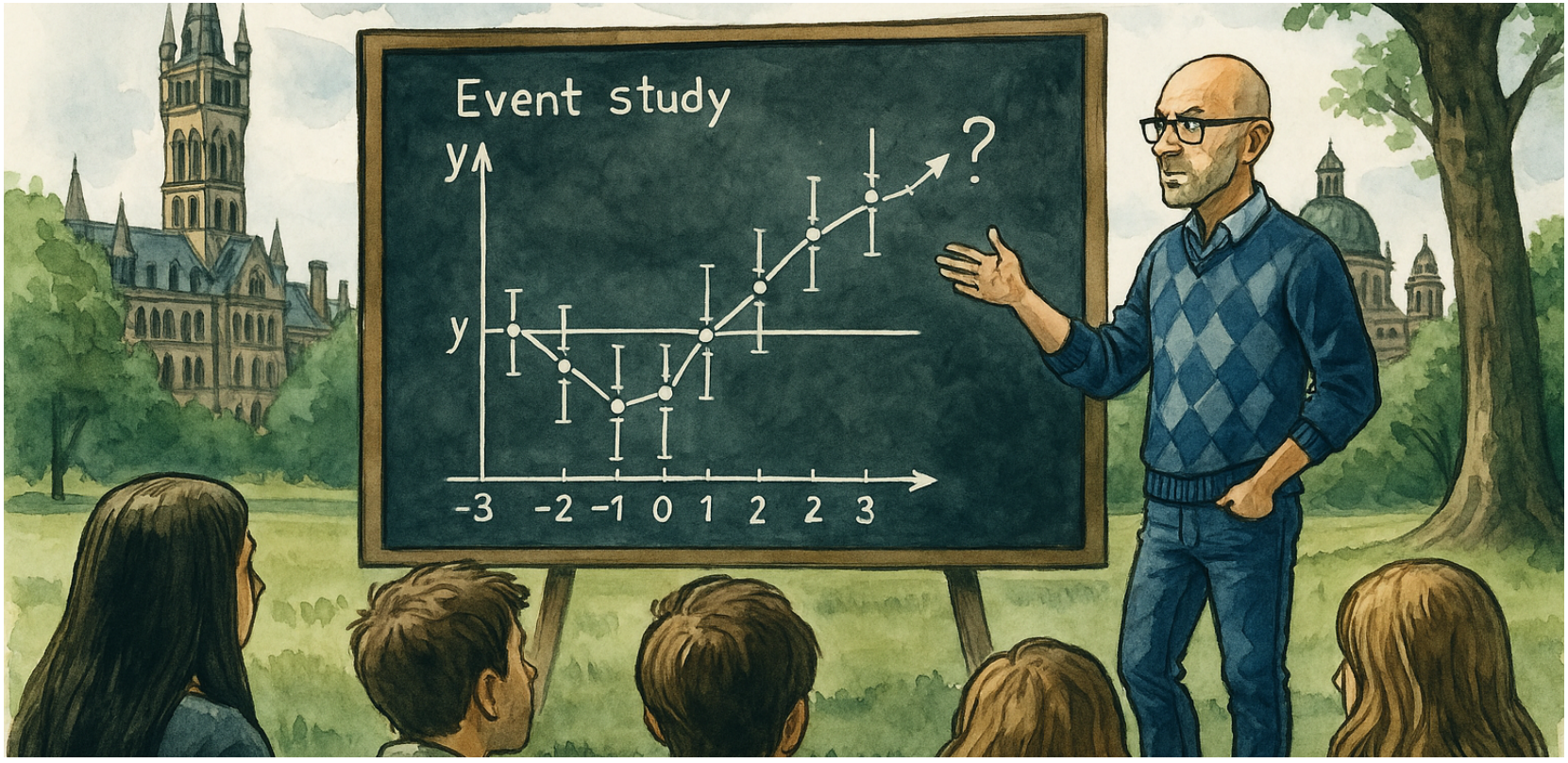
University of Glasgow

2025



Roadmap

Introduction

- Welcome to a week of advanced panel methods at the University of Glasgow!
- I'm Scott Cunningham, professor of economics at Baylor University (visiting this next year at Harvard University)
- Five days of difference-in-differences and other panel estimators!

What my pedagogy is like

- Workshop is intended to take someone from knowing nothing about difference-in-differences to a broad level of competency on advanced topics
- My stuff will mix basic things with subtle interpretation, my own subjective beliefs about ”good science” (but not necessarily good papers), and developments
- Lecture, discussion, exercises, application, coding along the way
- Ask questions at any point; I’ll do my best to answer them

Class goals

Pedagogical goal is to explain diff-in-diff and other panel estimators, but as a secondary objective to help foster in you:

1. **Confidence**: You will feel like you have a good enough understanding of diff-in-diff, both in its basics and some more contemporary issues, so that it seems like an intuitive and useful tool you could imagine using
2. **Comprehension**: You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation in the diff-in-diff
3. **Competency**: You will have more knowledge of programming syntax in Stata and R so that later you can apply this in your own work

Day 1 outline

Basics of difference-in-differences (Beginners but also my own personal take on this)

- The Core –
  - → what is difference-in-differences, how do you calculate it, what does it mean, what are its assumptions
  - → Why do we weight?
  - → Four Means or One Regression?
- Testing for parallel trends – event studies, falsifications
- Violations of parallel trends and solutions:
  - → Compositional change and corrections
  - → Triple differences
  - → Handling covariates with regression adjustment, IPW, double robust vs traditional OLS specifications

Day 2 outline

- Covariates if not done
- Traditional fixed effects regressions and Bacon decomposition
- Callaway and Sant'Anna (2021) for differential timing

Day 3 outline

- Callaway and Sant'Anna (2021) continued
- CS in action with an example from my own work
- Decomposition event study leads and lags under staggered adoption (Sun and Abraham 2021)

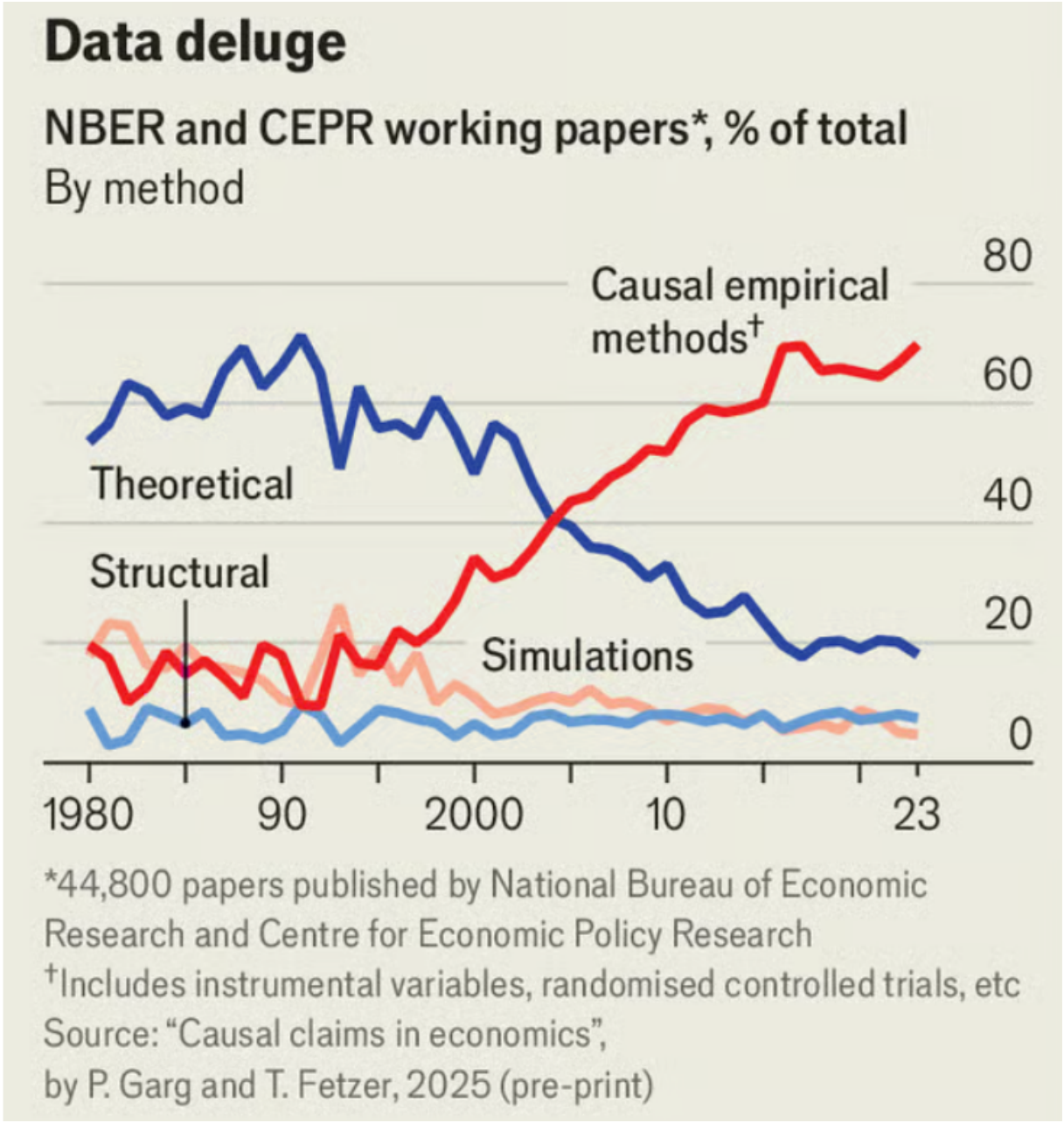
Day 4 outline

- de Chaisemartin and D'Haultfoueille (2020)
- Imputation methods (e.g., Borusyak, et al. 2024)
- Honest diff-in-diff
- Continuous treatments in diff-in-diff framework

