Causal Inference II

Mixtape Session

Roadmap

- Introduction
  - Managing expectations
  - Origins of diff-in-diff in public health
  - Potential outcomes
  - Assumptions and Estimation
  - Design Stage
- Parallel Trends Violations
  - Results versus Evidence
  - Event Studies
- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
  - Double Robust
  - Canonical TWFE with Additive Covariates

Introduction

- Scott Cunningham, Ben H. Williams Professor of Economics at Baylor University in Texas USA
- Today will talk about causal panel methods
- Causal panel uses longitudinal data to estimate causal effects
- We'll cover difference-in-differences and synthetic control methods

What is difference-in-differences (DiD)

- 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
- We'll do a quick run through the social history of diff-in-diff to set the stage for our workshop this week

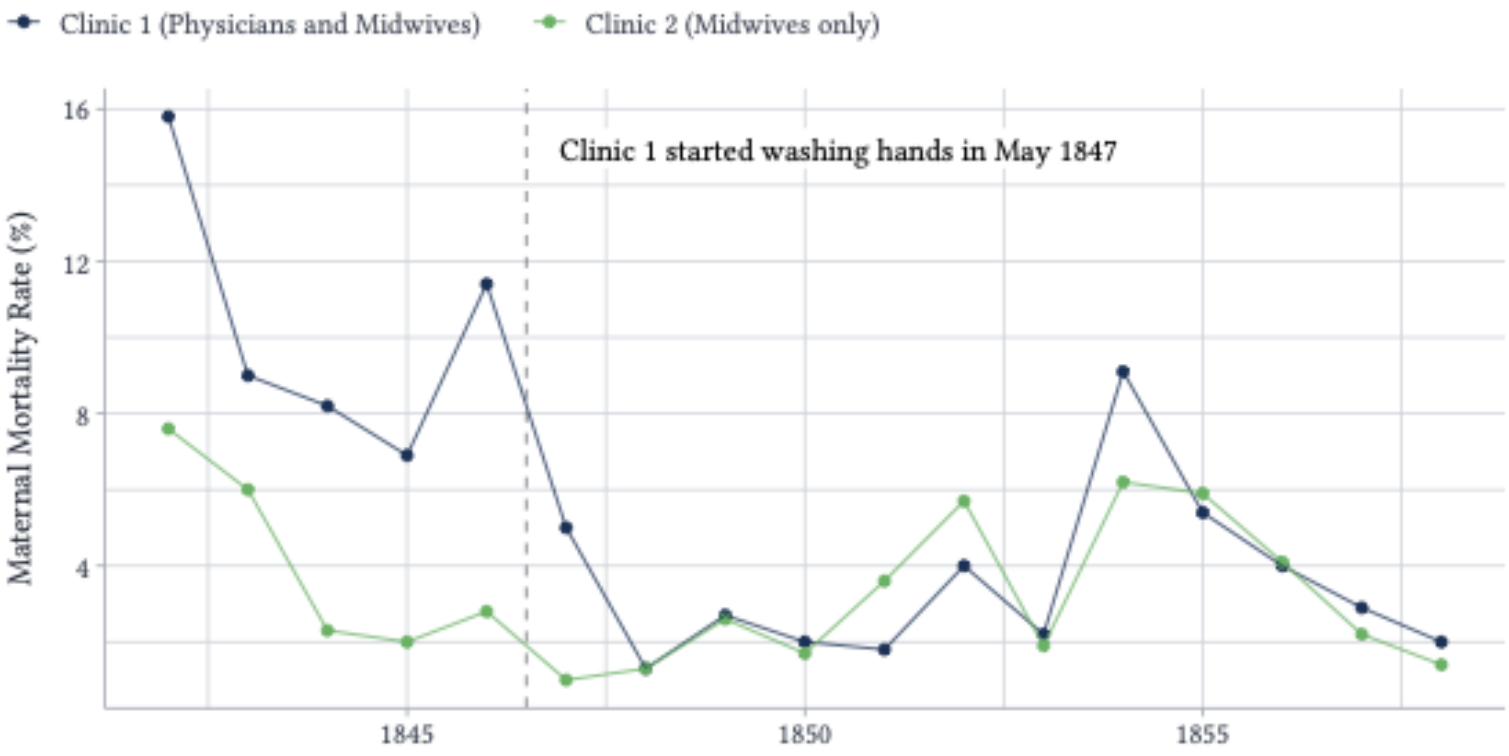
Ignaz Semmelweis and washing hands

- 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



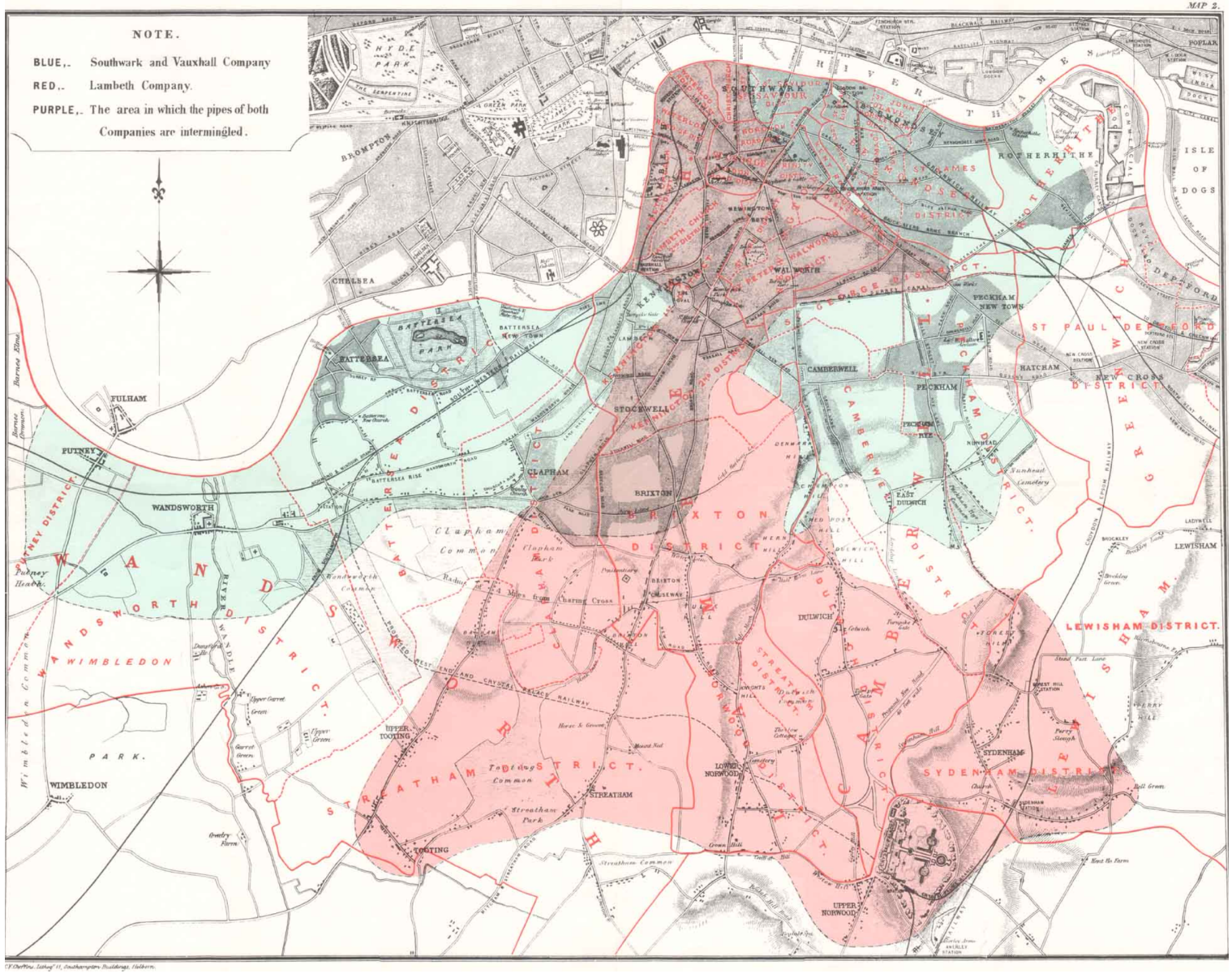
Evidence Rejected

- 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

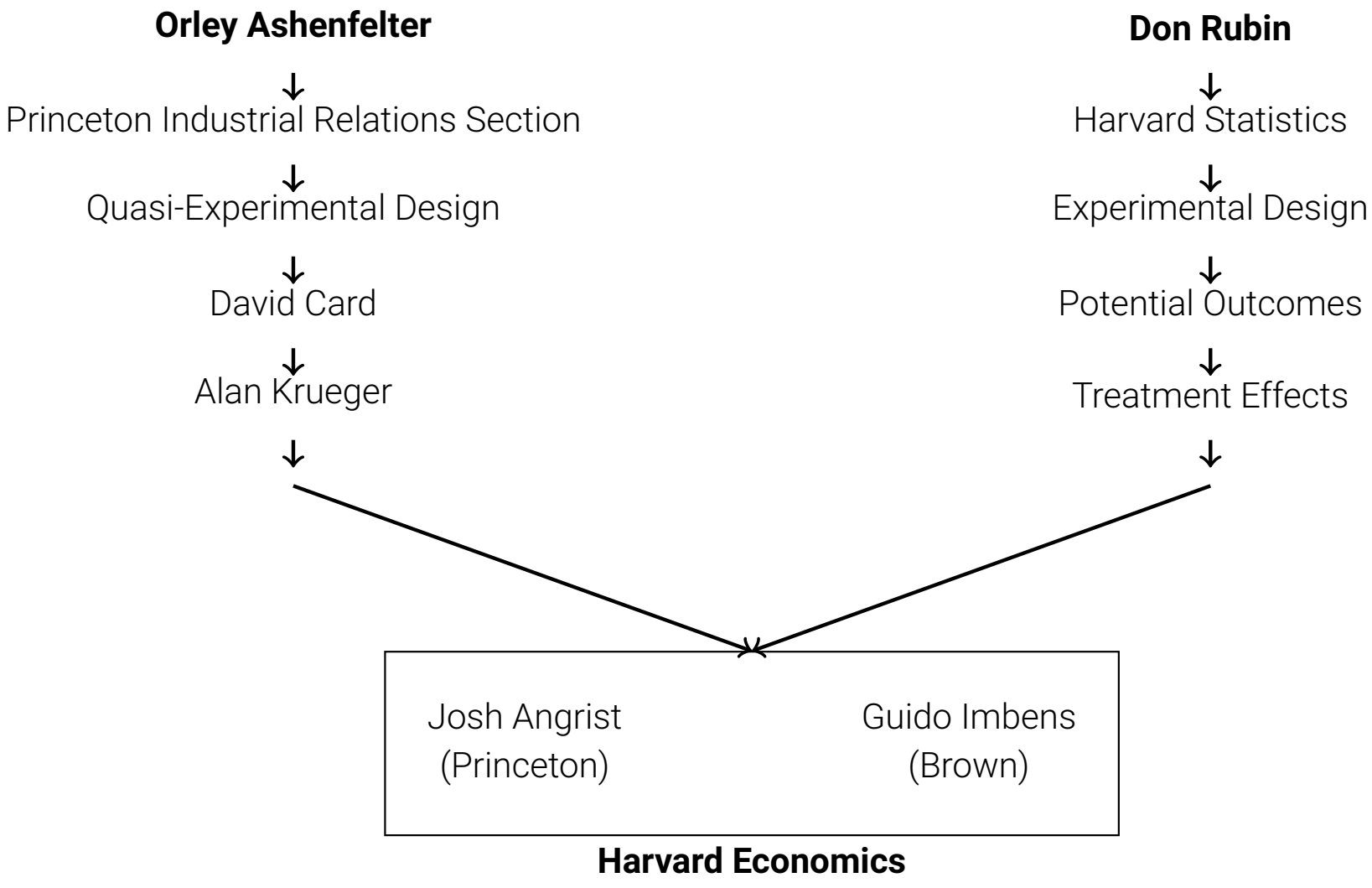
Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

