

Causal Inference II

MIXTAPE SESSION



Roadmap

Introduction

Managing expectations

Diff-in-diff background

Potential outcomes

Estimation

Types of parallel trends evidence

How parallel trends can get violated

Event studies

Falsifications

Conditional parallel trends with covariates

Inverse probability weighting

Outcome Regression and Double Robust

Lalonde lab

Introductions

- Thank you for welcoming me to Egham and Royal Holloway!
- Scott Cunningham, Professor at Baylor, Waco Texas (halfway between Dallas and Austin)
- Welcome to a two day workshop on difference-in-differences and if we have time synthetic control!

What my pedagogy is like

- Long days that don't feel long because it's high energy, with regular breaks including lunch
- Move between the econometrics, history of thought, videos, applications, code, spreadsheets, exercises
- Ask questions at any point; I'll do my best to answer them

Class goals

Pedagogical goal is to break down the procedures into plain English, rebuilding it into something you can and want to use, but also:

1. **Confidence:** You will feel like you have a good enough understanding of diff-in-diff and synthetic control, both in its basics and some more contemporary issues, so that by the end of the week it a very intuitive, friendly, and useful tool
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 and synth context
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

Introduction to DiD basics

- Potential outcomes review and the ATT parameter
- DiD equation (“four averages and three differences”), parallel trends and estimation with OLS
- Evaluating parallel trends with falsifications, event studies
- Compositional changes, triple differences and covariates

Day 2 outline

Differential timing

- TWFE Pathologies in static and dynamic specifications ("event study")
- Aggregating group-time ATT

Synthetic control (non-negative and negative weighting)

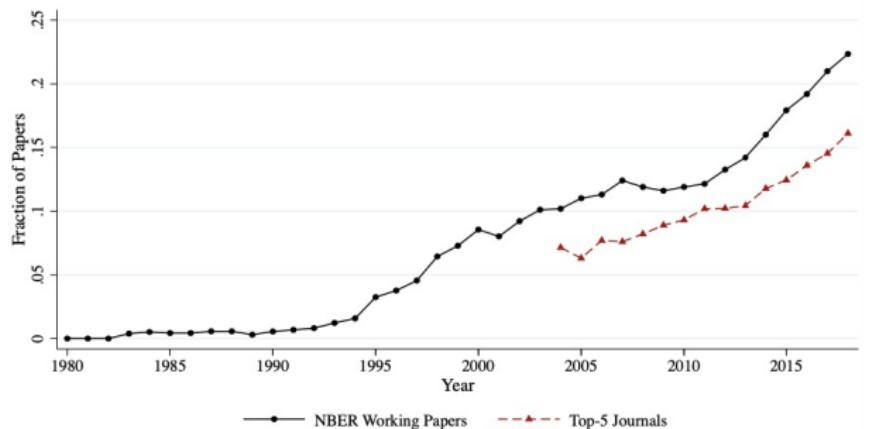
What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Mostly associated with “big shocks” happening in space over time

*“A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us.” – Daniel McFadden
(Nobel Laureate recipient with Heckman 2002)*

Figure: Currie, et al. (2020)

A: Difference-in-Differences



Origins of diff-in-diff

- Difference-in-differences (DiD) was quietly and largely unnoticed introduced in the 19th century as a way to convince skeptics in health policy arguments
- Dominant disease theory in 19th century was *miasma* – disease caused by smelly vapor
- Keep in mind – microorganisms would not be identified until much later, partly caused by poor resolution in microscopes (Freedman 2007)

Miasma I: Ignaz Semmelweis and washing hands

- 1840s, Vienna maternity wards had high postpartum infections in one wing compared to other wings
- One division had doctors and trainee doctors, but another had midwives and trainee midwives

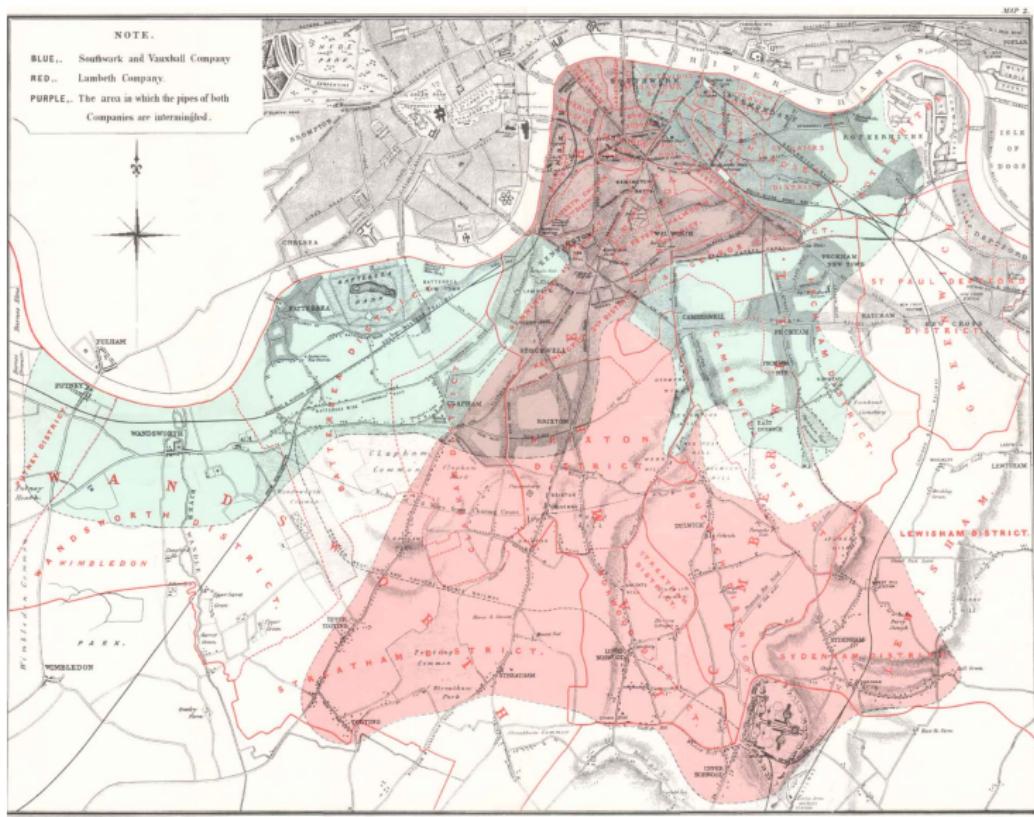
Miasma I: Ignaz Semmelweis and washing hands