Day 5 outline

- Synthetic control (Abadie)
- Imbalanced synthetic control and bias correction
- Matrix completion with nuclear norm regularization
- TBD

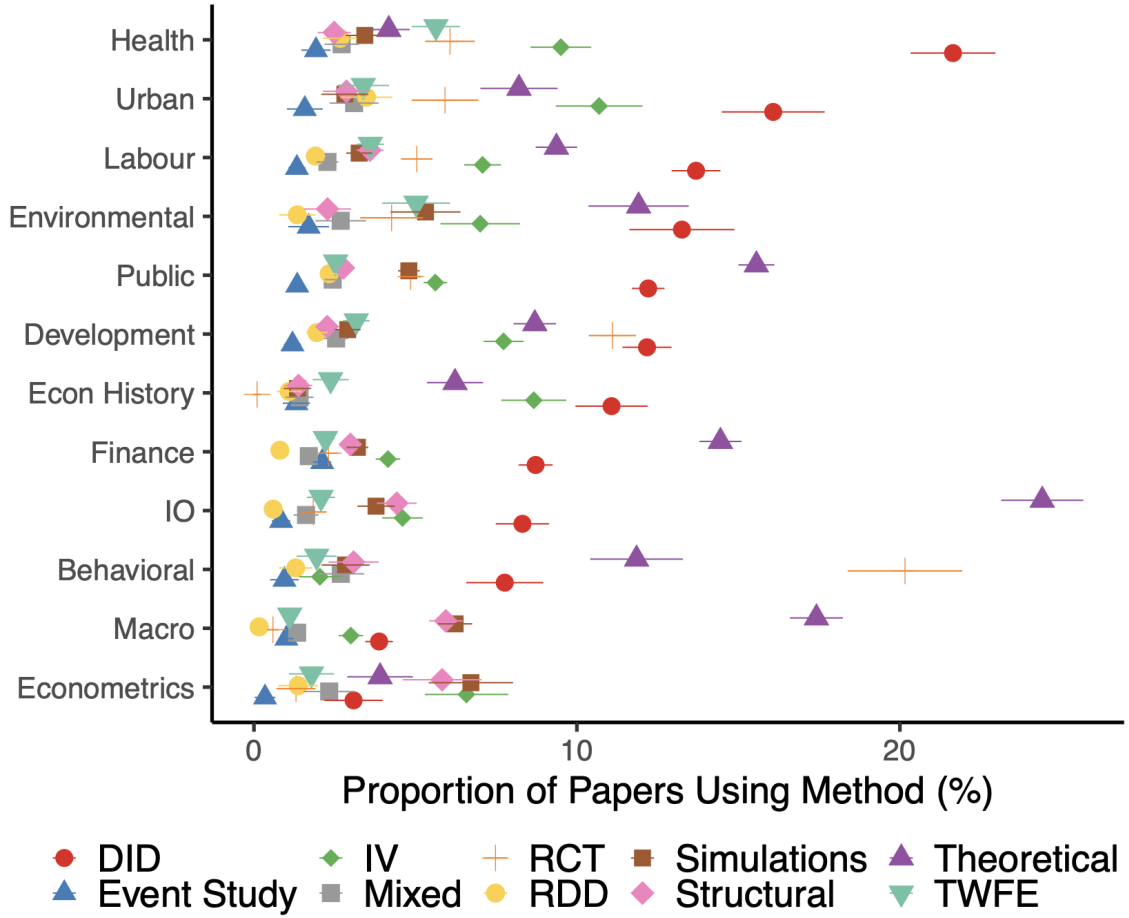
Causal Claims in Economics



Diff-in-Diff in the Cross-Section by Field

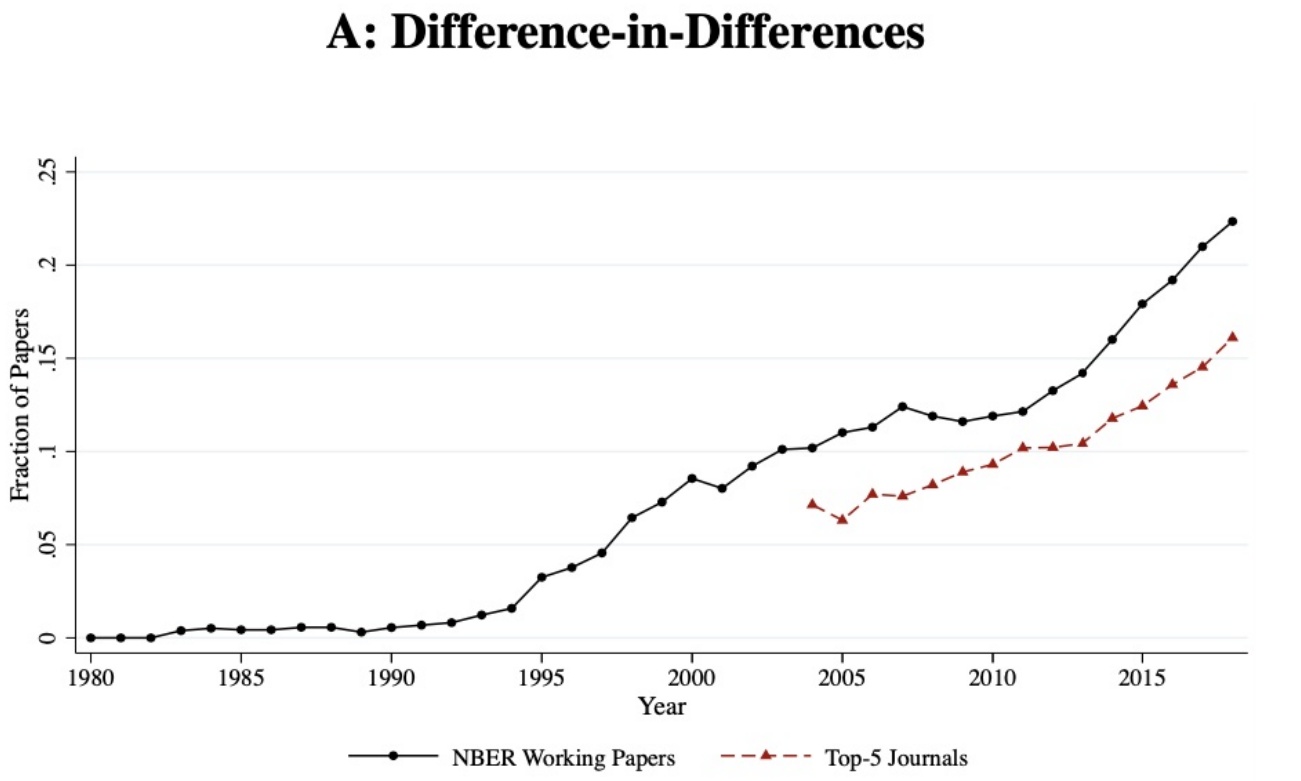
Figure: Garg and Fetzer (2025)

Figure 5: Cross-Sectional Breakdown of Empirical Methods by Field in NBER and CEPR Working Papers



Diff-in-Diff Over Time

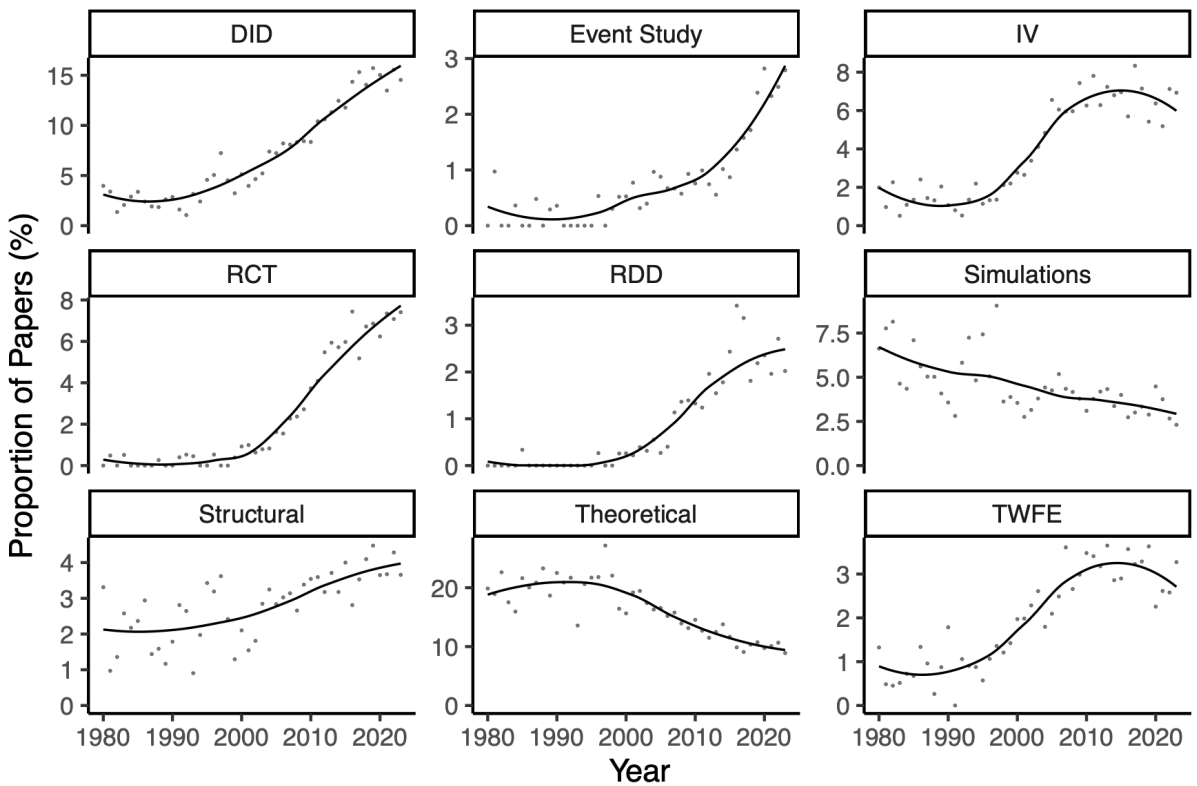
Figure: Currie, et al. (2020)