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

Two rivers into causal inference



Background I: Harvard Stats and Potential Outcomes

- Don Rubin, former chair of Harvard stats, is the main source of potential outcomes, building on Jerzy Neyman's 1923 work.
- Rubin's influential 1970s papers advocated for causal inference using contrasts of $Y(1)$ and $Y(0)$.
- Neyman's notation, initially in Polish, was translated into English in 1990, likely due to Rubin.
- Rubin expanded Neyman's ideas from experiments to observational studies, leading to developments like propensity score methods.
- Economics was slow to adopt these methods initially.

Background II: Princeton Industrial Relations Section

- Late 1970s and early 1980s: little ”credibility” in empirical labor studies.
- Princeton Industrial Relations Section: older than the economics dept, rigorous, non-partisan focus on US ”manpower”, highly empirical.
- Key faculty are Orley Ashenfelter, David Card, Alan Krueger.
- Key students include Bob Lalonde, Josh Angrist, Steve Pischke, John Dinardo, Janet Currie, Anne Case, and many more.
- Listen to David Card: https://youtu.be/1soLdywFb_Q?si=BCVqYeRz6jYiwHTQ&t=1580

Background II: Princeton Industrial Relations Section

Example of Princeton Paradigm Emerging

- Lalonde (1986) was a groundbreaking study, recently reviewed by Guido Imbens and Yiqing Xu (2024)
- Lalonde, a student of Card and Orley, analyzed an RCT on a job training program, finding an average treatment effect of +$800.
- He then replaced the experimental control group with survey data, reran econometric methods, and couldn't replicate the results.
- Orley and Card emphasized randomization in their 1985 Restat article, advocating for its exploitation in studies.

Orley Ashenfelter and diff-in-diff

- Diff-in-diff gets rediscovered by Orley Ashenfelter from Princeton
- Leaves academia to work in Washington DC to study job training programs for low skill workers
- Coins the phrase ”difference-in-differences” so as to avoid having to explain regressions to bureaucrats (3:53)
  https://youtu.be/WnB3EJ8K7lg?si=uE4clqUIPzvbxm0r&t=2
- More associated with David Card (Mariel boatlift, minimum wage), but it was earlier that he and Orley worked with the method, and ironically largely, rejected its usefulness for the questions they were working on

- 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 DiD, they do not find a negative effect of the minimum wage on employment leading to complex reactions from economists
- Orley's describes his understanding of people's reaction to the paper. https://youtu.be/MOtbuRX4eyQ?t=1882

Binyamin Appelbaum
@BCAppelbaum

Replying to @BCAppelbaum

The Nobel laureate James Buchanan wrote in the Wall Street Journal that Card and Krueger were undermining the credibility of economics as a discipline. He called them and their allies "a bevy of camp-following whores."

3:49 PM · Mar 18, 2019

179 Reply Share

Read 18 replies

Reaction to the paper

Lots of anecdotes in this interview with Card, but here are just two. First, Card and Krueger received a lot of personal hostility from their peers (1:07 to 1:10)

https://youtu.be/1soLdywFb_Q?si=laAVYf_E2KBZKywG&t=4020

Later Card says Sherwin Rosen accused them of having an agenda. But the worst is what happens to Alan Krueger maybe (1:16 to 1:17)

https://youtu.be/1soLdywFb_Q?si=jsb8h50ZosGDnKrv&t=4556

Card on that study

"I've subsequently stayed away from the minimum wage literature for a number of reasons. First, it cost me a lot of friends. People that I had known for many years, for instance, some of the ones I met at my first job at the University of Chicago, became very angry or disappointed. They thought that in publishing our work we were being traitors to the cause of economics as a whole."

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 numerical calculation like Orley was doing so let's do that

Four Averages and Three Subtractions

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

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

- Orley claims that the OLS estimator of δ and the "four averages and three subtractions" calculation are numerically identical
  https://youtu.be/WnB3EJ8K7lg?t=126
- They are and I want you to see that with a numerical example in `equivalence.do` and `equivalence.R`

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

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

- Imagine there are 100 cities – 25 of them raise the minimum wage, 75 do not
  - → ATE is the average treatment effect across all 100 cities
  - → ATT is the average treatment effect for the 25 cities that raised the minimum wage
- If you want to know the ATE, diff-in-diff is not appropriate as it only identifies the ATT
- Just keep in mind the *population* – you can average the treatment effects for everyone or just some people, but whichever affects the interpretation

Potential Outcomes vs Realized Outcomes

- When you're analyzing data, you are working with the realized outcomes
- When you are thinking about your treatment parameters, you are working with potential outcomes
- When treatment does occur, then only one of the potential outcomes happens ("realized outcome") but the other one becomes "counterfactual" or missing
- We represent this with the switching equation

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$$

- How those treatment get assigned is called the treatment assignment mechanism, but let's review this together first before we discuss that

DiD equation is the 2x2

Orley's "four averages and three subtractions" uses two groups, two time periods, or 2x2

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

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.

When will $\widehat{\delta}$ equal the ATT? When will it not?

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:

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

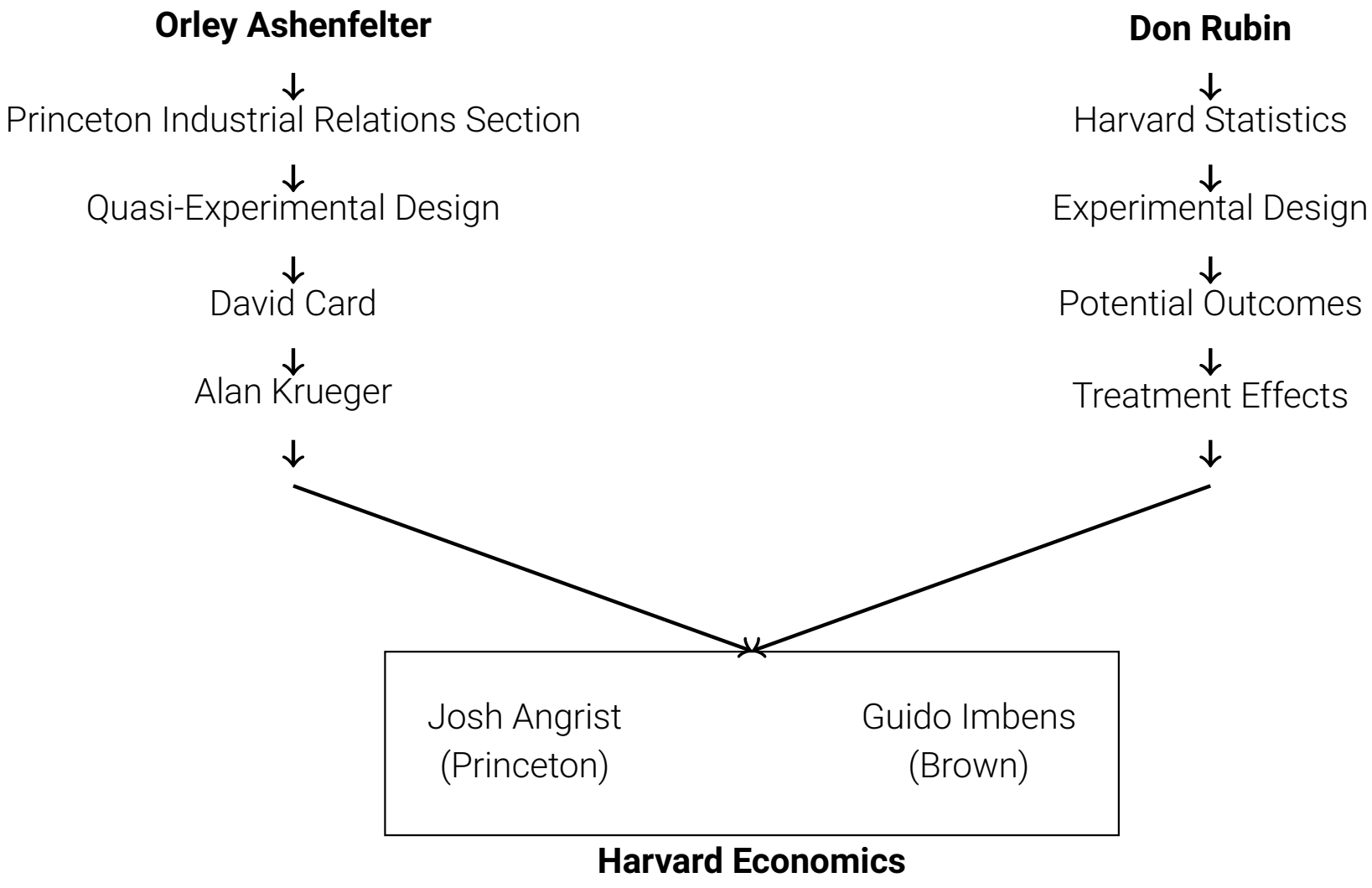
In words: "The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees".

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"

Design stage in diff-in-diff

- There is a period of time before you *ever* touch your data on outcomes where you design your study
- You want to be thinking in this stage of things like this:
  - → What does a treatment effect mean in my study?
  - → Diff-in-diff identifies the ATT – what does that mean in my study?
  - → If I ran a randomized experiment on my population, what would I be randomizing?
  - → Do I think unconditional parallel trends (i.e., no covariates) holds? Why?
- Because many people come to causal inference from outside the experimental design tradition, this step is confusing, but keep this picture in mind

Two rivers into causal inference



Causal Inference in Design Tradition

- So you have to learn the design tradition and adopt its approach
- Does not mean you abandon your prior human capital
- Practical implication is that you spend more time in the design stage thinking about treatment assignment mechanism, covariate imbalance
- But it also means you don't "peek at the outcomes" to avoid biasing *yourself* until you're ready
- Imagine you were pre-registering an RCT – try to mimic that even in your diff-in-diff

What We Should Be Concerned About

- **Time-varying unobservables**
  - → Unobservable factors that change over time can create biases if they affect treated and control groups differently.
  - → This can lead to violations of the parallel trends assumption if such unobservables are correlated with treatment.
- **Selection based on post-treatment information**
  - → If individuals select into treatment based on information they gain after the treatment starts (foresight), this undermines the assumption.
  - → Essential heterogeneity means that the potential outcomes vary systematically with unobserved traits, which complicates inference.