- Ignaz Semmelweis notes the difference in 1841 when hospitals moved to “anatomical” training involving cadavers (Pamela Jakeila lecture notes on DiD)
- New training happens to one but not the other and Semmelweis thinks the mortality is caused by working with cadavers
- Proposes washing hands with chlorine in 1847 in the midwives’ wing and uses a DiD design of pre and post

Miasma II: John Snow and cholera

- Three major waves of cholera in the early to mid 1800s in London
- John Snow believed cholera was spread through the Thames water supply which contradicted dominant theory about “dirty air” transmission
- Grand experiment: Lambeth moves its pipe between 1849 and 1854; Southwark and Vauxhall delay
- He can evaluate the effect in three ways (one of which is DiD)

Figure: Two water utility companies in London 1854



1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\hat{\delta}_{cs} = D + (L - SV)$$

What is L and SV ?

1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\widehat{\delta}_{cs} = D + (L - SV)$$

This is biased if $L \neq SV$ (selection bias). Give an example when we're pretty sure they are equal.

2) Interrupted time series design

Table: Lambeth, 1849 and 1854

Company	Time	Cholera mortality
Lambeth	1849	$Y = L$
	1854	$Y = L + (T + D)$

$$\hat{\delta}_{its} = D + T$$

What is required for this estimator to be unbiased?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$	$T_L + D$	
	After	$Y = L + T_L + D$		
Southwark and Vauxhall	Before	$Y = SV$	T_{SV}	D
	After	$Y = SV + T_{SV}$		

$$\widehat{\delta}_{did} = D + (T_L - T_{SV})$$

This method yields an unbiased estimate of D if $T_{SV} = \textcolor{red}{T}_L$

Orley goes to Washington

- Orley Ashenfelter graduated from Princeton in the 1970s, takes a job in Washington DC and begins studying “job trainings programs”
- Empirical crisis in empirical macro and empirical labor back in the 1970s – Orley, David Card, Bob Lalonde, Alan Krueger at Princeton all helped bring attention to it and began pushing for solutions, one of which was RCTs in labor but also diff-in-diff as well as better instruments
- Listen to Orley explain the connection he made between two way fixed effects and difference-in-differences; it was born out of a need to explain OLS to an American bureaucrat

<https://youtu.be/WnB3EJ8K7lg?t=126>

Steps of a project

1. Convert research question into causal parameter
2. Deduce beliefs needed to estimate that causal parameter with data
3. Create a calculator that will use data and estimate the causal parameter

Most of us skip (1) and maybe even (2) and instead simply “run regressions” and cross our fingers that that coefficient is causal, but is it? And why is it? And what is it? Let’s dig into Orley’s comment a little more.

Equivalence

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

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

- Orley claims that the OLS estimator of δ and the “four averages and three subtractions” are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

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 counterfactual state of the world

Treatment effect definitions

Individual treatment effect

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

Missing data problem: I don't know my own counterfactual

Conditional Average Treatment Effects

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:

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

This is one of the most important policy parameters, if not the most important, and coincidentally it's also the parameter you get with diff-in-diff (even with heterogeneity)

Potential outcomes vs data

- ATT is expressed in terms of potential outcomes, but we do not use potential outcomes for estimation; we use data
- Potential outcomes are unknown and *hypothetical* possibilities describing states of the world but our data are realized outcomes, or "data", that actually occurred
- Potential outcomes become realized under treatment assignment

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

- Depending on how the treatment is assigned really dictates whether correlations reveal causal effects or bias

DiD equation

Orley's "four averages and three subtractions", or what Bacon will call the 2x2

$$\hat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

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.

Does $\hat{\delta}$ equal the ATT? If so when? If not why not?

Potential outcomes and the switching equation

$$\hat{\delta} = \underbrace{\left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)}_{\text{Switching equation}} + \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

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

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”

What is parallel trends

- Parallel trends assumes away the selection bias associated with comparisons
- The assumption is thought to be more plausible than simply assuming simple comparisons held equal
$$E[Y^0|D = 0] = E[Y^0|D = 1]$$
- But it is still a strong assumption, and differs from the assumptions have in the RCT which though also untestable, is nearly guaranteed by randomization
- Most of the hard part of the work involves the old fashioned detective work and the work of making good arguments with good exhibits (tables and figures)

Understanding parallel trends through worksheets

Before we move into regression, let's go through a simple exercise to really pin down these core ideas with simple calculations

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

OLS Specification

- Simple DiD equation will identify ATT under parallel trends
- But so will a particular OLS specification (two groups and no covariates)
- OLS was historically 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

Minimum wages

- Card and Krueger (1994) have a famous study estimating causal effect (ATT) of minimum wages on employment
- Exploited a policy change in New Jersey between February and November in mid-1990s where minimum wage was increased, but neighbor PA did not
- Using DiD, they do not find a negative effect of the minimum wage on employment which is part of its legacy today, but I mainly present it to illustrate the history and the design principles



Binyamin Appelbaum

@BCAppelbaum



Replies 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](#)

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."