Diff-in-Diff Compared to Others Over Time

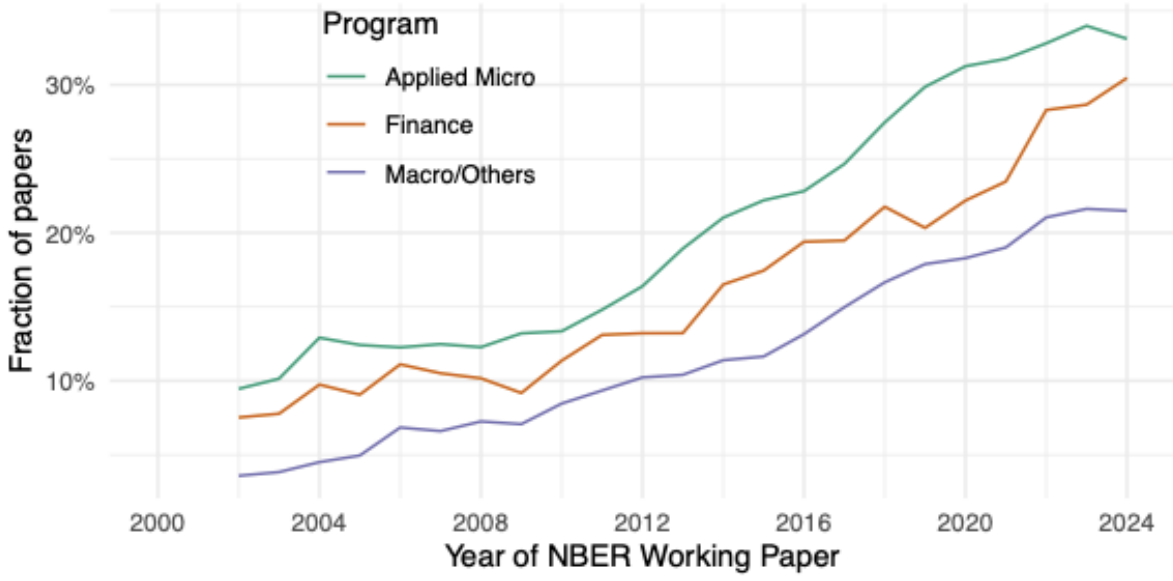
Figure: Garg and Fetzer (2025)

Figure 4: Proliferation of Empirical Methods Over Time in NBER and CEPR Working Papers



Diff-in-Diff by Field

Figure: Goldsmith-Pinkham (2024)



(a) Difference-in-differences

What is difference-in-differences

- DiD is when a group of units are assigned some treatment and then compared to a group of units that weren't before and after
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Predates the randomized experiment by 80 years, but uses basic experimental ideas about treatment and control groups (just not randomized)
- Uses panel or repeated cross section datasets, binary treatments usually, and often covariates

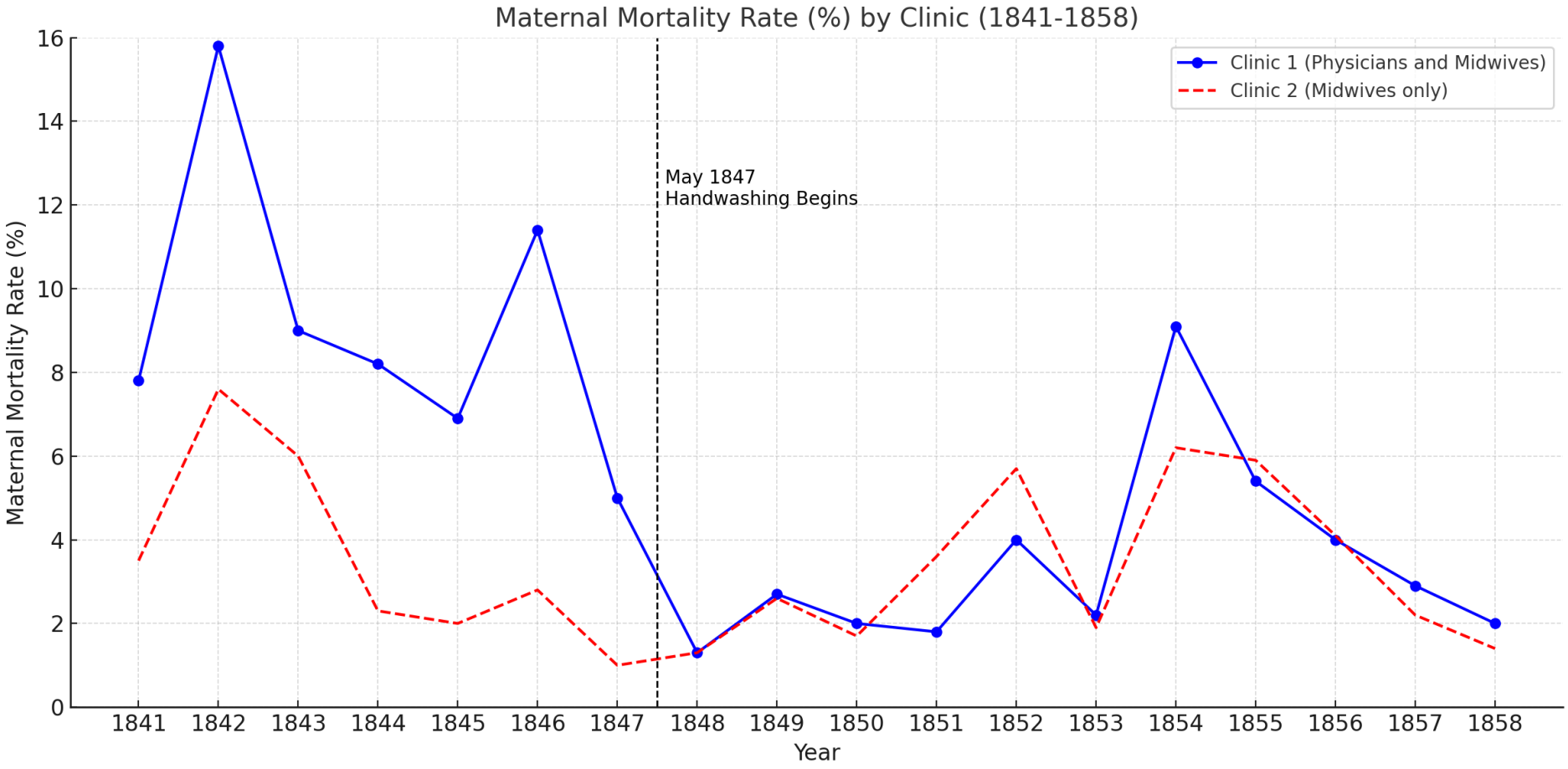
Ignaz Semmelweis and washing hands

- Early 1820s, Vienna passed legislation requiring that if a pregnant women giving birth went to a public hospital (free care), then depending on the day of week and time of day, she would be routed to either the midwife wing or the physician wing (most likely resulting in random assignment)
- But by the 1840s, Ignaz Semmelweis noticed that pregnant women died after delivery in the (male) wing at a rate of 13-18%, but only 3% in the (female) midwife wing – cause was puerperal or “childbed” fever
- Somehow this was also we known – women would give birth in the street rather than go to the physician if they were unlucky enough to have their water break on the wrong day and time

Ignaz Semmelweis and washing hands

- Ignaz Semmelweis conjectures after a lot of observation that the cause is the teaching faculty teaching anatomy using cadavers and then delivering babies *without washing hands*
- New training happens to one but not the other and Semmelweis thinks the mortality is caused by working with cadavers
- Convinced the hospital to have physicians wash their hands in chlorine but not the midwives, creating a type of difference-in-differences design