What We Should Be Concerned About

- **Non-stationarity**
  - → Non-stationarity refers to changing relationships between covariates and outcomes over time, which can lead to inconsistent treatment effect estimates.
  - → Trends in the data that are not stable or stationary across time are particularly problematic.
- **Martingale property violation**
  - → If the unobservable shocks affecting outcomes do not follow a martingale process, meaning future changes are not independent of past outcomes, the model's assumptions may fail.
  - → This can result in time-dependent correlations in unobservable factors that distort trend comparisons.

What is and is not parallel trends?

- Parallel trends does *not* mean treatments were randomly assigned (though random assignment guarantees parallel trends)
- Parallel trends does *not* require that the groups be similar at baseline on outcomes (though random assignment guarantees that would be)
- Parallel trends *does* require that the comparison group follows a trend in outcomes that is approximately the same as the counterfactual trend of the treatment group (what would have had happened had the treatment not occurred)

Three main DiD assumptions

- Parallel trends is the most common one and most well known
- But parallel trends is nested within a bundle of assumptions, and all of them are need for traditional difference-in-differences
- Other two lesser known assumptions are ”No anticipation” (or NA) and Stable Unit Treatment Value Assumption (SUTVA)

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`.

SUTVA

- Stable Unit Treatment Value Assumption (Imbens and Rubin 2015) focuses on what happens when in our analysis we are combining units (versus defining treatment effects)
    1. **No Interference**: a treated unit cannot impact a control unit such that their potential outcomes change (unstable treatment value)
    2. **No hidden variation in treatment**: When units are indexed to receive a treatment, their dose is the same as someone else with that same index
    3. **Scale**: If scaling causes interference or changes inputs in production process, then #1 or #2 are violated
- Shifts from defining treatment effects to estimating them, which means being careful about who is the control group, how you define treatments and what questions can and cannot be answered with this method

Do not use already treated controls

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

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

Similar to SUTVA but this time it isn't caused by a spillover – it's just an already treated group. Let's do this together now. It'll help to see it with your own eyes.

Already Treated Control Group

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

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

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

Summarizing

- Lots of restrictions placed on difference-in-differences
  - → NA: you chose a baseline that is not treated
  - → SUTVA: your comparison group is never treated during the course of the calculations
  - → PT: your comparison group has a trend in $E[Y^0]$ that is the same as the counterfactual
- Only when you have NA SUTVA and you use an untreated group as a control does DiD equal ATT + PT
- But it's crucial to remember: DiD and ATT are not the same thing

OLS Specification

- OLS specification is the same as four averages and three subtractions
- Assume unconditional parallel trends (i.e., no covariates are necessary), then OLS might be preferred because . . .
  - → OLS estimates the ATT under parallel trends
  - → Easy to calculate the standard errors
  - → Easy to include multiple periods
- People liked it also because of differential timing, continuous treatments and covariates, but those are more complex so we address them later

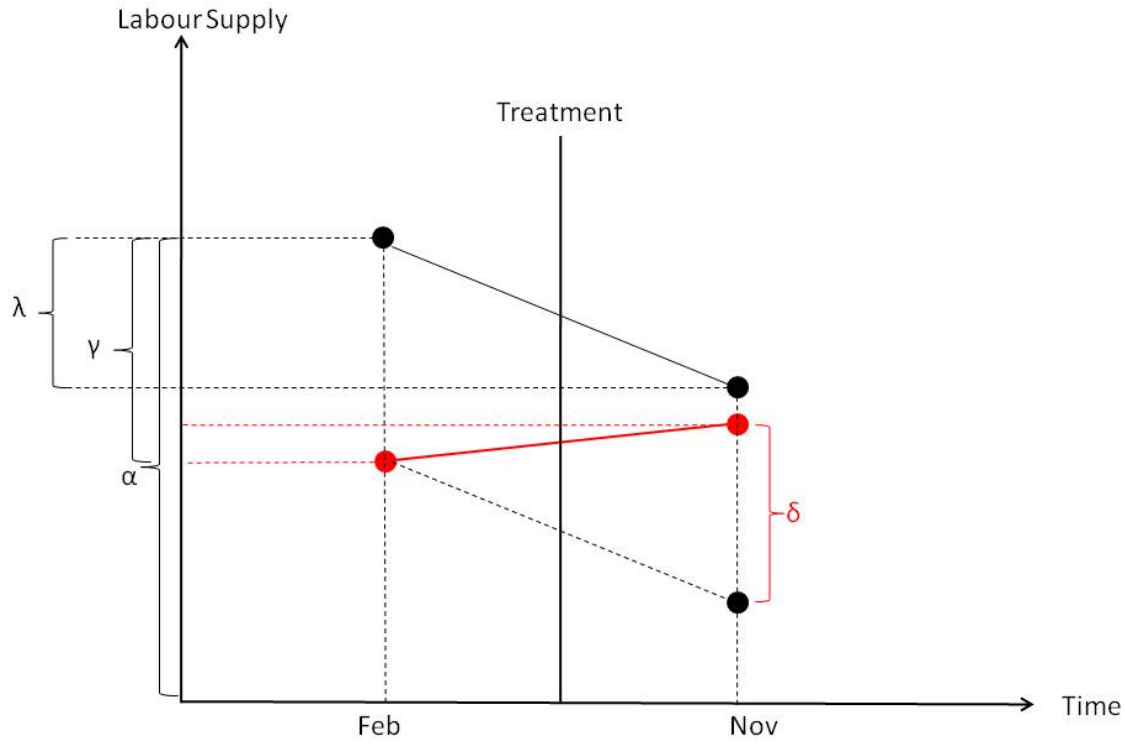
OLS specification of the DiD equation

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

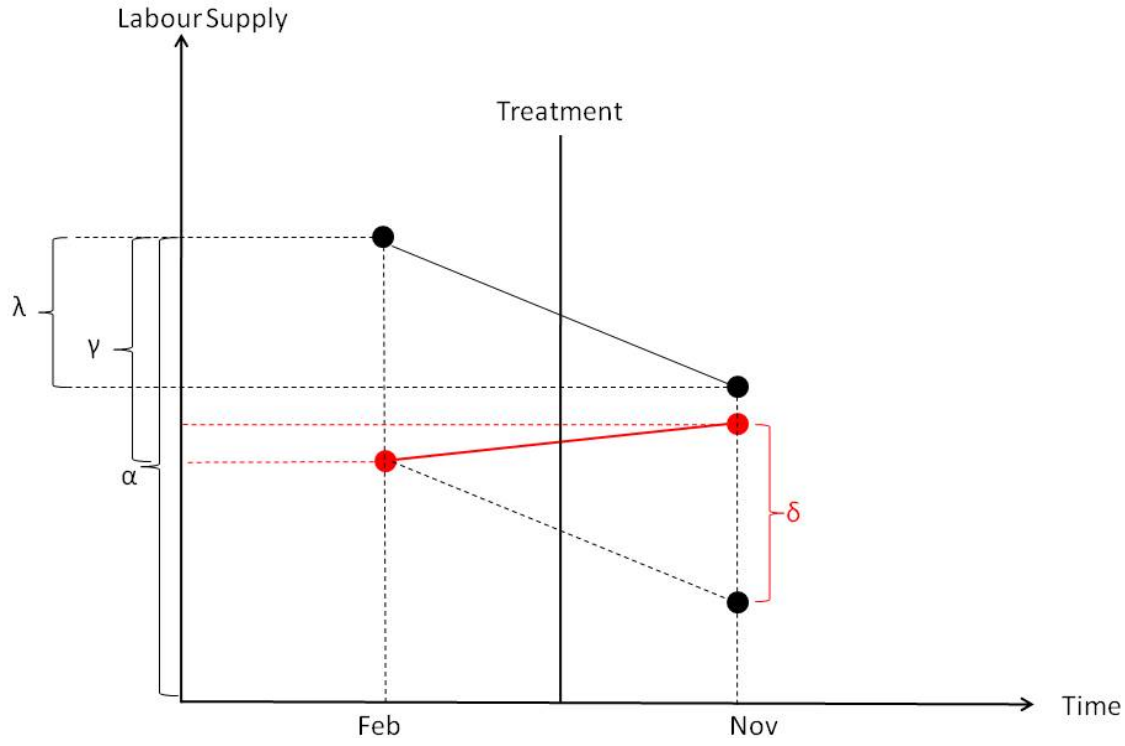
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

  - → 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

Inference with correlated errors

- Correlated errors occur when the unobserved errors are correlated within a cluster.
- This violates the assumption of independent errors, leading to possibly biased standard errors and higher over rejection rates
- Failing to account for correlated errors can lead to misleading inference.

Conservative inference in DiD

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely understate the standard deviation of the estimators
- They proposed three solutions, but most only use one of them (clustering)
- Clustering standard errors accounts for this within-cluster correlation and is a more conservative approach
- Clustering is typically recommended at the aggregate unit where the entire treatment occurred

Roadmap

- Introduction
  - Managing expectations
  - Origins of diff-in-diff in public health
  - Potential outcomes
  - Assumptions and Estimation
  - Design Stage
- Parallel Trends Violations
  - Results versus Evidence
  - Event Studies
- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
  - Double Robust
  - Canonical TWFE with Additive Covariates

Evidence versus the Main Result

- Causal inference is about *warranted beliefs* – should you or should you not believe the *causal claim*?
- Your DiD *results* are like the claim of guilt, but your DiD results are *not* the smoking gun
  - → Your table of regression coefficients *is not enough* for evidence
  - → You need to do more to provide a justification for parallel trends
- You need to provide evidence for parallel trends against several well known vulnerabilities
- Evidence will be bite, falsifications, mechanisms and event study data visualization
- We will mix the parallel trends violations with the evidence concept before getting into advanced estimators