But let's listen to Orley's opinion about the paper's controversy at the time. <https://youtu.be/M0tbuRX4eyQ?t=1882>

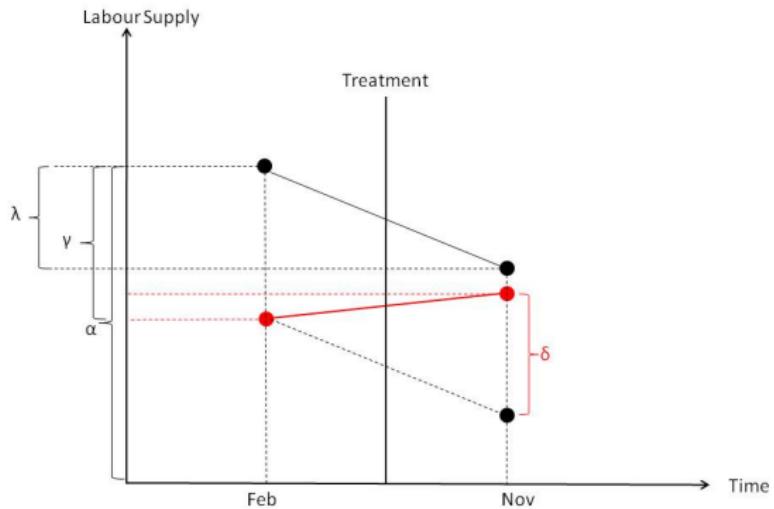
OLS specification of the DiD equation

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

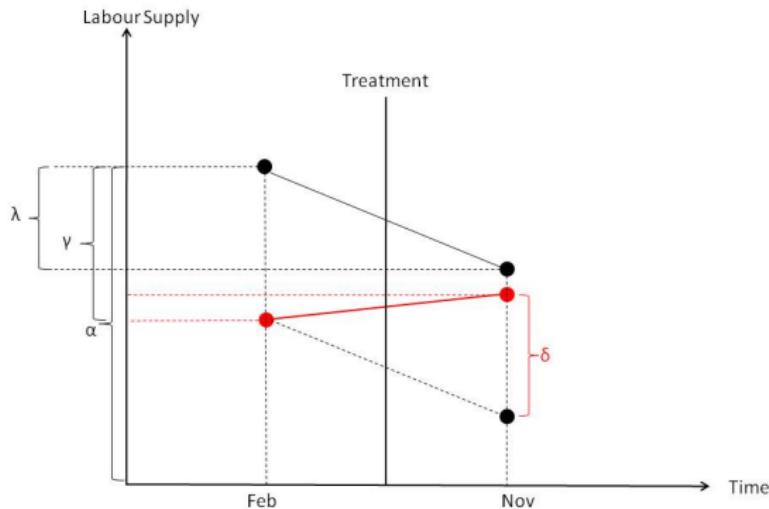
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

- 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: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$

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



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



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

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely underestimate the standard deviation of the estimators
- Standard errors are biased downward (i.e., too small, over reject)
- They proposed three solutions, but most only use one of them (clustering)

Inference

- 1 Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
- 2 Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)

Most people will simply cluster, but there are issues if you have too few clusters. They mention a third way but it's only a curiosity.

Sample code

Let's show that the "four averages and three differences" yields the same number as our saturated regression now using `equivalence.do` at github labs under "Basic DID"

Main DiD assumptions

There are actually three DiD assumptions in the basic design, but you usually only hear about the first:

1. Parallel trends – concerns changes in Y^0 , one of which is a fictional change because the post treatment Y^0 doesn't exist for the treated
2. No anticipation (next slide)
3. SUTVA (slide after next)

No Anticipation

- No anticipation means that the treatment effect happens only at the time that the treatment occurs or after, but not before
 - **Example 1:** Tomorrow I win the lottery, but don't get paid yet. I decide to buy a new house today. That violates NA
 - **Example 2:** Next year, a state lets you drive without a driver license and you know it. But you can't drive without a driver license today. This satisfies NA.
- We need this for a boring reason – baseline in the DiD *must* be Y^0 in order for DiD to equal ATT plus PT

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

Roadmap

Introduction

- Managing expectations

- Diff-in-diff background

- Potential outcomes

- Estimation

Types of parallel trends evidence

- How parallel trends can get violated

- Event studies

- Falsifications

Conditional parallel trends with covariates

- Inverse probability weighting

- Outcome Regression and Double Robust

- Lalonde lab

Violating parallel trends exercise

- Parallel trends are needed so we can impute the missing $E[Y^0|D = 1]$ with $E[Y^0|D = 0]$ either explicitly or implicitly
- Which means if parallel trends isn't true, then the imputation isn't correct and therefore estimates are biased
- To illustrate this, let's go through the document again – this time to tab 2

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

Violating parallel trends

- Parallel trends are in expectation only – we don't rely everybody to follow the same trend, just that the group average for Y^0 be approximately the same for treated and control
- Violations are a form of selection bias and there are two straightforward ways that parallel trends will be violated
 1. Compositional differences in samples associated with repeated cross-sections
 2. Policy endogeneity

Repeated cross-sections and compositional change

- One of the risks of a repeated cross-section is that the composition of the sample may have changed between the pre and post period in ways that are correlated with treatment
- Hong (2013) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napster (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX

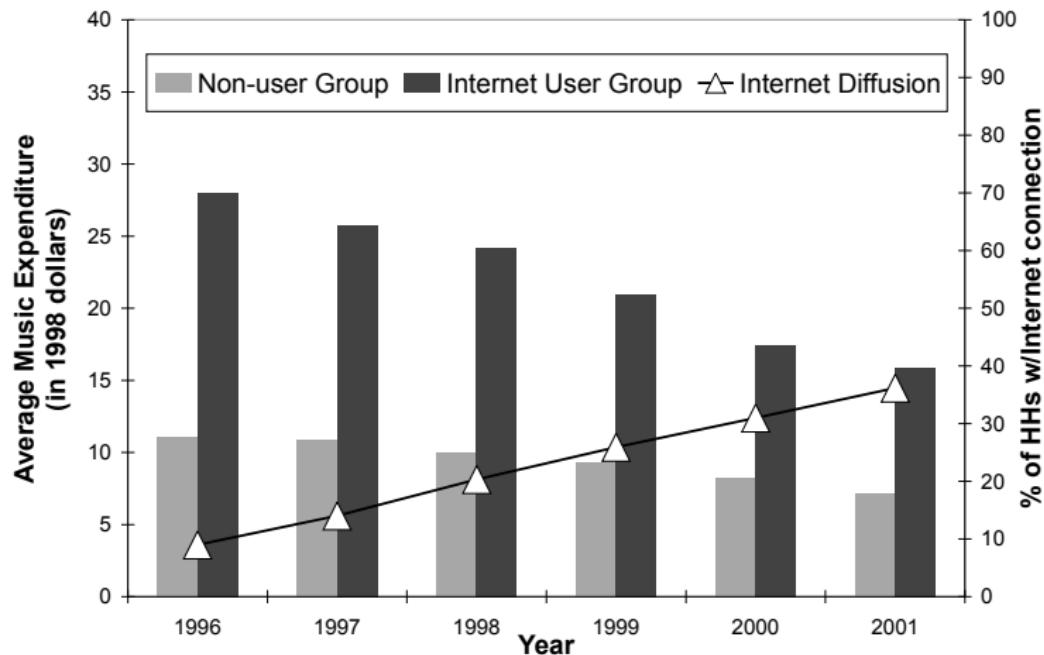


Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	\$52,887	\$30,459	\$51,995	\$28,169	\$49,970	\$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the Internet changes samples (e.g., younger music fans are early adopters)

Repeated cross-sections

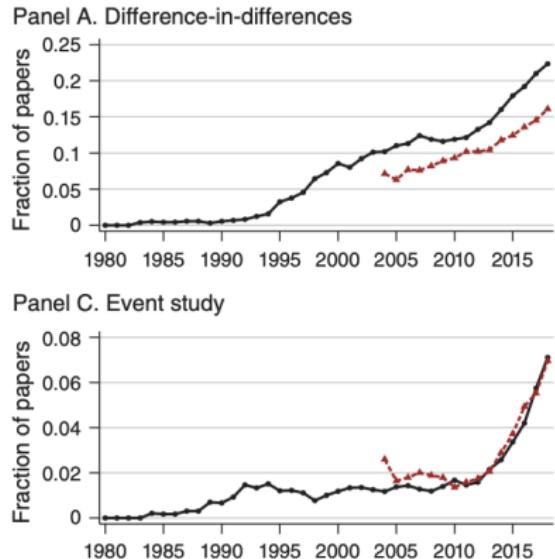
- Surprisingly underappreciated problem with almost no literature around it
- So what can you do? Check covariate balance by regressing the time-varying covariates instead of the outcome onto the treatment using your OLS specification
- They should be exogenous remember, so this covariate regression can be a helpful test of whether this is a problem
- “Difference-in-differences with Compositional Changes” by Pedro Sant’anna and Qi Xu (not yet released) is the only paper I’ve ever seen to look into it

Types of evidence

What kinds of evidence is convincing?

- You are building a case, the prosecutor before a judge and jury, always in battle with the defense attorney
- Evidence has particular broadly defined forms that can help you on the front end
- Your goal in my humble opinion should be reasonable and theoretically relevant falsifications with particular kinds of data visualization, starting with the event study

Event studies have become mandatory in DiD



Event studies are about pre-trends not parallel trends

- Parallel trends involves, Y^0 , specifically
 $\Delta E[Y^0|D = 1] = \Delta E[Y^0|D = 0]$
 - Notice that parallel trends is about Y^0 in other words, not Y^1
- We cannot verify the red term, because the change is post-treatment and thus counterfactual (fictional)
- But there are other non-red $\Delta E[Y^0|D = 1]$ that aren't fictional which we can investigate, but where?

"Pre-trends" are also $\Delta E[Y^0|D = 1]$, just non-fictional in nature

Testing for parallel pre-trends is a type of falsification for selection bias

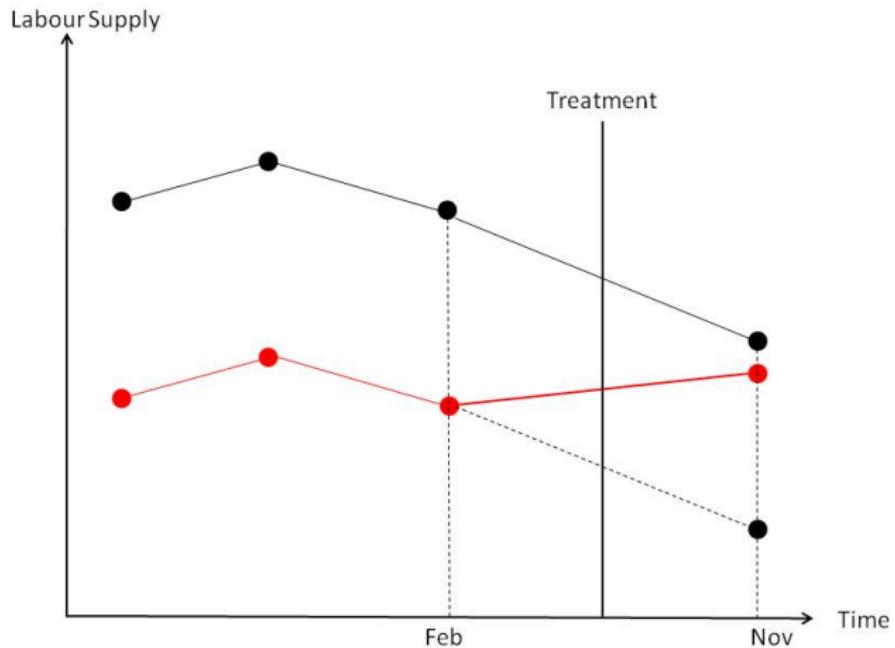
Intuition behind event studies

- Checking pre-trends is **not** a test for parallel trends as there is no formal test for parallel trends
- It's akin to finding a smoking gun – maybe someone planted it, but dismiss it is irresponsible
- Do not overweight nor underweight parallel pre-trends
- 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)

How do we do it?