Semmelweis diff-in-diff evidence



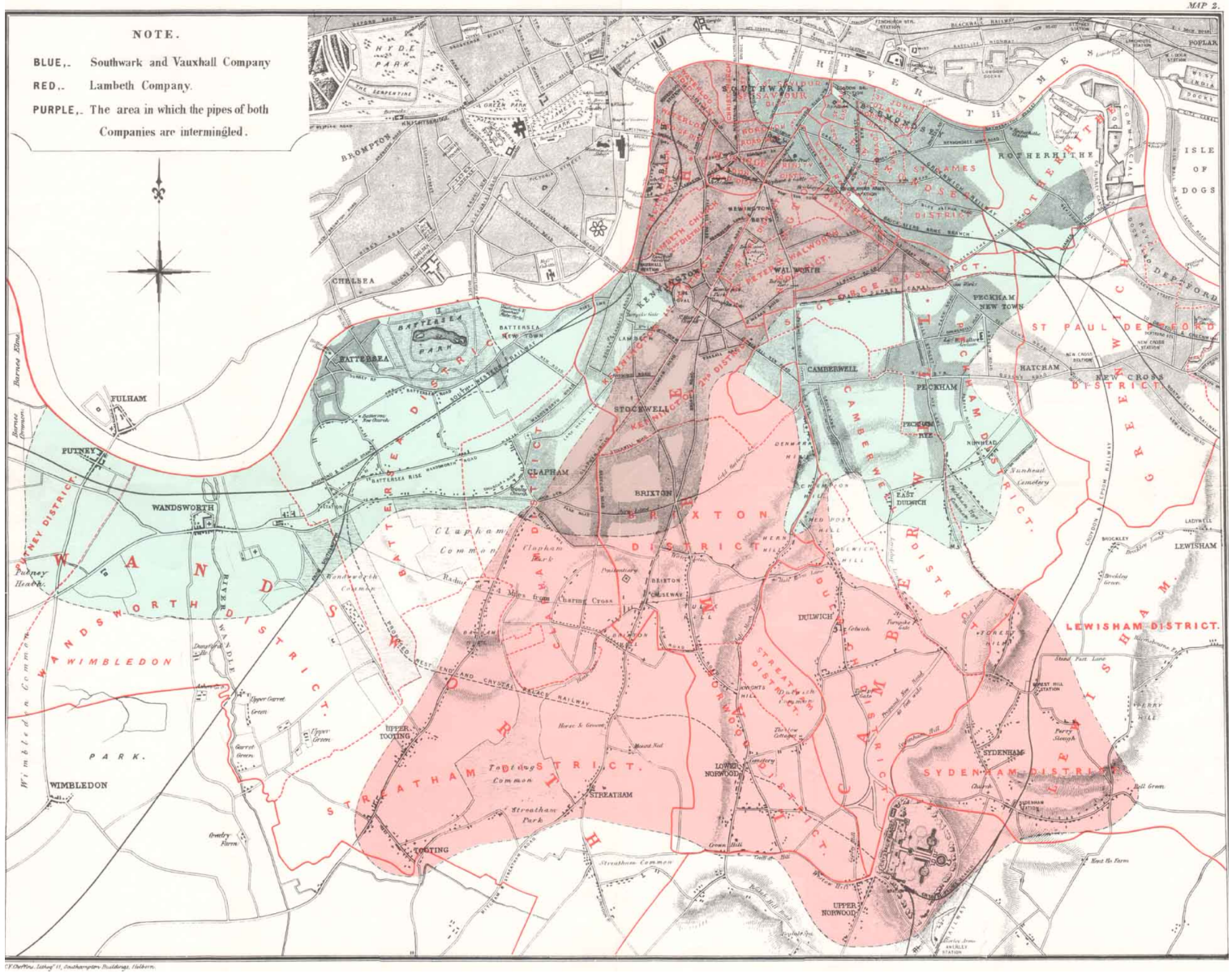
Evidence Rejected

- Diff-in-diff evidence was rejected by Semmelweis' superiors claiming it was the hospital's new ventilation system
- Dominant theory of disease spread was caused by "odors" or miasma or "humors"
- Semmelweis began showing signs of irritability, perhaps onset of dementia, became publicly abusive, was committed to a mental hospital and within two weeks died from wounds he received while in residence
- Despite the strength of evidence, difference-in-differences was rejected – a theme we will see continue

John Snow and cholera

- Three major waves of cholera in the early to mid 1800s in London, largely thought to be spread by miasma (“dirty air”)
- John Snow believed cholera was spread through the Thames water supply through an invisible creature that entered the body through food and drink, caused the body to expel water, placing the creature back in the Thames and causing epidemic waves
- London passes ordinance requiring water utility companies to move inlet pipe further up the Thames, above the city center, but not everyone complies
- Natural experiment: Lambeth water company moves its pipe between 1849 and 1854; Southwark and Vauxhall water company delayed

Figure: Two water utility companies in London 1854



Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	**Time**	**Outcome**	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + L_t + D$	$L_t + D$	
				$D + (L_t - SV_t)$
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + SV_t$	SV_t	

$$\widehat{\delta}_{did} = D + (L_t - SV_t)$$

This method yields an unbiased estimate of D if $L_t = SV_t$, but note that L_t is a counterfactual trend and therefore not known

Roadmap

- Introduction
  - Managing expectations
  - Diff-in-Diff Popularity
  - Origins of diff-in-diff in public health
- Diff-in-Diff Fundamentals
  - Average Treatment Effect on the Treated
  - Core Diff-in-Diff Assumptions
  - Clearly Defined Control Status
  - Data Requirements
- Weighting and Target Parameters
- Four Means or One Regression

Introducing Potential Outcomes to DiD

- Research question versus causal question – not the same thing
  - → Research question would be you are wanting to know effect of job training programs on earnings
  - → Causal question is expressed using potential outcomes
- Causal questions are usually averages of individual treatment effects for a specific population of units

Identification vs Estimation

- We must start by making a distinction between the parameter we are attempting to identify and the manner in which we will estimate it
- Identification requires first stating explicitly our goal expressed using potential outcomes
- But often people skip this step and go directly to the 2×2 calculation

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if in job training program } t \\ 0 \text{ if not in job training program at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1\text{: wages at time } t \text{ if trained} \\ 0\text{: wages at time } t \text{ if not trained} \end{cases}$$

  where j indexes a state of the world where the treatment happened or did not happen

Treatment effect definitions

Individual treatment effect

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

Missing data problem: No data on the counterfactual

Average Treatment Effects for the Treated Subpopulation

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:

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

It's the average causal effect but only for the people exposed to some intervention; notice we can't calculate it, also, because we are missing the red term

ATT vs ATE

- Potential outcomes is subtle and easily people are overconfident about its interpretation (probably because counterfactuals are seemingly easy to understand)
- Let's look at the "Potential Outcomes" tab at our Diff-in-Diff Worksheet; fill out all the instructions
- https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing

Deriving identification assumptions

- Diff-in-diff is two things
    1. It's always a calculation (i.e., four averages)
    2. It sometimes has a causal interpretation
- I'm going to walk us through two main assumptions using potential outcomes and algebra
- Then we will look at what happens when our control group has been treated (not an assumption, more like a promise)

DiD equation is the 2x2

- The building block of diff-in-diff calculations is “four averages and three subtractions”
- Needs two groups, two time periods, which is four averages
- Often called the 2×2 for that reason (Goodman-Bacon 2021)

$$\widehat{\delta} = \Big(E[Y_k|Post] - E[Y_k|Pre] \Big) - \Big(E[Y_U|Post] - E[Y_U|Pre] \Big)$$

k are the people in the job training program, U are the untreated people not in the program, $Post$ is after the trainees took the class, Pre is the period just before they took the class, and $E[y]$ is mean earnings.

- See https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing

Replace with potential outcomes and add a zero

$$\widehat{\delta} = \underbrace{\Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)}_{\text{Switching equation}}$$

$$+ \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

Identification through parallel trends

Parallel trends

Assume two groups, treated and comparison group, then we define parallel trends as:

$$E(\Delta Y_k^0) = E(\Delta Y_U^0)$$

In words: "The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees".

It's in red because parallel trends is untestable and critically important to estimation of the ATT using any method, OLS or "four averages and three subtractions"

Don’t Need More Than Two Periods

- Notice that diff-in-diff identifies the ATT *with only* two time periods
- Diff-in-diff does not ”need” a long pre-treatment time series (unlike synthetic control which does) – two periods and two groups is enough
- But there is another assumption aside from parallel trends we want to learn

No Anticipation

- First interpretation of No Anticipation (NA) is this:

$$Y_{i,t-1} = Y^0_{i,t-1}$$

- Future treatments do not spill over to the past – that is, a unit is not treated until the day it is treated
- A better word might be ”no pre-treatment”, rather than ”no anticipation” because ”anticipation” is a behavior and that confuses things
- What if a government announces a minimum wage increase that will take place in one year?

No Anticipation

- Second interpretation of NA means this:

$$Y^1_{i,t-1} - Y^0_{i,t-1} = 0$$

  which means ”all pre-treatment treatment effects are zero.
- This means even if future minimum wage increases were known, their pre-treatment treatment effects were zero
- So no – knowing that something is going to happen does not automatically violate NA (but it absolutely could)
- Let’s formalize what happens when NA is violated

No Anticipation Violation

$$\widehat{\delta} = \Big(E[Y_k^1|Post] - E[Y_k^1|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)$$

What if the k group had been treated at the baseline "pre" period in our 2x2?

Add in **two zeroes** instead of one, substitute and rearrange.

$$+ \quad E[Y_k^0|Post] - E[Y_k^0|Post]$$
$$+ \quad E[Y_k^0|Pre] - E[Y_k^0|Pre]$$

No Anticipation Violation

If the baseline period is treated, then the simple 2x2 identifies the following three terms:

$$\begin{aligned} \delta \quad = \quad & ATT_k(Post) \\ & + \text{Non PT bias} \\ & - ATT_k(Pre) \end{aligned}$$

First row is the ATT in the post period; middle row is parallel trends; third row subtracts the baseline ATT from the calculation. If treatment effects are constant, then the DiD coefficient will be zero despite positive treatment effects. Let's look in `na.do`.

Do not use already treated controls

Using already treated units as controls is bad practice and here's why:

$$\widehat{\delta} = \Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^1|Post] - E[Y_U^1|Pre]\Big)$$

What if the U group had always been treated in both periods? Is parallel trends enough to identify the ATT?