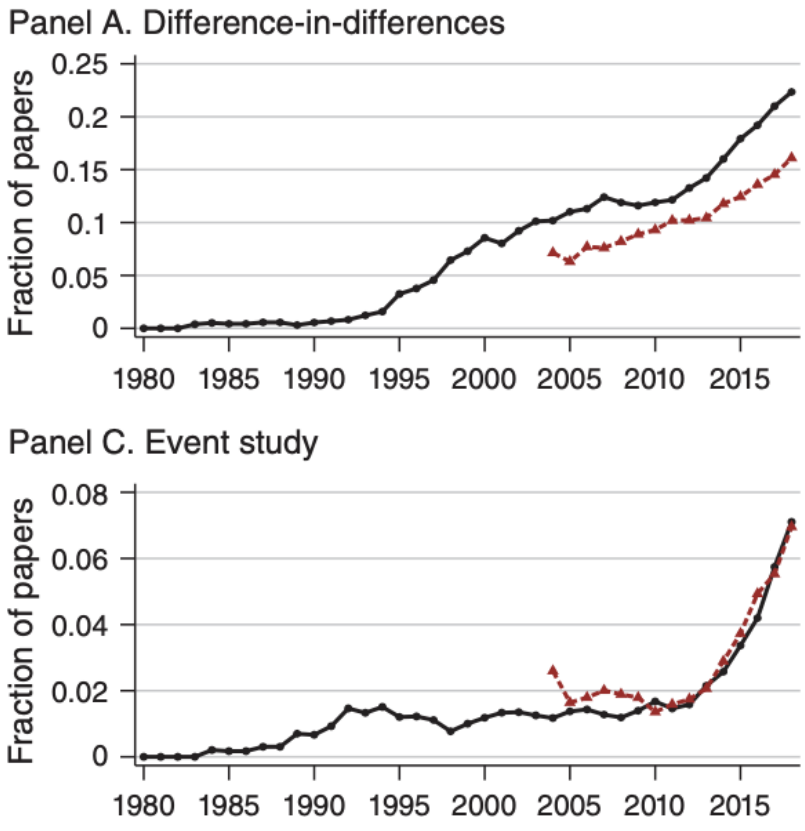
Court metaphor

- Think of yourself as a prosecutor arguing against a defense attorney to convince a judge and jury of a defendant's guilt
- The claim the defendant is guilty is your table of main results
- But the claim is not the evidence – you have to back up that claim
- Your evidence of guilt is the smoking gun, the fingerprints, the eye witnesses, the footprints in the mud outside the house
- If your claim is supported by weak evidence, then no one *should* convict – it would be borderline corruption if they did

Three classic parallel trends violations

1. Compositional change with repeated cross-sections
2. Policy endogeneity
3. Omitted covariates needed for parallel trends to hold

Event studies have become mandatory in DiD

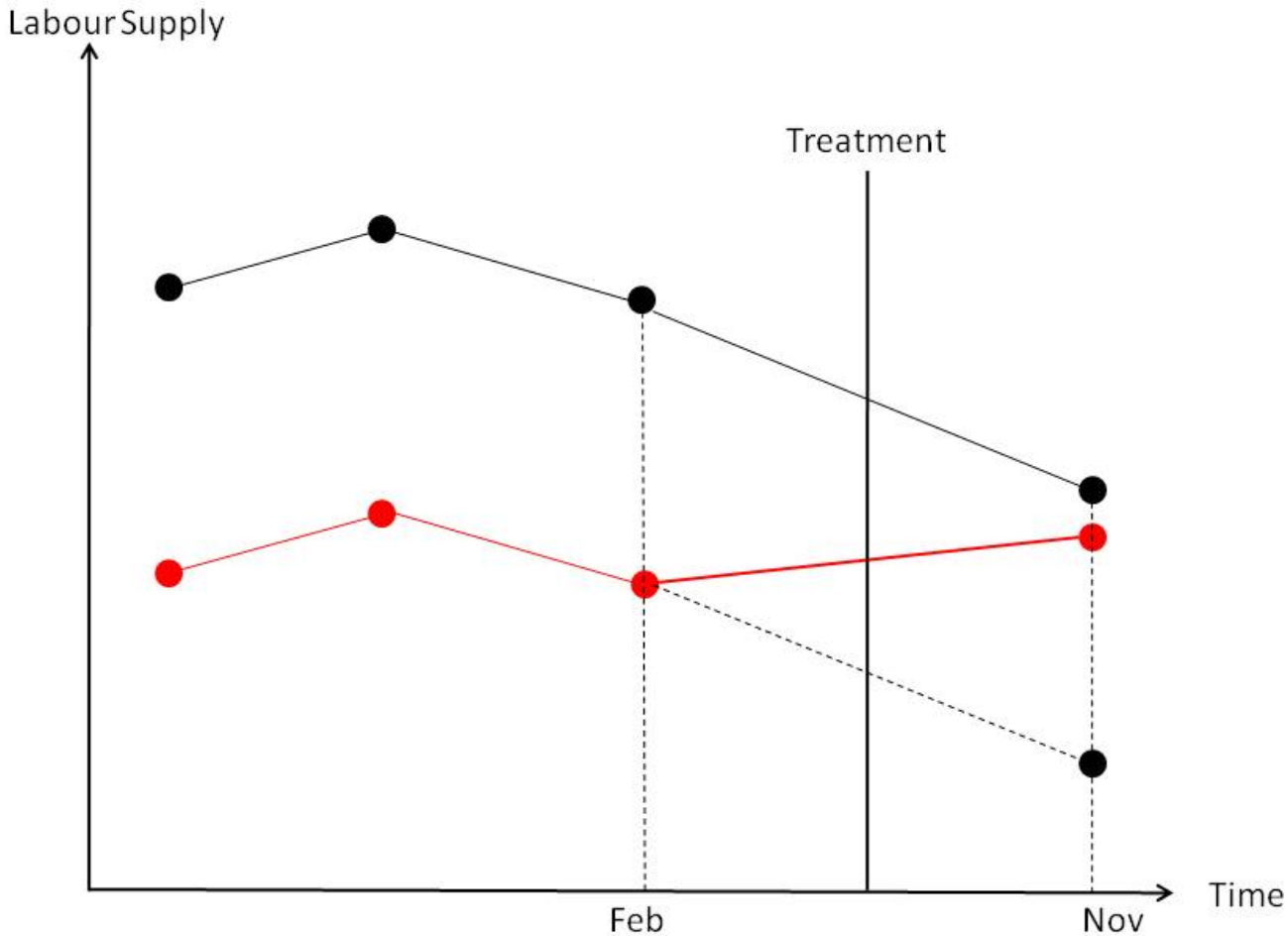


Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends (e.g., Ashenfelter's Dip)
- Parallel pre-trends are not the same as parallel counterfactual post-trends, but this is the smoking gun we typically look for nonetheless
- Think of it as a type of check for selection bias, but imperfect with false positives and false negatives
- Even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias is a parallel trends violation)

Creating event studies

- You want to visualize a particularly set of regression coefficients which we will show
- But if you can show the raw data, do that too as that will show difference in levels as well which will matter for the next section on covariates


Labour Supply
Treatment
Time
Feb
Nov

Event study regression

- Alternatively, present estimated coefficients from a dynamic regression specification:

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau (D_s \times \tau_t) + \sum_{\tau=0}^{m} \delta_\tau (D_s \times \tau_t) + \tau_t + D_s + \varepsilon_{ist}$$

  - → With a simple 2x2, you are interacting treatment indicator with calendar year dummies
  - → Includes q leads (dropping the $t-1$ as baseline) and m lags
  - → Since treatment did not happen until $\tau = 0$, then pre-treatment coefficients only capture differential trends
- Estimated $\widehat{\delta}_\tau$ coefficients are estimated ATT for each year under parallel trends but $\widehat{\mu}_\tau$ is your smoking gun evidence
- Just remember that $\mu = 0$ is not the same as parallel trends as parallel trends is untestable.

Reviewing previous slide for emphasis

- Under NA, SUTVA and parallel pre-trends, then mechanically $\widehat{\mu_\tau}$ will be zero as everything cancels out
  - $\rightarrow$ There are still specification and power issues that Jon Roth has written about, but I will skip that
- But also under NA, SUTVA and parallel trends (post trends), then $\widehat{\delta}$ are estimates of the ATT at points in time
- Typically you'll plot the coefficients and 95% CI on all leads and lags

Normal DiD coefficient

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

But this was *post*-treatment. Still, put that aside – diff-in-diff equations *always* identify the sum of those terms, even in the pre-period

Pre-treatment DiD coefficient

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

Under NA, then the $t-1$ period is untreated. But then so are the other pre-periods so the ATT is implicitly zero and the *only* thing that you can be measuring with pre-trend DiD coefficients is differential trends.

Event study coefficients

- Remember that the OLS specification we discuss collapses to ATT plus parallel trends bias
- This is *always* true because it's an identity and holds even in the pre-period as much in the post
- It's just in the pre period, you do not have the missing $E[Y^0|D=1]$ term as no one and nothing is treated in pre-period under NA
- This means pre-period is basically an opportunity to directly verify parallel pre-trends – but it's the past's pre-trends, not the counterfactual pre-trend of the present/future
- And that's how people use the pre-period – they use the pre-period to evaluate whether they think this is a good control group

Event study example

- The notion is really simple: if PT held then, you'll argue that it's reasonable it would've still held
- But this is an assertion, and you need to build the case as we said
- At this point, it's a lot easier to show you what I'm talking about – where the art and the science meet – with a great paper

Medicaid and Affordable Care Act example

Volume 136, Issue 3
August 2021

< Previous Next >

Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data

Get access >

Sarah Miller, Norman Johnson, Laura R Wherry

The Quarterly Journal of Economics, Volume 136, Issue 3, August 2021, Pages 1783–1829,
https://doi.org/10.1093/qje/qjab004
Published: 30 January 2021

Cite Permissions Share ▾

Abstract

We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults most likely to benefit using survey information on socioeconomic status, citizenship status, and public program participation. We find that prior to the ACA expansions, mortality rates across expansion and nonexpansion states trended similarly, but beginning in the first year of the policy, there were significant reductions in mortality in states that opted to expand relative to nonexpanders. Individuals in expansion states experienced a 0.132 percentage point decline in annual mortality, a 9.4% reduction over the sample mean, as a result of the Medicaid expansions. The effect is driven by a reduction in disease-related deaths and grows over time. A variety of alternative specifications, methods of inference, placebo tests, and sample definitions confirm our main result.