- Formula is “four averages and three differences” on each lead and lag relative to baseline so calculate it using OLS or manually
- Some prefer to also plotting raw data

Plot the raw data when there's only two groups



Event study regression

- Event studies have a simple OLS specification with only one treatment group and one never-treated group

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

- where D is an interaction of the treatment dummy with the calendar year
- Treatment occurs in year 0, no anticipation, drop baseline $t - 1$
- All “four averages and three differences” calculations will use $t - 1$ as “pre” which is why it must be untreated (no anticipation)
- Includes q leads or anticipatory effects and m lags or post treatment effects

Event study regression

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

Typically you'll plot the coefficients and 95% CI on all leads and lags
(binned or not, trimmed or not)

Under no anticipation, then you expect $\hat{\mu}$ coefficients to be zero, which gives you confidence that parallel trends holds (but is not a guarantee, and there are still specification issues – see Jon Roth's work)

Under parallel trends, $\hat{\delta}$ are estimates of the ATT at points in time

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

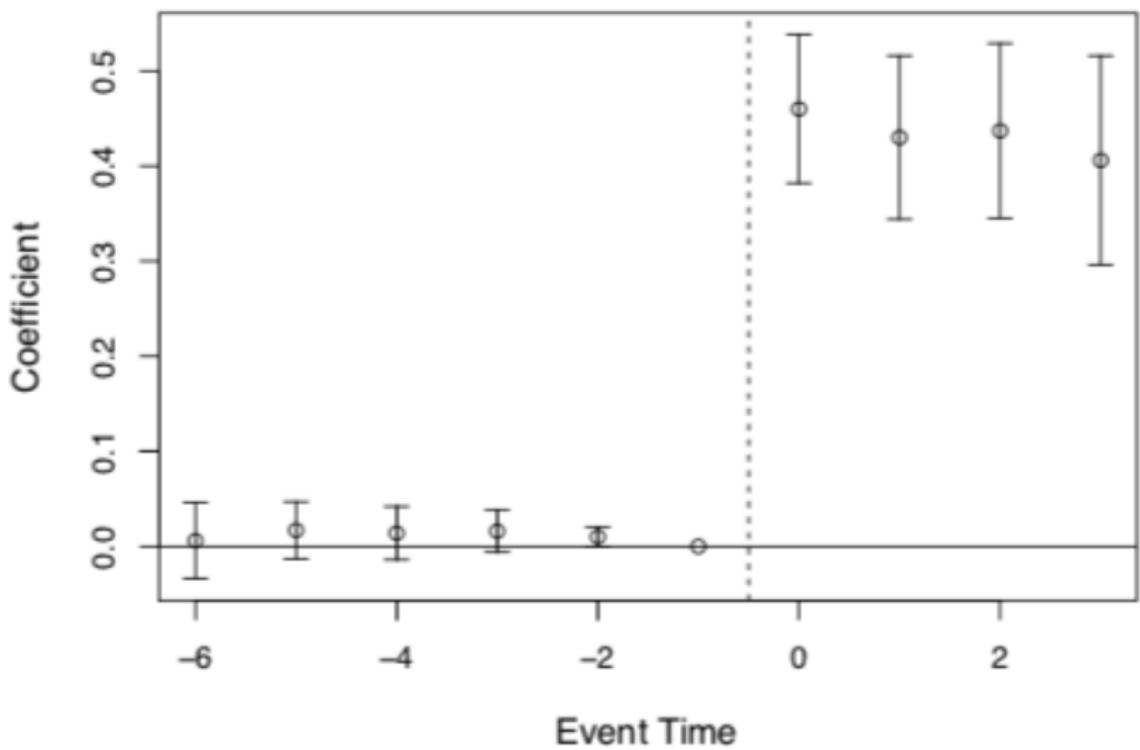
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults 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 nonexpander states. 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