Add in **three zeroes** instead of one, substitute and rearrange.

$$\begin{aligned} &+ \quad E[Y_k^0|Post] - E[Y_k^0|Post] \\ &+ \quad E[Y_U^0|Post] - E[Y_U^0|Post] \\ &+ \quad E[Y_U^0|Pre] - E[Y_U^0|Pre] \end{aligned}$$

Already Treated Control Group

If the baseline period is treated, then the simple 2x2 identifies the sum of the following three terms:

$$\begin{aligned} \delta &= ATT_k(Post) \\ &\quad + \text{Non PT bias} \\ &\quad - \Delta ATT_U \end{aligned}$$

Again, first row is the target parameter, plus parallel trends term, minus the changing ATT in our control group

Clearly Defined Control Status

Look again at the result when you have NA and no treated control group:

$$\widehat{\delta} = \left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)$$

What does the zero mean exactly? At baseline, before k was treated, it was in the Y^0 state of the world – that wasn't "not treated", it was rather a different treatment status. What was it? Whatever it was, group U had the same one

Clearly Defined Control Status

Make our substitutions

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\left[E[Y_k^0|Post] - E[Y_k^0|Pre]\right] - \left[E[Y_U^0|Post] - E[Y_U^0|Pre]\right]}_{\text{Non-parallel trends bias in 2x2 case}}$$

1. Notice that embedded in parallel trends is NA, untreated controls, and clearly defined control status.
2. Notice that your target parameter is based on the same Y^0 as the Y^0 at baseline *and* a clearly defined control state

Clearly Defined Control Status

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

- Your control group is untreated Y^0 in both periods and so under parallel trends will be a valid *imputation*
- But notice the role of Y_U^0 – parallel trends, clearly defined control state, NA – is complex

Example: Clearly Defined Control Status

- United States has prohibited cannabis use (e.g., Texas), medical marijuana (e.g., Oklahoma) and recreational cannabis (e.g., California)
- Researcher wants to know the effect of *recreational marijuana* on opiate addiction, but relative to what?

$$\delta_i = Y_i^1 - Y_i^0$$

- Note: All states with legalized recreational cannabis started with medical marijuana. Does that matter?

Different Target Parameters

1. $ATT_1 = E[Y_k^1|Pre] - E[Y_k^0|Post]$ where Y^0 is "treated with medical marijuana"
2. $ATT_2 = E[Y_k^1|Pre] - E[Y_k^0|Post]$ where Y^0 is "treated with prohibition"

Do you think these *must* be the same thing? Why/why not?

What data do you need for ATT_1 vs ATT_2?

Clearly Defined Control Status

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

- Diff-in-diff *can* estimate the effect of recreational marijuana relative to prohibition so long as states transitioned from medical marijuana to recreational marijuana. Why?

Clearly Defined Control Status

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

- Diff-in-diff *can* estimate the effect of recreational marijuana relative to prohibition so long as states transitioned from medical marijuana to recreational marijuana. Why?
- Because of clearly defined control status, NA and parallel trends which is imputing counterfactuals

Clearly Defined Control Status

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

- Diff-in-diff *can* estimate the effect of recreational marijuana relative to prohibition so long as states transitioned from medical marijuana to recreational marijuana. Why?
- Because of clearly defined control status, NA and parallel trends which is imputing counterfactuals
- Therefore diff-in-diff can only identify ATT_1 (i.e., recreational cannabis vs medical marijuana) but not ATT_2 (i.e., recreational cannabis vs prohibition) *in the US*

Target Parameter

- Matching and regression adjustment can technically estimate either ATT_1 or ATT_2 so long as its own assumptions hold (do you see why that is?)
- But diff-in-diff has a limitation that not all the other designs suffer from – the control group *must* have the same control status as the treatment group at baseline
- So you must decide ahead of time:
  1. What is your target parameter and why?
  2. Can you estimate it with diff-in-diff in your data? How do you know?

Summarizing the basics

- With two time periods (before and after treatment), and two groups (one treated and one not), then the 2×2 identifies the ATT if parallel trends and no anticipation holds so long as you do not have an already treated group as a control
- Let's now look at some basic data requirements and what happens when you don't meet those basic things

Longitudinal Data

- Diff-in-diff requires four means – pre and post for two groups
- Traditionally, the "pre" is a baseline mean at year just prior to treatment, $t-1$ or b depending on author
  - → Though sometimes you will see people present any interaction as diff-in-diff, we will focus only on time in our workshop
  - → Just remember the interaction regression we presented is calculating four means and three subtractions
  - → Interpreting parallel trends is a bit stranger otherwise

Longitudinal Data

- Two types of longitudinal data:
  - → Panel Data: same units tracked over time (e.g., National Longitudinal Survey of Youth 1997)
  - → Repeated Cross-Sections: different units sampled at each time (e.g., Census, Current Population Survey)
- Violations of parallel trends can arise differently across data types.

What is an Imbalanced Panel?

- Balanced panel is when all units observed in every period.
- Imbalanced panel is when units missing in some periods.
  - → Anthony is in waves 1-3,
  - → Bob is in waves 1-3,
  - → Inez is in waves 1 and 3 only,
  - → Dignan is in waves 1 and 2 only
- Missingness alone does not violate parallel trends, though it does change the parameter.

Example of Balanced ATE

Name	Year	Y0	Y1	Delta
Anthony	1	10	6	4
Anthony	2	12	7	5
Anthony	3	14	8	6
Bob	1	5	5	0
Bob	2	6	5	1
Bob	3	7	5	2
Inez	1	20	10	10
Inez	2	25	20	5
Inez	3	30	30	0
Dignan	1	5	0	5
Dignan	2	6	0	6
Dignan	3	7	0	7
			ATE	**4.25**

How does it change the parameter?

- Remember – in the potential outcomes framework, a treatment effect is defined at the individual level, δ_{it}
- So if you are missing a person, i, in a period, t, then it does not contribute
- The more heterogeneity in the treatment effects, the more the broken panel will shift away from what you think you're after

Imbalanced ATE

Let's work together on an example of the effect of imbalanced panels versus balancing imbalanced panels at https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing ("Balancing")

Missing at Random (MAR)

- Wooldridge (201) notes that missing at random (MAR) does not bias estimates under large samples
- This is because missingness is independent of $E[\Delta Y^0]$, therefore parallel trends will still hold in the large sample.
- This is somewhat testable, too, because if MAR holds, then baseline covariates for the $M = 1$ and $M = 0$ groups should on average be the same
- But maybe this is not plausible always so is there anything else we can do?

Conditional Missing at Random

- What if the missing is conditionally random

$$Y^0 \perp M \mid X$$

- This implies:

$$E[Y^0|M=1,X] = E[Y^0|M=0,X]$$

- So you can impute $\widehat{Y^0}$ for missing units using $Y^0 \sim X$ with your non-missing data with a regression estimator
- More needs to be done to think about the Y^1 imputation so I'll just focus on this for now

Conditional Missing at Random

- Note though – this is a ”missing based on observables” type of unconfoundedness assumption
- You will need an overlap assumption or a functional form assumption for regression-based imputation
- Should you assume it? It is technically weaker than assuming MAR, and imputing missing yet real Y seems far less problematic than imputing missing yet fictional Y^0
- But are you sure you know the drivers of missingness in your data?

Roadmap

Choosing Target Parameter

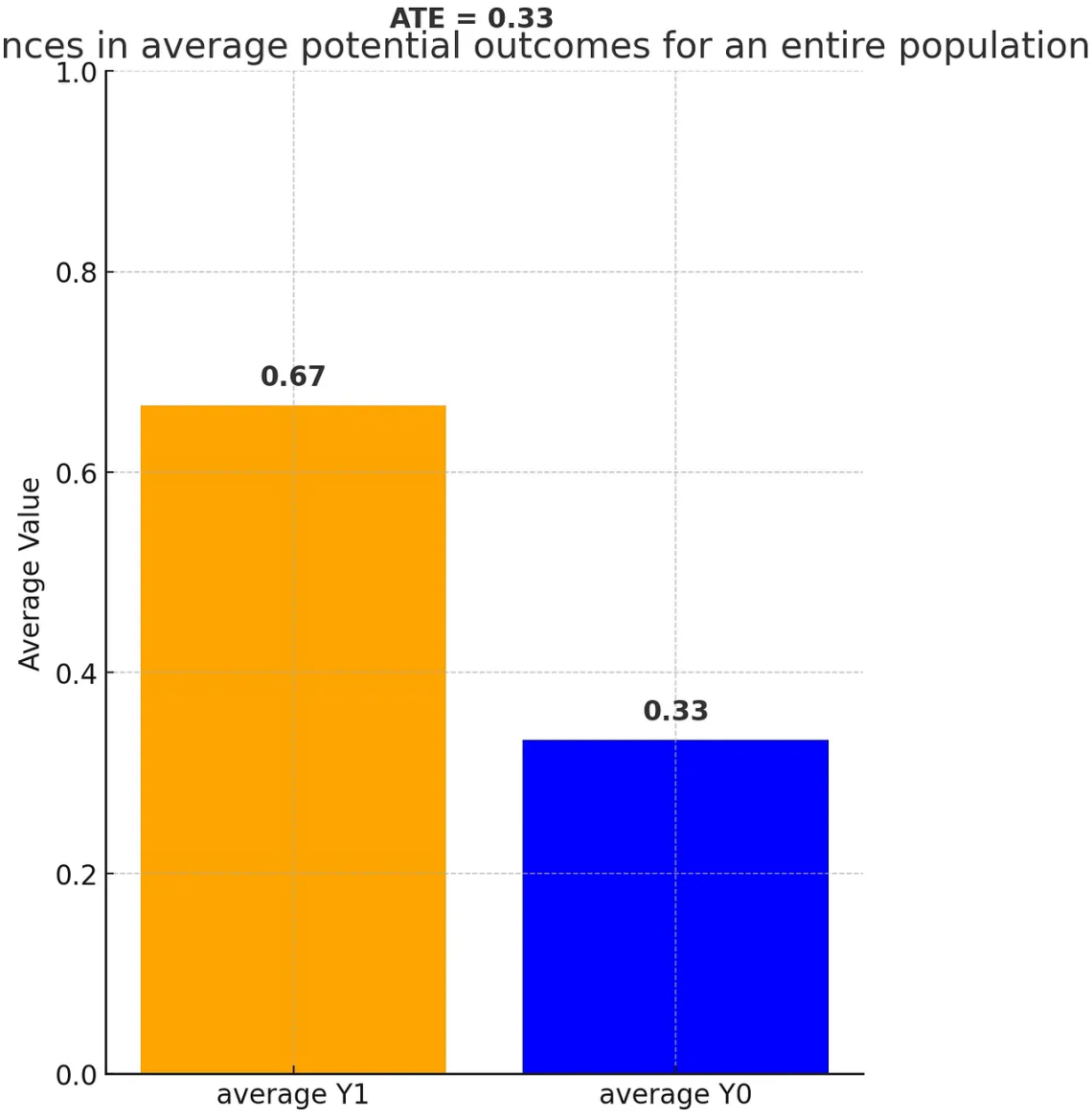
- Causal parameters are *averaged treatment effects* and that has two elements
  - → What population's treatment effects are you averaging
  - → What weight are you using to do that averaging?
- I'll use as an example a gun law in the United States called "concealed carry" or "right to carry" that lets people carry weapons "concealed" on their bodies or in their vehicles