JEL: H75 - State and Local Government: Health; Education; Welfare; Public Pensions, I13 - Health Insurance, Public and Private, I18 - Government Policy; Regulation; Public Health

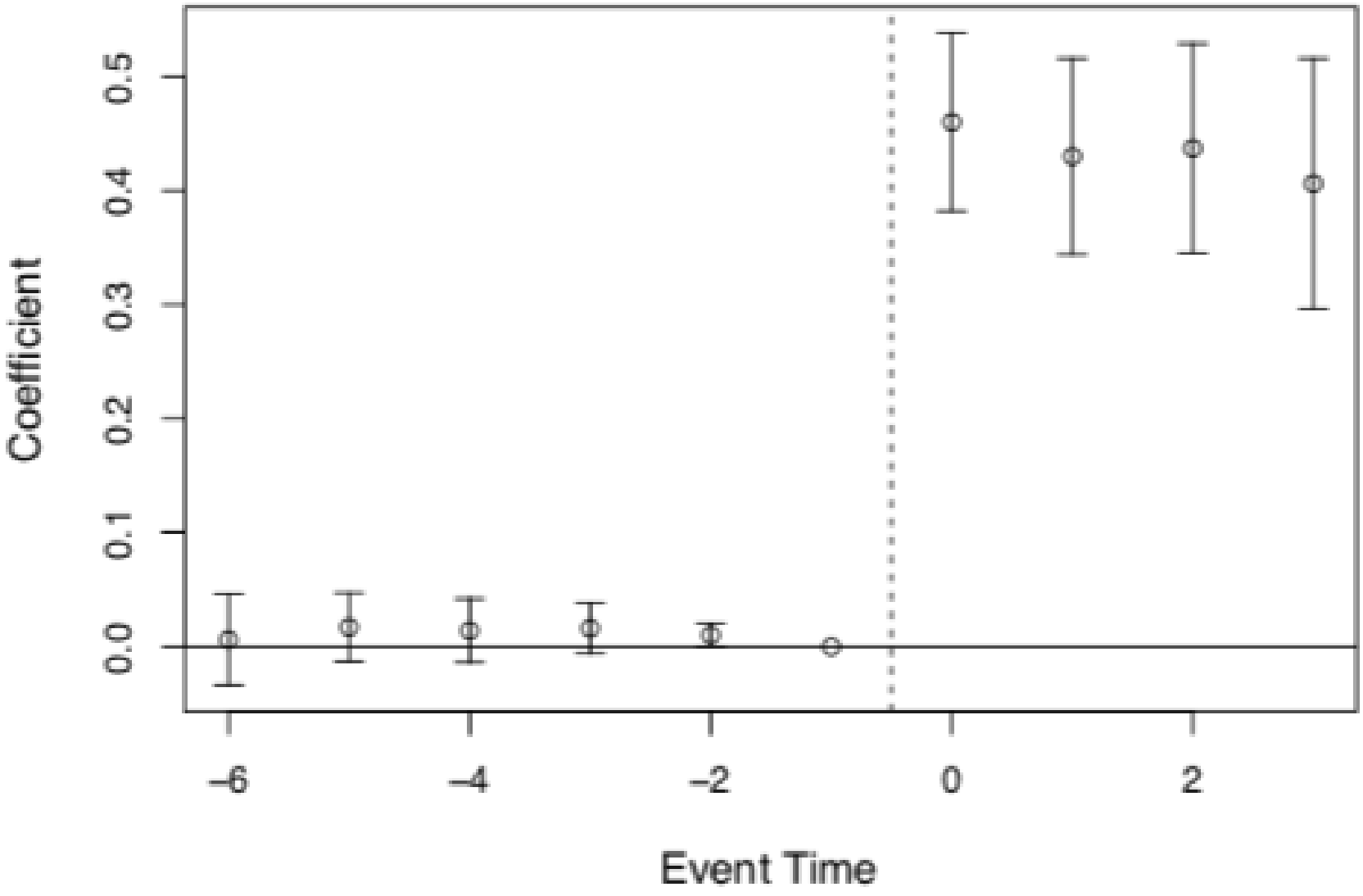
Issue Section: Article

Their Evidence versus Their Result

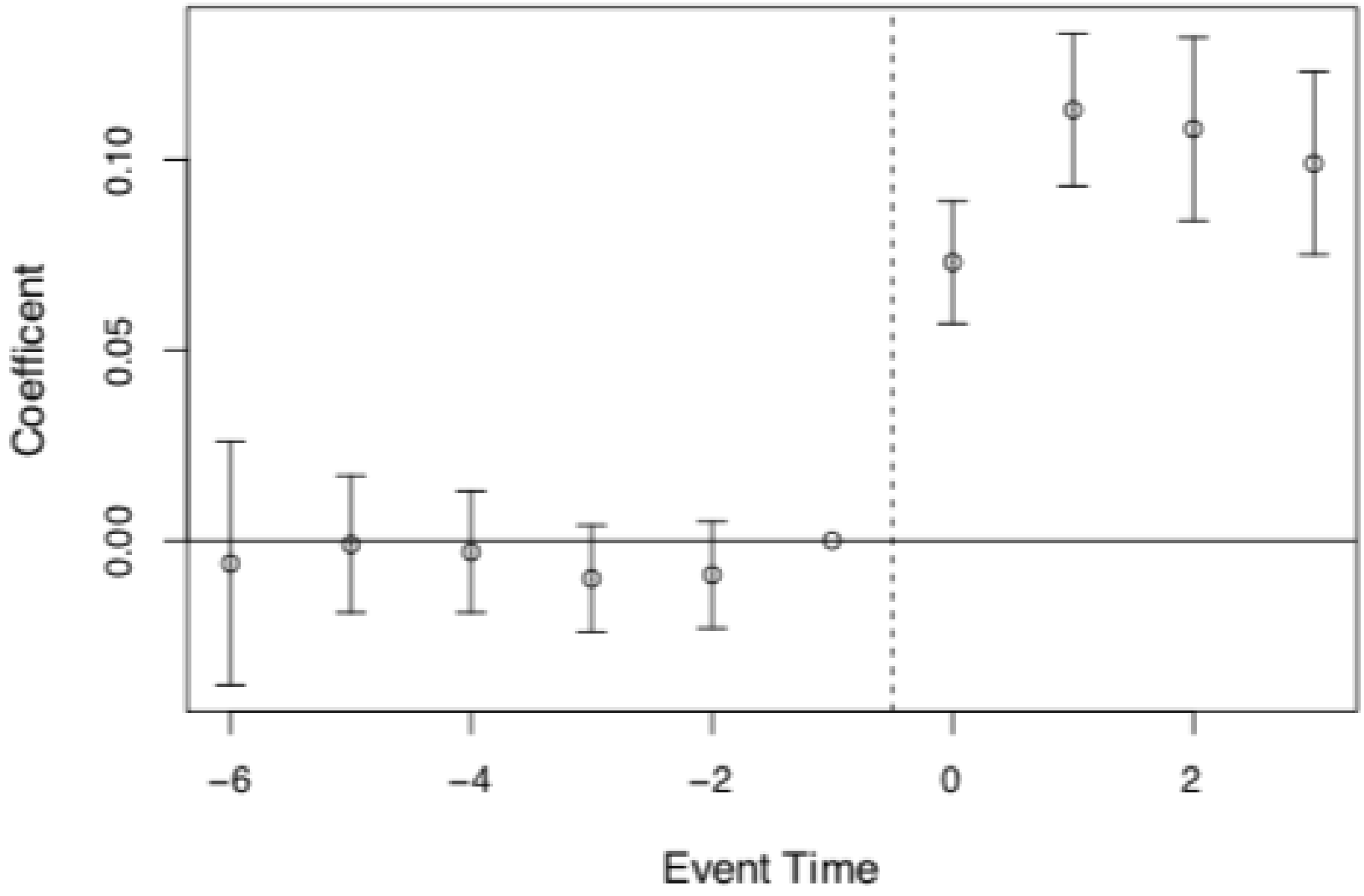
- **Bite** – they will show that the expansion shifted people into Medicaid and out of uninsured status
- **Placebos** – they show that there's no effect of Medicaid on a similar group that didn't enroll
- **Event study** – they will lean hard on those dynamic plots
- **Main results** – with all of this, they will show Medicaid expansion caused near elderly mortality to fall
- **Mechanisms** – they think they can show it's coming from people treating diseases causing mortality declines to compound over time

Bite

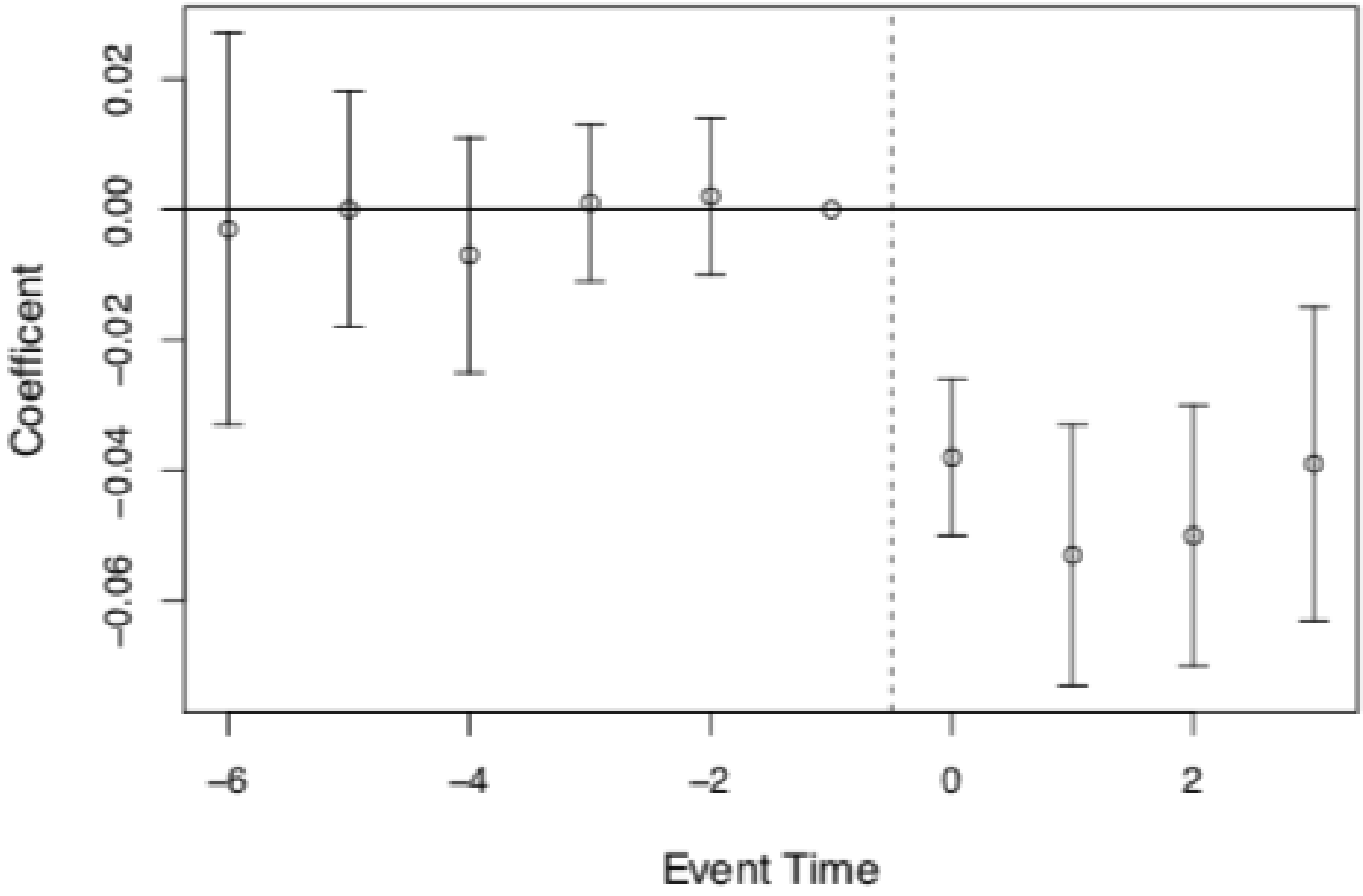
- Bite is a labor economist's phrase, often used with the minimum wage, to say that the minimum wage actually was binding in the first place
- Here it means when US states made Medicaid more generous, people got on Medicaid who would not have been on it otherwise
- And as a bonus, would not have been insured at all without it
- Not the most exciting result, but imagine if the main results on mortality were shown but there was no evidence for bite – is it believable?