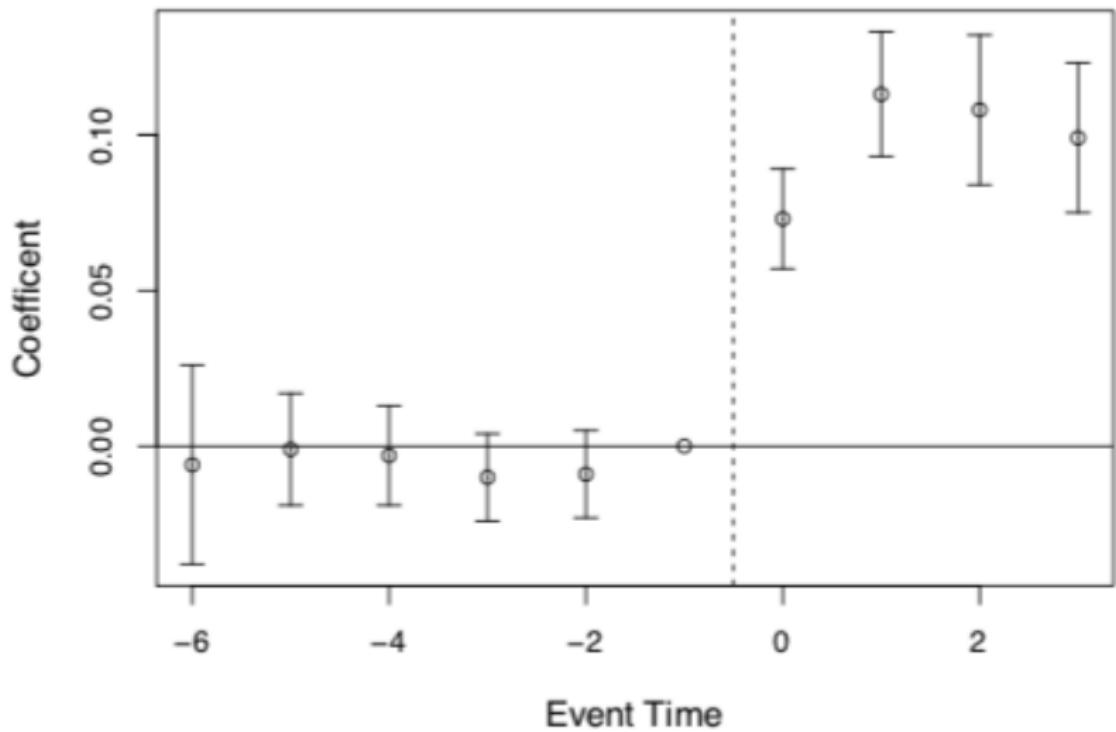
Issue Section: Article

Types of evidence

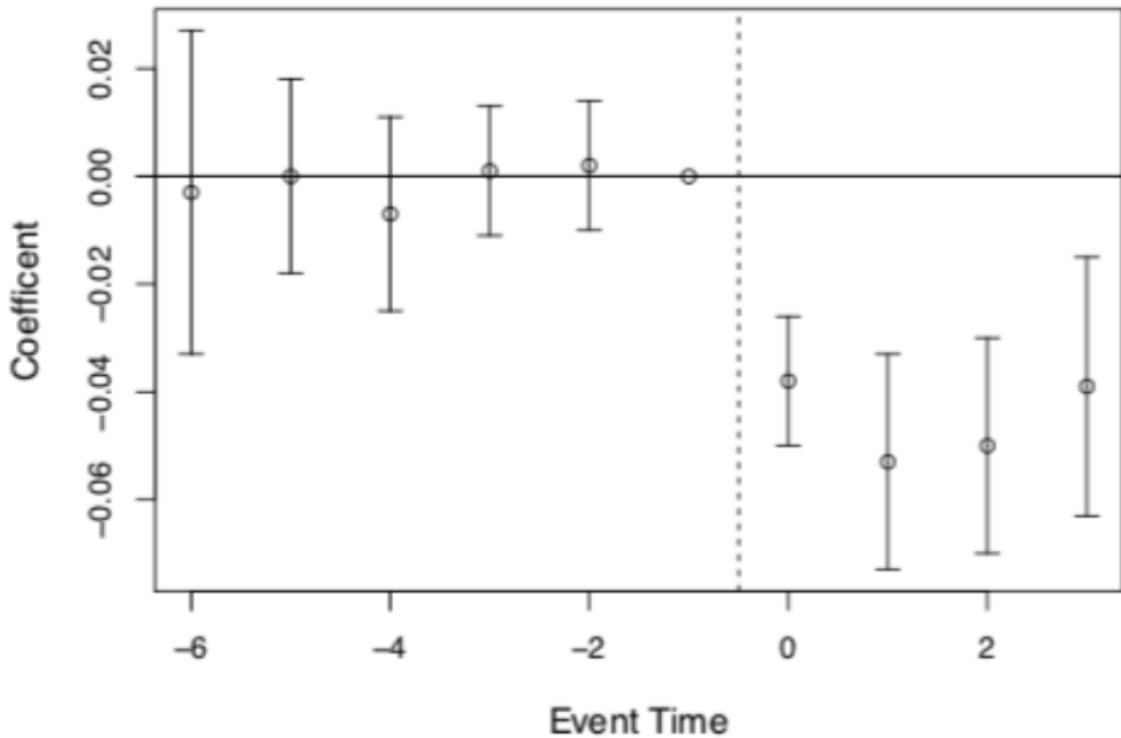
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Main Results** – Show your main results (the point of the paper)
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Mechanisms** – Find some reason explaining why the treatment affects the outcome via some “mechanism”
- **Event study** – Show leads and lags on mortality