Average Treatment Effects

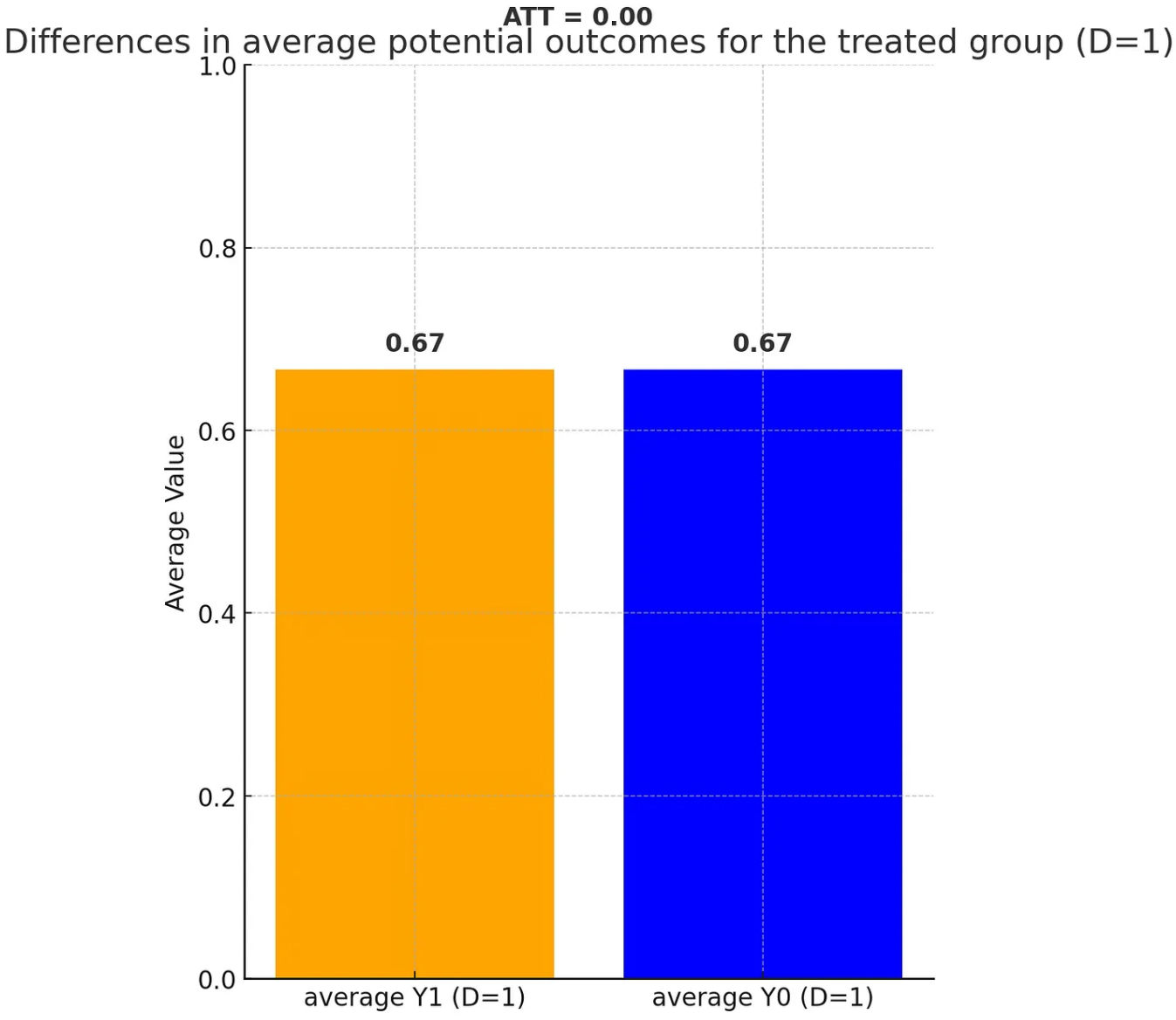
Table: Illustration of Potential Outcomes and Treatment Effects

Name	Y^1	Y^0	δ	**D**
Alan	1	0	1	1
Betty	0	1	-1	1
Chad	1	1	0	1
Daniel	0	0	0	0
Edith	1	0	1	0
Frank	1	0	1	0
$ATE =$			0.33	
$ATT =$			0	

Average Treatment Effects



Average Treatment Effect on the Treated



Which causal effect do you want?

- We know diff-in-diff identifies the ATT not the ATE but there are still more than one ATT!
- Solon, Haider and Wooldridge (2015), ”What are we weighting for?”
- Why you weight in surveys and why you weight in causal inference are different
  - → Survey weights are to make estimates nationally representative
  - → Population weighting in causal inference is because you want a different parameter
- How do we interpret adjustments made with population weights versus not?

Different Levels of Aggregation, Different Weights

Individuals	Y_1	Y_0	δ	County
Alan	1	0	1	1
Betty	1	1	0	1
Chad	1	1	0	1
Daniel	0	0	0	1
Edith	1	0	1	1
Frank	1	0	1	1
George	0	0	0	1
Hank	1	0	1	1
Ida	0	1	-1	2
Janet	0	1	-1	2

County	ATE_c
1	0.5
2	-1

ATE for average county	-0.25
ATE for average person	0.2

Weighting formula

- What if you have city or county level data but you want the ATE for the average person?
- Then that is when you weight by population – because you want to know the effect for the *average person*

$$ATE_{\text{people}} = \frac{\sum_c ATE_c \cdot N_c}{\sum_c N_c} \tag{1}$$

$$\begin{aligned} ATE_p &= \frac{(0.5 \times 8) + (-1 \times 2)}{8 + 2} \\ &= \frac{4 - 2}{10} \\ &= 0.2 \end{aligned}$$

Different Levels of Aggregation, Different Weights

- Heterogenous treatment effects is causing this, but so is weighting
- Consider Texas
  - → Texas has 31 million residents
  - → Texas 254 counties
- Where do they live?
  - → 13 million live in Harris, Dallas, Fort Worth, San Antonio and Austin, or rather 41%
- What if concealed carry increases firearm deaths in cities, but reduces them in counties, because of sorting by treatment effects?