(a) Medicaid Eligibility



(b) Medicaid Coverage



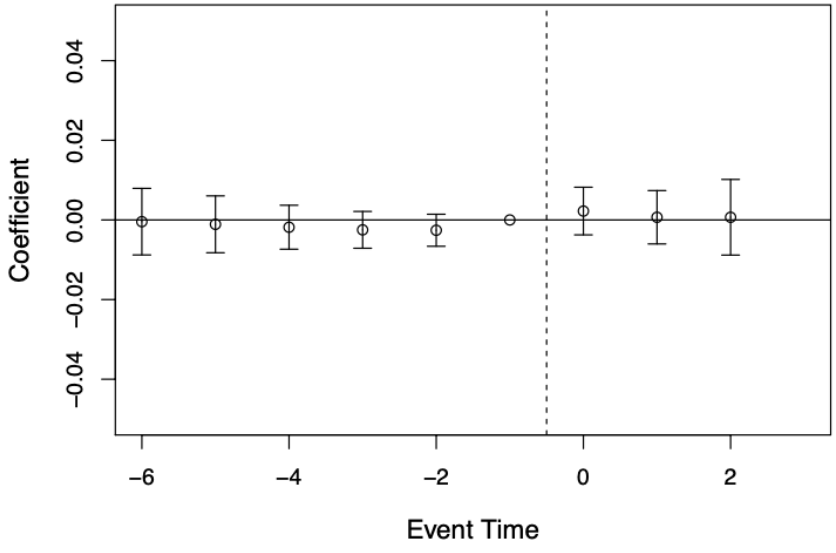
(c) Uninsured

Falsification

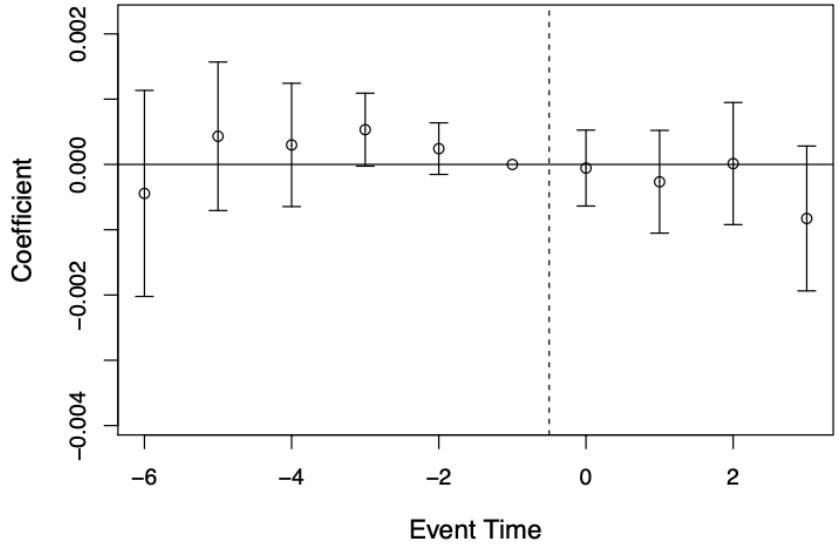
- Their study focuses on “near elderly”, which means just under 65
- They choose just under 65 because in the US, 65 and older are eligible for Medicare so more generous Medicaid is irrelevant
- *But* probably the near elderly and the elderly are equally susceptible to unobserved factors correlated with the treatment
- So they painstakingly examine the effects on elderly as a falsification as this will strengthen the parallel trends assumption on the near elderly

Falsifications on elderly

Age 65+ in 2014



(c) Medicaid Coverage

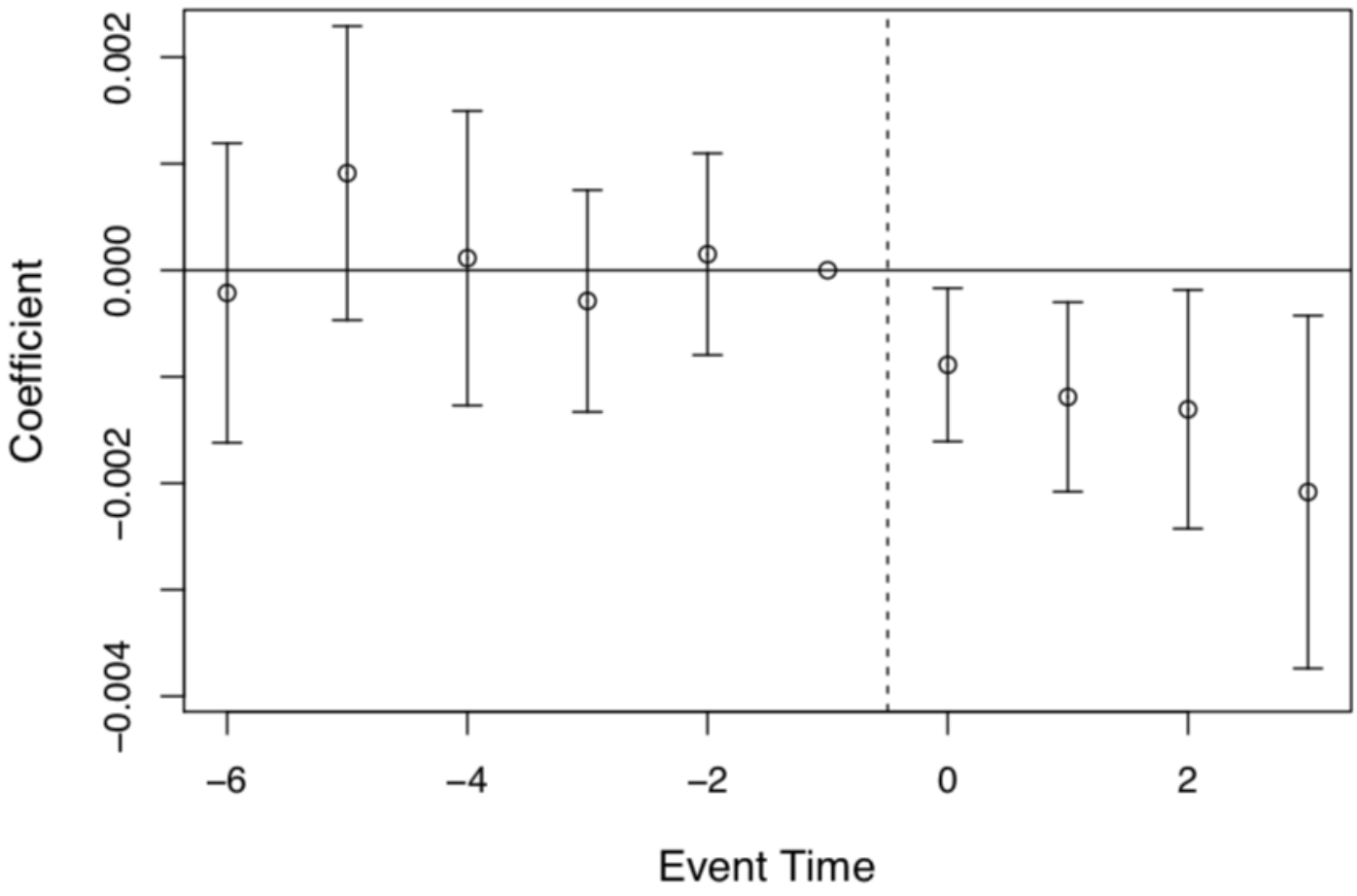


(d) Annual Mortality

Main result

- Finally they focus on the main result – and there's more in the paper than I'm showing
- Event study plots with same specification as the rest allowing us to look at the pre-trends and the post-treatment coefficients
- If parallel trends holds, then the post-treatment coefficients are interpreted as ATT parameter estimates for each time period
- The result alone isn't nearly as strong the result in combination with the rest, but it could still be wrong as parallel trends is ultimately not verfiable

Near elderly mortality and Medicaid expansion



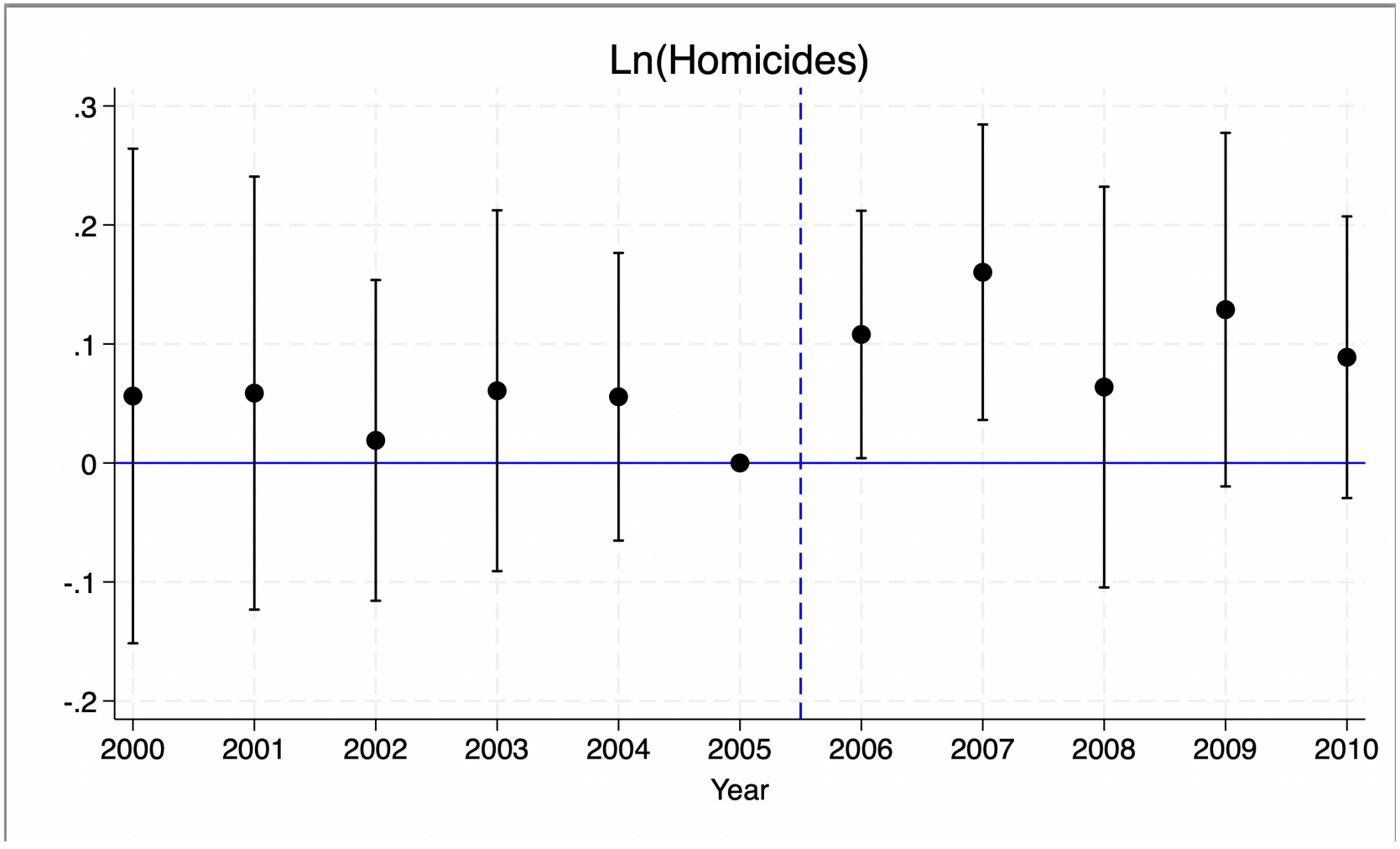
Summarizing evidence and results

- **Bite**: Increases in enrollment and reductions in uninsured support that there is adoption of the treatment
- **Event studies**: Compelling graphics showing similarities between treatment and control
- **Falsifications**: no effect on a similar group who isn't eligible
- Main results: 9.2% reduction in mortality among the near-elderly
- Mechanism: "The effect is driven by a reduction in disease-related deaths and grows over time."