(a) Medicaid Eligibility



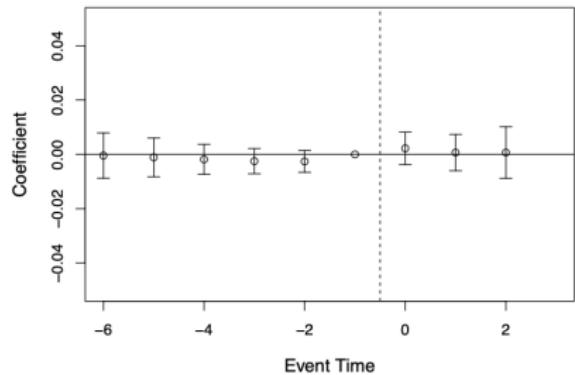
(b) Medicaid Coverage



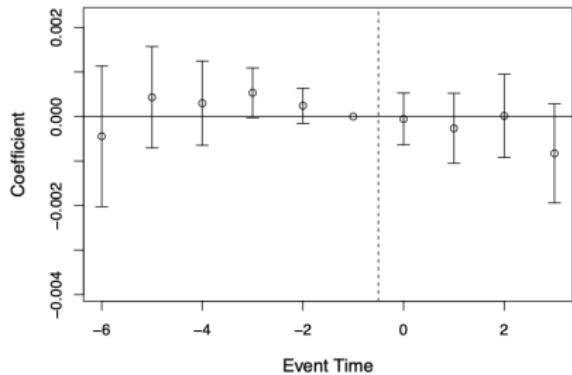
(c) Uninsured

Falsifications on elderly

Age 65+ in 2014



(c) Medicaid Coverage



(d) Annual Mortality

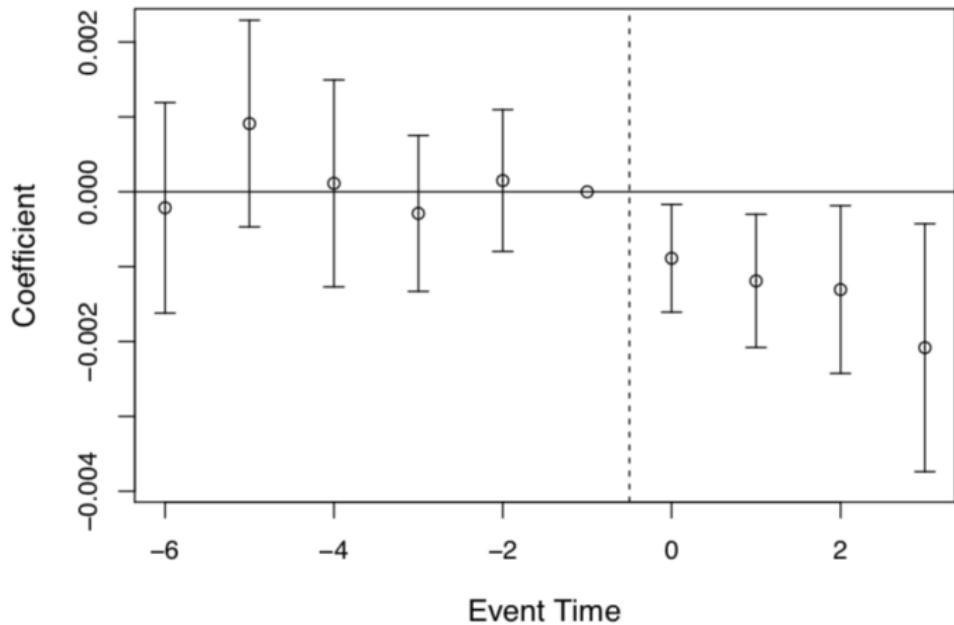


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on annual mortality

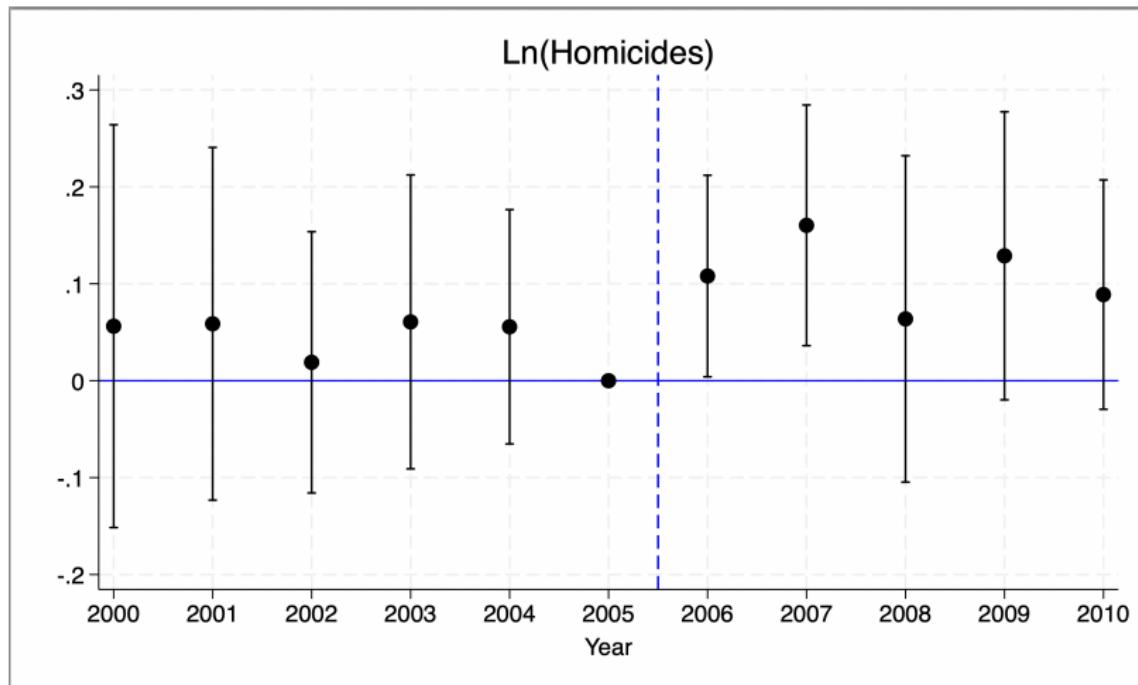
Reviewing the evidence

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

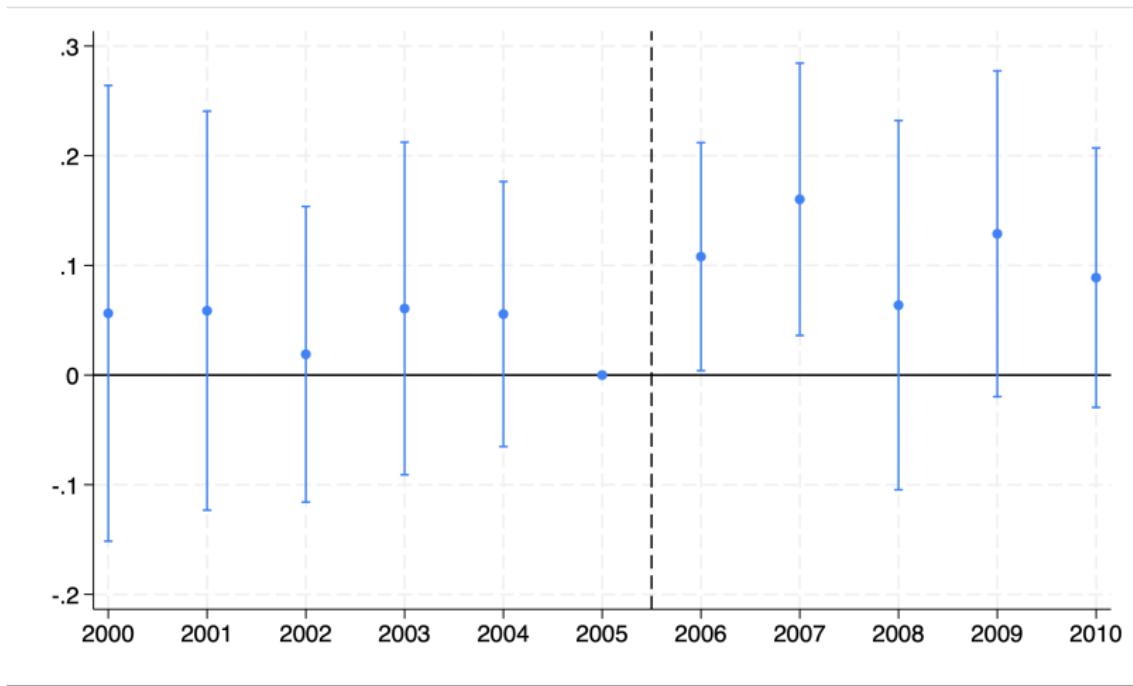
Making event study

- All the simple event study is an interaction between the treatment group dummy and the calendar year dummies
- You must drop a $t - \tau$ as the baseline (e.g., $t - 1$) must be Y^0 untreated comparisons recall
- I have included in a do file that will do it for you either manually or using coefplot in `simple_eventstudy.do` at github labs