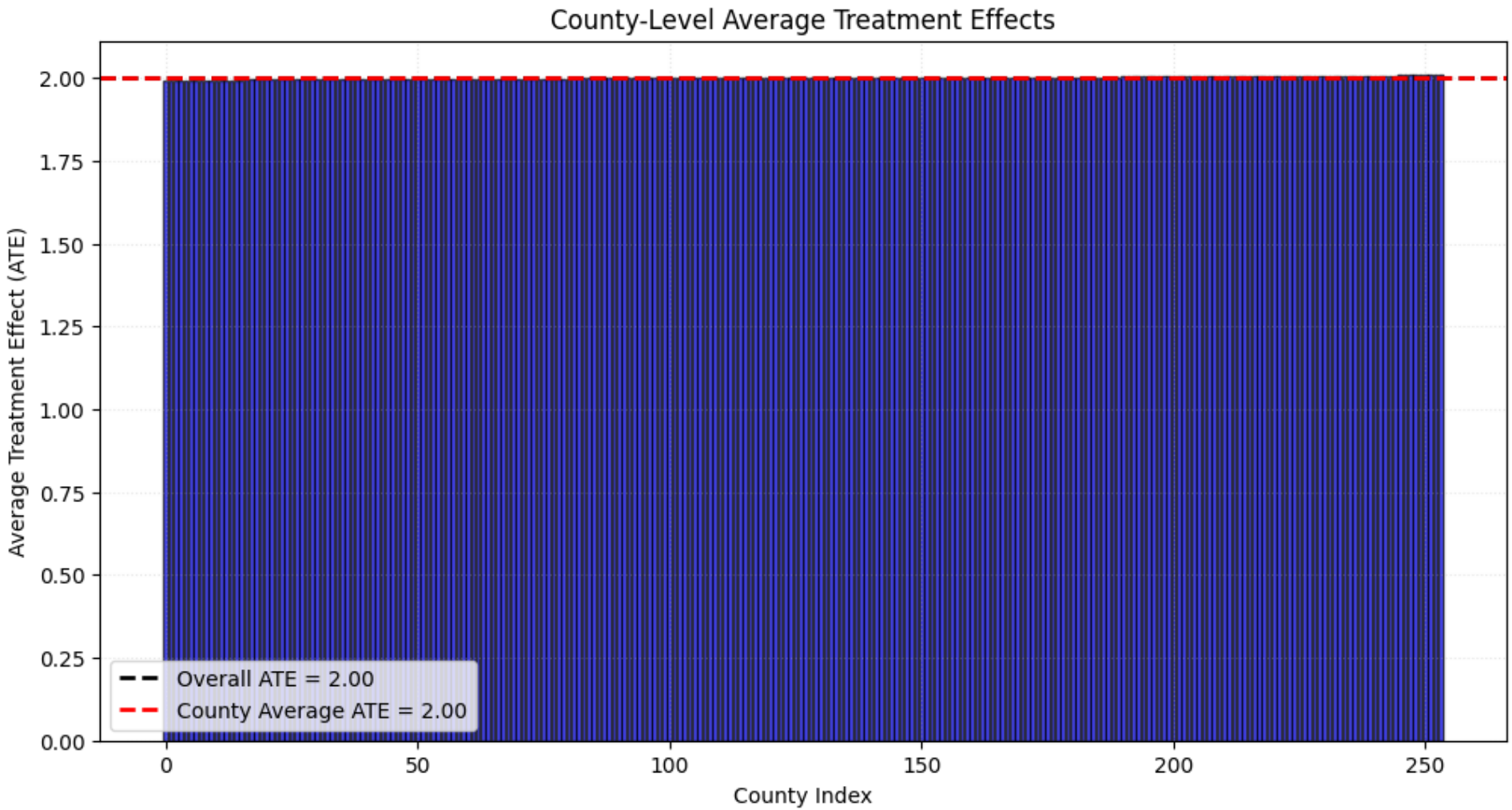
Simulation

- Assume a state with 30 million people and 254 counties
  - → 15 million live in 5 counties
  - → 15 million live equally spread in the other 249 counties (around 60,000 each)
- Assume that δ_i varies, sometimes positive and sometimes negative and $E[Y^1 - Y^0] = 2$

Selection into Counties is Random

- People choose where to live in Texas, but the mechanism by which they do so will have implications for our datasets
- What if they sort into counties (i.e., where they live) by lottery
- Every county will therefore have the same distribution of Y^1 and Y^0
- Every county will have an ATE of 2

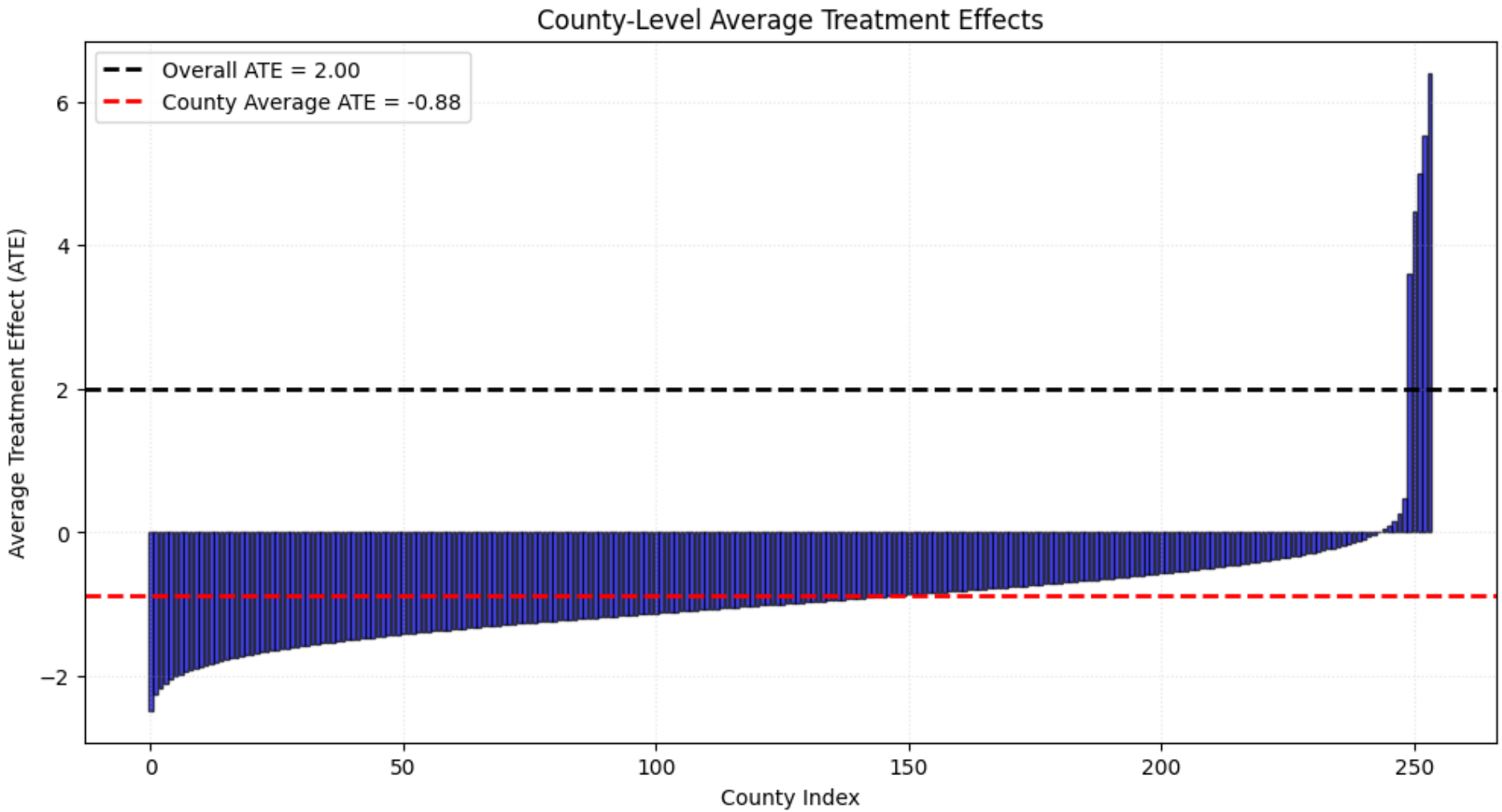
Average County ATE and Overall ATE Are the Same



Selection into Counties is Based on Potential Outcomes

- Now assume that people sort into five largest counties have the highest treatment effects (δ_i)
- Thus the five largest counties are those people for whom concealed carry guns law cause homicides
- Remaining 249 small counties, in decreasing order, get residents with lower treatment effects (i.e., for whom concealed carry reduces murders)

Average County ATE and Overall ATE Differ



Both Are Valid, But Not Necessarily Your Research Question

- This has implications for what to learn from your dataset which is likely to be aggregated at some level (e.g., individual, city, county, state, country)
- Do you want to know the ATT for the *average person* or the *average county*?
  - → It depends on what your study is about
  - → If it about the average person, then you want the overall ATT (i.e., the first case)
  - → If it is about the average county, then you want the county average ATT (i.e., the second case)
- Since you're averaging over *units* in *data*, it's imperative you make a decision early on as it changes what you decide
- You can always use population weights but in causal inference, you ask what your target parameter is, and then decide your weights

Choosing your ATT parameter

- This means that even diff-in-diff has more than one ATT – the *average county* or *average person*
- As a rule, I think your goal is probably to imagine who your audience is, and what you both agree the policy levers are
- In the US, local municipalities have a lot of discretion to pass their own laws – even in areas where you might think it was impossible to imagine like the decriminalization of drugs
- It may be *local* policy makers want to know the causal effect on *local communities* in which case you **don't weight**

Choosing your ATT parameter

- But the argument isn't to weight to make it nationally representative – it's "is the average effect on the average county theoretically important, policy relevant and how one is planning to use it
- Others are also possible
- Quantile treatment effects (Athey and Imbens 2006; Callaway and Li 2019) and distribution regression target features of the marginal distributions of $Y_{i,t}^1$ and $Y_{i,t}^0$(Fernandez-Val, Meier, van Vuuren and Vella 2024a)

Roadmap

Why do Diff-in-Diff

- Appeal of diff-in-diff has been its simplicity, its transparency, and its ease of conveying analysis to an audience
  - → Orley Ashenfelter used it in the 1970s to explain regressions with fixed effects to Bureaucrats in DC
- Diff-in-diff is four averages and three subtractions and everyone knows what those are

$$\widehat{\delta} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$

- But $\widehat{\delta}$ is just the OLS coefficient in this regression:

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist}$$

Minimum wages

- Card and Krueger (1994) have a famous study estimating causal effect of minimum wages on employment
- New Jersey raises its minimum wage in April 1992 (between February and November) but neighboring Pennsylvania does not
- Using diff-in-diff, they do not find a negative effect of the minimum wage on employment leading to complex reactions from economists