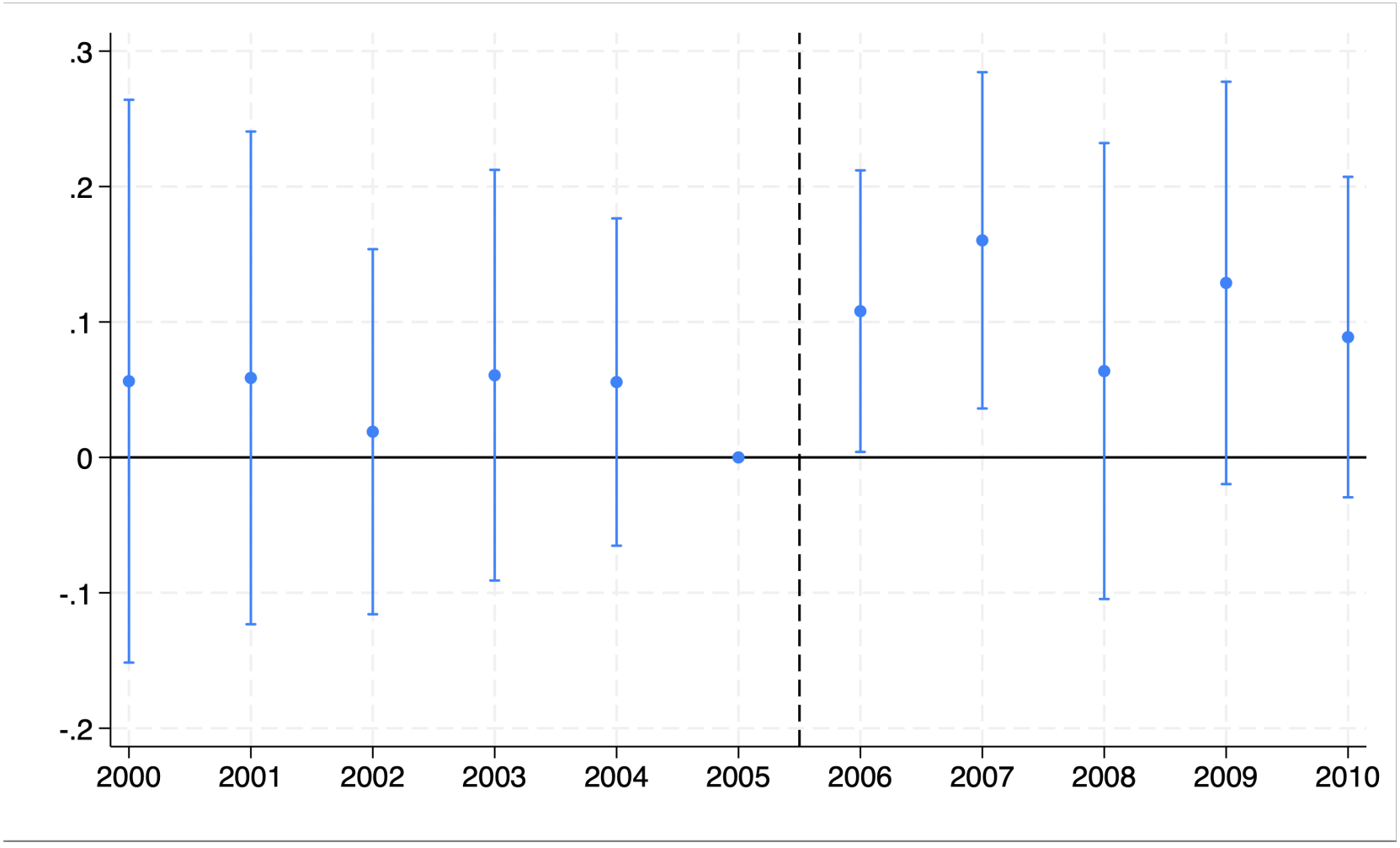
Making event study

- When there is only one treatment group and one comparison group, then you run a regression with an interaction of the treatment group dummy and the calendar year dummies (plus both separately)
- You must drop $t - \tau$ as the baseline (e.g., $t - 1$) and it must be Y^0 untreated comparisons (No Anticipation)
- I have included in a do file that will do it for you either manually or using coefplot in `simple_eventstudy.do` at the shared github labs directory

Manually creating the event study



Creating the event study with Ben Jann's `coefplot`



Parallel Trends

- Of the three assumptions we discussed – SUTVA and no treated control group, unconditional parallel trends, and no anticipation – it's usually parallel trends we worry about
- We discussed Event Studies
- But we were focused on *unconditional parallel trends* (i.e., no covariates), but is unconditional parallel trends plausible in your study?
- Next we will explore when you think you need covariates – why do you need them and how do you choose them?

Roadmap

- Introduction
  - Managing expectations
  - Origins of diff-in-diff in public health
  - Potential outcomes
  - Assumptions and Estimation
  - Design Stage
- Parallel Trends Violations
  - Results versus Evidence
  - Event Studies
- Conditional Parallel Trends
  - Introducing Covariates
  - Choosing Covariates
  - Checking for Imbalance
  - Double Robust
  - Canonical TWFE with Additive Covariates

Review

- Diff in Diff is a quasi-experimental method that identifies the aggregate causal parameter, ATT, if:
  1. Parallel trends holds
  2. And also no anticipation and if we use comparisons that aren't treated
- But we worked with "unconditional parallel trends" which means this:

$$\Big(E[Y_k^0|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)$$

- This means that the *average* treatment group potential outcome trends for our treatment group is the same as our control group

Is Unconditional Parallel Trends Plausible?

- But what if these two groups are different in ways that predict Y^0?
- Let's say for instance that your study looks at an intervention that primarily effects cities, but your only control group are rural counties
- These groups differ in a ways that might, depending on your study, not be plausibly suggesting they would've been on the same trends

Conditional parallel trends

- While it isn't required that two groups – group k and group U – be similar on observables for parallel trends to hold:
  1. Remember, if treatment had been random, then they would have the average covariates
  2. And if they aren't, then you may need to provide more evidence that these differences are irrelevant
  3. So you should still check and see and if they aren't, you either adjust or have a good justification for why they're the same on trends in $E[Y^0]$
- Including covariates weakens the unconditional parallel trends assumption by only requiring that units with similar covariate values, as opposed to everyone, follow similar counterfactual trends in Y^0
- This is means switching out the *unconditional parallel trends* assumption for the *conditional parallel trends* assumption

Why covariates?

- The inclusion of covariates in diff-in-diff models is not about trying to find random variation in the treatment within values of the dimension of X
- It is based on the claim that the inclusion of covariates is necessary to re-establish parallel trends
- This is itself different than how covariates will be used in synthetic control, too

Correcting the missingness problem

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

We were always missing Y^0 values for the treatment group units, but parallel trends allowed us to impute it using the change in $[Y^0]|D=0]$ as a guide

But if that trend is not a good guide, then we cannot.

Conditional parallel trends

The DiD equation yields:

$$\begin{aligned}\widehat{\delta} &= \Big(E[Y_k|Post] - E[Y_k|Pre]\Big) - \Big(E[Y_U|Post] - E[Y_U|Pre]\Big) \\ &= \text{ATT} + \text{Non-parallel trends bias}\end{aligned}$$

If we believe that conditional on covariates, parallel trends holds, but only within values of X, then there are methods we can use that incorporate covariates into the DiD equation and unbiasedness returns

The inclusion of covariates has particular regression specifications, plus there are alternative methods too, and we will review them

Picking covariates

- You want to select confounders, which are baseline untreated covariates that predict both the treatment (and thus cause imbalance) but also the potential outcome, Y^0
- You can use your common sense, logic, graphical models (i.e., DAGs) and familiarity of the subject
  - $\rightarrow$ Our focus is on the covariates that are determinants of the outcome, and less so the treatment itself, because note our concern is the untreated potential outcome Y^0
- But you can also use some simple approaches that I'll suggest now

Three ideas

- Note that we are missing $E[Y^0|D=1]$, and so we want covariates that are highly predictive of that missing potential outcome – problem is, it's missing
- But what if we did this:
  1. Drop all treated (post treatment treatment group) so that all outcomes are now Y^0 and regress that against candidate covariates, and select from there
  2. Drop post treatment treatment group units and drop the control units and regress Y^0 for the treated unit against baseline covariates and select from there
  3. Use $t-2$ and $t-1$ period, calculate $\Delta E[Y^0|D=1]$ and regress that against $t-1$ covariates and select from there
- This is a more data driven procedure and is helping you figure out which covariates might be candidate confounders – but note this still is not checking for imbalance

Covariate Balance and Parallel Trends

- The parallel trends assumption is untestable but can be evaluated using observed covariates.
- Balance in key covariates helps assess whether treated and control groups would follow similar trends.
- Differences in demographics or economic conditions between groups can indicate violations of parallel trends.
- Use covariates that are unlikely to be affected by treatment to check balance.

Baseline Covariates and Normalized Difference

- Baseline covariates are measured before treatment ($t-1$). Check for balance between treatment and control groups.
- Report the averages of covariates for both groups in a table.
- The normalized difference is calculated as:

$$\text{Norm. Diff}_{\omega} = \frac{\overline{X}_{\omega,T} - \overline{X}_{\omega,C}}{\sqrt{(S^2_{\omega,T} + S^2_{\omega,C})/2}}$$

- The normalized difference measures imbalance; it should be less than 0.25 to avoid problematic imbalance Imbens and Rubin (2015).