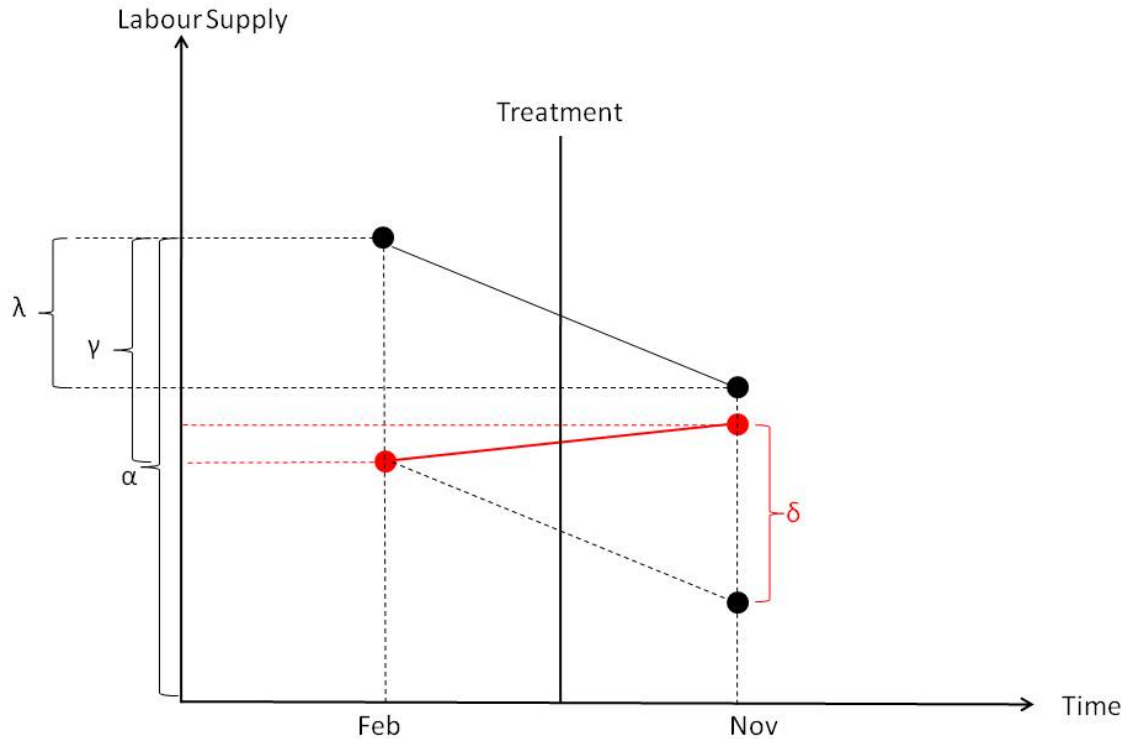
OLS specification of the DiD equation

- The correctly specified OLS regression is an interaction with time and group fixed effects:

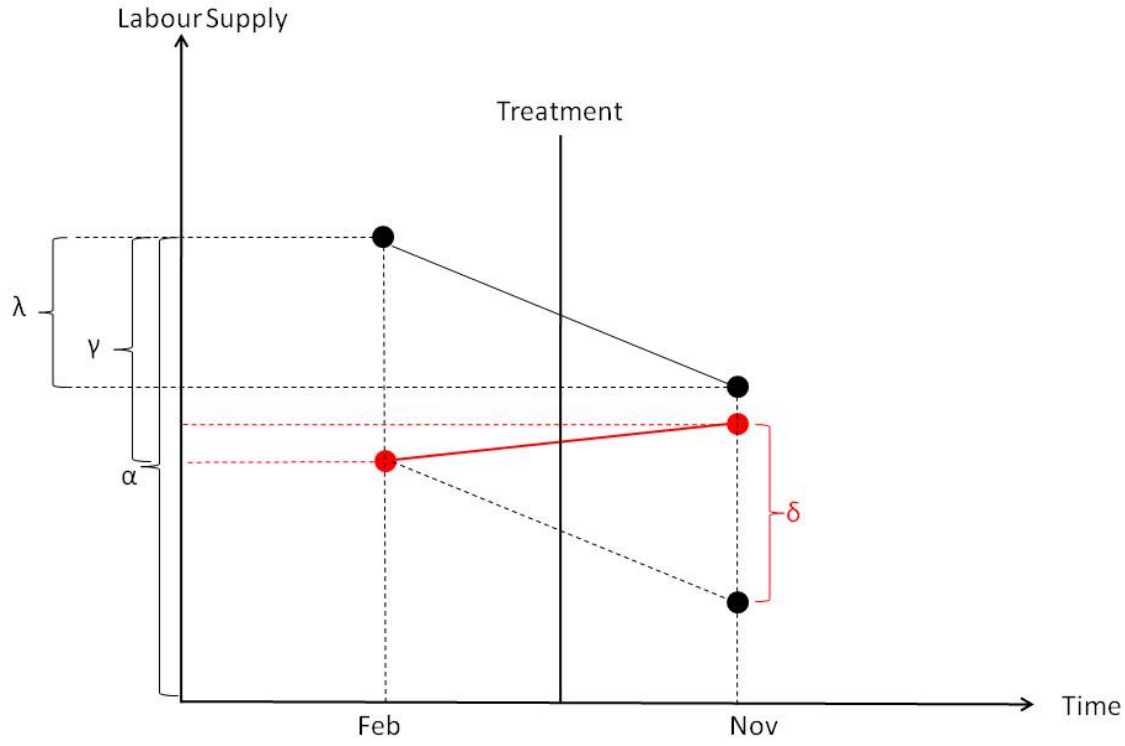
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

  - → 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: $\alpha + \gamma + \lambda + \delta$
- DiD equation: (NJ Post - NJ Pre) - (PA Post - PA Pre) $= \delta$

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Notice how OLS is "imputing" $E[Y^0|D=1, Post]$ for the treatment group in the post period? It is only "correct", though, if parallel trends is a good approximation

Data example

- Medicaid is a US policy for the poor – provides access to healthcare to the poor – and it's been a research question whether it has any effect on mortality
  - → Finkelstein, et al. (2012) found no effect on mortality from RCT, but it was probably underpowered to find small effects
  - → Borgschulte and Vogler (2020) used county level publicly available mortality data and found evidence that it did
  - → Miller, Johnson and Wherry (2022) used linked administrative data and also found it did
- Let's review the Borgschulte and Vogler (2020) approach

Simple 2x2 DD

Table 2: Simple 2 × 2 DiD

	Unweighted Averages			Weighted Averages		
	Expansion	No Expansion	*Gap/DiD*	Expansion	No Expansion	*Gap/DiD*
2013	419.2	474.0	*-54.8*	322.7	376.4	*-53.7*
2014	428.5	483.1	*-54.7*	326.5	382.7	*-56.2*
Trend/DiD	*9.3*	*9.1*	0.1	*3.7*	*6.3*	-2.6

Notes: This table reports average county-level mortality rates (deaths among adults aged 20-64 per 100,000 adults) in 2013 (row 1) and 2014 (row 2) in states that expanded adult Medicaid eligibility in 2014 (columns 1 and 4) and states that have not expanded by 2019 (columns 2 and 5). The first three columns present unweighted averages and the second three columns present population-weighted averages. Columns 1, 2, 4, and 5 in the third row show time trends in mortality between 2013 and 2014 for each group of states. The first two rows of columns 3 and 6 show the cross-sectional gap in mortality between expansion and non-expansion states in 2013 and 2014. The entries in bold red text in row 3 show the simple 2 × 2 difference-in-differences estimates without weights (column 3) and with them (column 6)

Three Regressions

- Three regression specifications give you those exact same numbers
  1. Regress mortality onto treatment dummy, post dummy and interaction (no fixed effects)
  2. Regress mortality onto interaction with county and year fixed effects (but no constant)
  3. Regress long difference (i.e., post value minus pre) onto a treatment dummy
- Those are all numerically identical to ”four averages and three subtractions”

Table 3: Regression 2 × 2 DiD

	Unweighted			Weighted		
	Crude Mortality Rate		Δ	Crude Mortality Rate		Δ
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	474.0***		9.1***	376.4***		6.3***
	(4.3)		(2.6)	(7.6)		(1.1)
Medicaid Expansion	-54.8***			-53.7***		
	(6.3)			(11.5)		
Post	9.1***			6.3***		
	(2.6)			(1.1)		
Medicaid Expansion × Post	0.1	0.1	0.1	-2.6*	-2.6*	-2.6*
	(3.7)	(3.7)	(3.7)	(1.5)	(1.5)	(1.5)
County fixed effects	No	Yes	No	No	Yes	No
Year fixed effects	No	Yes	No	No	Yes	No

Notes: This table reports the regression 2 × 2 DiD estimate comparing counties that expand Medicaid in 2014 to counties that do not expand Medicaid by 2019, using only data for the years 2013 and 2014. Columns 1-3 report unweighted regression results, while columns 4-6 weight by county population aged 20-64 in 2013. Columns 1 and 4 report results from regressing the crude mortality rate for adults ages 20-64 on indicators for expansion states (Treat) and post-expansion year (Post), with the DiD estimate being the coefficient on the interaction term. Columns 2 and 5 report the corresponding results for the interaction term using county and year fixed effects. Finally, Columns 3 and 6 report the results of the long difference in county mortality rates on a treatment indicator. Standard errors (in parentheses) are clustered at the county level.

Equivalence

- Equivalence between calculating these 2x2s by hand or with a regression has appealing features
- Regressions are simple to run, and they do the averaging behind the scenes
- They also allow us to use statistical inference tools from OLS like clustering decisions
- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely understate the standard deviation of the estimators and propose clustering standard errors at the aggregate unit level of treatment