Normalized Difference vs. Z-Score

- **Normalized Difference:**
  - → Measures difference in group means (e.g., treatment vs. control)
  - → Interprets difference relative to pooled variance
  - → Used for comparing balance between groups
- **Z-Score:**
  - → Measures distance of a single observation from the mean
  - → Expresses this distance in standard deviation units
  - → Used for assessing how unusual a single observation is
- Both metrics standardize differences using standard deviation, making them unit-free and comparable across variables, but in our case we need to focus on differences between two groups, so we use the normalized difference and focus on a threshold of 0.25.

Table 4: Covariate Balance Statistics

	Unweighted			Weighted		
Variable	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	27.94	28.32	0.19	29.89	30.13	0.14
% White	47.43	52.50	0.62	46.07	47.76	0.21
% Hispanic	5.86	4.75	-0.14	10.06	11.35	0.13
Unemployment Rate	7.10	7.77	0.25	6.98	8.00	0.50
Poverty Rate	18.54	16.22	-0.35	17.22	15.29	-0.37
Median Income	43.38	48.00	0.41	49.28	57.83	0.68
2014 - 2013 Covariate Differences						
% Female	-0.11	-0.13	-0.06	-0.05	-0.04	0.10
% White	-0.29	-0.35	-0.14	-0.29	-0.27	0.07
% Hispanic	0.11	0.11	-0.01	0.14	0.20	0.33
Unemployment Rate	-1.09	-1.26	-0.24	-1.08	-1.36	-0.54
Poverty Rate	-0.51	-0.28	0.13	-0.41	-0.35	0.04
Median Income	1.13	1.04	-0.04	1.11	1.73	0.32

This table reports the covariate balance between adopting and non-adopting states. In the top panel, we report the averages and standardized differences of each variable, measured in 2013, by adoption status. In the bottom panel we report the average and standardized differences of the county-level long differences between 2014 and 2013 of each variable. We report both weighted and unweighted measures of the averages to correspond to the different estimation methods of including covariates in a 2 $\times$ 2 setting.

Intuition for Checking Covariate Balance

- Remember that confounders are variables that do two things
  1. $X \rightarrow D$. Confounders have different distributions in treatment than control and these balance checks are about seeing if that's true
  2. $X \rightarrow \Delta E[Y^0]$. But it's only a problem if the imbalance is on variables that also predict potential outcome trends, and we choose covariates based on things we think that's true
- So you select covariates that you think are theoretically predictive of Y^0 trends, but then you check imbalance to see if they are differentially distributed

Example: Murder

- Say that there a set of cities pass gun law ordinances but smaller towns don't and you want to know its effect on homicides
- Maybe as much as 80% of homicides are in cities – the homicides in small towns are rare, and the trends are probably pretty noisy and different
- So you might think population size, could be a very important covariate to include or urbanization

Three covariate DiD papers

Three papers (though sometimes you see others) about covariate adjustment in DiD:

1. Abadie (2005) on semiparametric DiD – reweights the comparison group part of the DID equation using a propensity score based on X
2. Heckman, Ichimura and Todd (1997) on outcome regression uses baseline X and control group only to impute the missing counterfactual Y^0 for treatment group units in a DiD equation
3. Sant'Anna and Zhou (2020) is double robust which means the method does both of these at the same time so that you don't have to choose between them

We will discuss both of them and then compare their performance with the more straightforward fixed effects model

Identification assumptions I: Data

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is possible but assumes stationarity which is a kind of stability assumption, but I'll use panel representation.

Cross-sections will be potentially violated with changing sample compositions (e.g., the Napster example).

Identification assumptions II: Modification to parallel trends

Assumption 2: Conditional parallel trends

Counterfactual trends for the treatment group are the same as the control group for all values of X

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Identification assumptions III: Common support

Assumption 3: Common support

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 $\leq 1 - e$.

Heckman, et al doesn't use the propensity score so we need a more general expression of support

Outcome regression

This is the Heckman, et al. (1997) approach where the potential outcome evolution for the treatment group is imputed with a regression based only on X_b for the control group *only*

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

where $\overline{Y}$ is the sample average of Y among units in the treatment group at time t and $\widehat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$.

Outcome regression

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

1. Regress changes ΔY on X among untreated groups using baseline covariates only
2. Get fitted values of the regression using all X from $D = 1$ only. Average those
3. Calculate change in this fitted Y among treated with the average fitted values

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\widehat{\delta}^{ipw} = \frac{1}{E_N[D]} E\left[\frac{D - \widehat{p}(X)}{1 - \widehat{p}(X)}(Y_1 - Y_0)\right]$$

where $\widehat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

These models cannot be ranked

- Outcome regression needs $\widehat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\widehat{p}(X)$ 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
- But what if you could do both of them at the same time and not pay for it?

Double Robust DR

- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one
- Two papers:
  1. Chang (2020) incorporates DR with double/debiased ML
  2. Sant'Anna and Zhau (2020) is based on the IPW (Abadie 2005) and OR (Heckman, Ichimura and Todd 1997)
- For now, I've prepped the latter, but will soon get Chang (2020) incorporated – I just have been relying on Brigham Frandsen to teach the DML material

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

$p(x)$: propensity score model

$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}(X),$ where $\mu(X)$ is a model for

$m_{d,t} = E[Y_t|D = d, X = x]$

So that means $\mu_{0,\Delta}$ is just the control group's change in average Y for each $X = x$

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

Notice how the model controls for X: you're weighting the adjusted outcomes using the propensity score

The reason you control for X twice is because you don't know which model is right. DR DiD frees you from making a choice without making you pay too much for it

Efficiency

- 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

Standard TWFE Model

Consider our earlier 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 three *more* assumptions

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

$$E[Y_1^0 - Y_0^0|D=1,X] = E[Y_1^0 - Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] - E[Y_0^0|D=1,X] = E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0^0|D=1,X] + E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0|D=1,X] + E[Y_1|D=0,X] - E[Y_0|D=0,X]$$

Switching equation substitution

Last line from the switching equation. This gives us:

$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Now compare this with our earlier Y^1 expression

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

We can define our target parameter, the ATT, now in terms of the fixed effects representation

Collecting terms

TWFE representation of our conditional expectations of the potential outcomes

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$
$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

Substitute these into our target parameter

$$\begin{aligned} ATT &= E[Y_1^1|D=1,X] - E[Y_1^0|D=1,X] \\ &= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X) \\ &= \delta + (\theta_1 X - \theta_2 X) \end{aligned}$$

What if $\theta_1 X \neq \theta_2 X$?

Assumption 4: Homogeneous treatment effects in X

TWFE requires homogenous treatment effects in X (i.e., the treatment effect is the same for all X)

If X is sex, then effects are the same for males and females.

If X is continuous, like income, then the effect is the same whether someone makes $1 or $1 million.

X-specific trends

TWFE also places restrictions on covariate trends for the two groups too. Take conditional expectations of our TWFE equation.

$$
\begin{aligned}
E[Y_1|D=1] &= \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11} \\
E[Y_0|D=1] &= \alpha_1 + \alpha_3 + \theta X_{10} \\
E[Y_1|D=0] &= \alpha_1 + \alpha_2 + \theta X_{01} \\
E[Y_0|D=0] &= \alpha_1 + \theta X_{00}
\end{aligned}
$$

X-specific trends

Now take the DiD formula:

$$\delta^{DD} = \Big((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10})\Big) - \Big((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00})\Big)$$

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

We need “no X-specific trends” for the treatment group (assumption 5) and comparison group (assumption 6)

Intuition: No X-specific trends means the evolution of potential outcome Y^0 is the same regardless of X. This would mean you cannot allow rich people to be on a different trend than poor people, for instance.

Without these six, in general TWFE will not identify ATT.

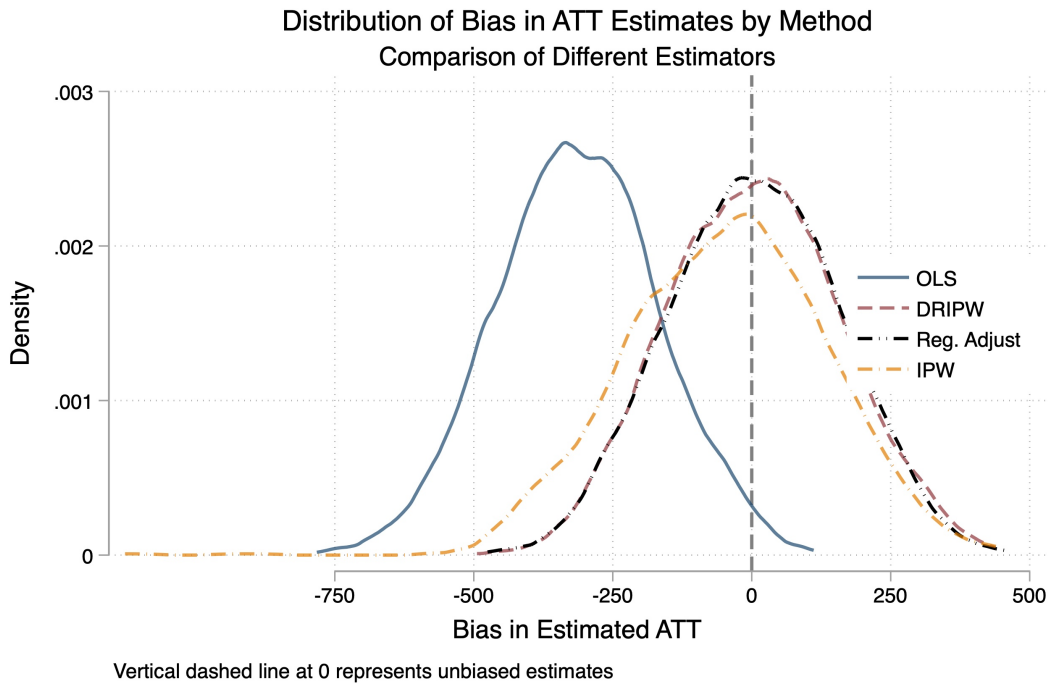
Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
  1. Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
  2. Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
  3. TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Let's look at a simulation

Simulation

- First we will look at the use of these estimators using a simulation named `covariates.do` and `covariates.R`
- We will do it both with a single run, as that's faster, and then run a simulation of 1,000 simulated regenerated data (i.e., Monte Carlo simulation) to get a distribution
- We will examine all four estimators: (1) OLS, (2) IPW, (3) OR and (4) DR

Simulation


Distribution of Bias in ATT Estimates by Method
Comparison of Different Estimators
.003
.002
.001
0
Density
-750
-500
-250
0
250
500
Bias in Estimated ATT
OLS
DRIPW
Reg. Adjust
IPW
Vertical dashed line at 0 represents unbiased estimates

R and Stata Code

There is code in R and Stata (all DiD estimators are now beautifully arranged at a website hosted by Asjad Naqvi)

- Stata: **drdid**
- R: **drdid**

https://asjadnaqvi.github.io/DiD/docs/01_stata/

Remember – it's for 2x2 with covariates (i.e., one treatment group).