Manually creating the event study



Creating the event study with Ben Jann's coefplot



What are falsifications

- Falsifications are indirect evidence that parallel trends probably holds by testing competing hypotheses directly
- These are outcomes that shouldn't be affected ("falsification on unaffected outcomes") or groups that shouldn't be affected ("falsification on comparison groups")
- Good falsifications require a shared belief they are good falsifications (requires buy-in)
- Can be hard if you work in an obscure area in which you have deep institutional knowledge but that's the art of the paper

Falsifications on more comparison groups

- Very common for readers and others to request a variety of “robustness checks” from a DD design
- We saw some of these just now (e.g., falsification test using data for alternative control group, the Medicare population)
- Triple differences is a formal design based on the idea that another group is unaffected by which is treated with competing forces
- Not a true falsification because the diff-in-diff need not be zero – triple diff has its own parallel trends assumption

Triple differences by Gruber (1995)

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES
ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. Treatment Individuals: Married Women, 20–40 Years Old:			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:		−0.062 (0.022)	
B. Control Group: Over 40 and Single Males 20–40:			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:		−0.008 (0.014)	
DDD:		−0.054 (0.026)	

Table: Difference-in-Difference-in-Differences numerical example

States	Group	Period	Outcomes	D_1	D_2	D_3	
Treatment states	Married women, 20-40yo	After	$NJ + T + NJ_t + wt + D$	$T + NJ_t + wt + D$	$D + wt - mt$	\hat{D}	
		Before	NJ				
	Single men 20-40yo	After	$NJ + T + NJ_t + st$	$T + NJ_t + mt$	$w_t - mt$		
		Before	NJ				
Comparison states	Married women, 20-40yo	After	$PA + T + PA_t + wt$	$T + PA_t + wt$	$w_t - mt$		
		Before	PA				
	Single men 20-40yo	After	$PA + T + PA_t + mt$	$T + PA_t + mt$	$w_t - mt$		
		Before	PA				

Triple diff estimation and identification:

Estimation: Eight averages and seven differences because it's differencing two diff-in-diff

Parallel trend assumption: $wt - mt$ for experimental states equals $w_t - mt$ for non-experimental states

Intuition: Whatever the difference in change in wages for men and women in PA, that's the change in wages gap in NJ counterfactual (which we don't know)

DDD in Regression

$$\begin{aligned} Y_{ijt} = & \alpha + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ & + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt} \end{aligned}$$

- Your panel is now a group j state i (e.g., AR women 1991, AR women 1992, etc.)
- Each term is a dummy relative to some baseline (pre), omitted group (men) and treatment state (PA)
- Most use DDD by assuming DD for control must be zero, but as we saw, formally it just has its own assumption and that need not be zero
- Assumption is that its change is the fictional change of our treated group, just like always

Great new paper to learn more



Econometrics Journal (2022), volume 00, pp. 1–23.
<https://doi.org/10.1093/econj/utac010>

The triple difference estimator

ANDREAS OLDEN AND JARLE MØEN

*Dept. of Business and Management Science, NHH Norwegian School of Economics, Hellevn.
30, N-5045 Bergen, Norway.*
Email: andreasolden@gmail.com, jarle.moen@nhh.no

First version received: 14 May 2020; final version accepted: 10 May 2021.

Summary: Triple difference has become a widely used estimator in empirical work. A close reading of articles in top economics journals reveals that the use of the estimator to a large extent rests on intuition. The identifying assumptions are neither formally derived nor generally agreed on. We give a complete presentation of the triple difference estimator, and show that even though the estimator can be computed as the difference between two difference-in-differences estimators, it does not require two parallel trend assumptions to have a causal interpretation. The reason is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. This requires only one parallel trend assumption to hold.

Keywords: DD, DDD, DID, DiDID, difference-in-difference-in-differences, difference-in-differences, parallel trend assumption, triple difference.

JEL Codes: C10, C18, C21.

1. INTRODUCTION

The triple difference estimator is widely used, either under the name ‘triple difference’ (TD) or the name ‘difference-in-difference-in-differences’ (DDD), or with minor variations of these spellings. Triple difference is an extension of double differences and was introduced by Gruber (1994). Even though Gruber’s paper is well cited, very few modern users of triple difference credit him for his methodological contribution. One reason may be that the properties of the triple difference estimator are considered obvious. Another reason may be that triple difference was little more than a curiosity in the first ten years after Gruber’s paper. On Google Scholar, the annual number of references to triple difference did not pass one hundred until year 2007. Since then, the use of the estimator has grown rapidly and reached 928 unique works referencing it in the year 2017.¹

Looking only at the core economics journals *American Economic Review* (AER), *Journal of Political Economy* (JPE), and *Quarterly Journal of Economics* (QJE), we have found 32 articles using triple difference between 2010 and 2017, see Table A1 in Appendix A. A close reading of these articles reveals that the use of the triple difference estimator to a large extent rests on

¹ More details on the historical development of the use of the triple difference estimator can be found in the working paper version of Olden and Møen (2020, fig. 1). In the working paper, we also analyse naming conventions and suggest that there is a need to unify terminology. We recommend the terms ‘triple difference’ and ‘difference-in-difference-in-differences’.

Falsification on outcomes

- The within-group control group (DDD) is a form of placebo analysis using the same *outcome*
- But there are also placebos using a *different outcome* – but you need a hypothesis of mechanisms to figure out what is in fact a *different outcome*
- Figure out what those are, and test them – finding no effect on placebo outcomes tends to help people your other results interestingly enough
- Cheng and Hoekstra (2013) examine the effect of castle doctrine gun laws on non-gun related offenses like grand theft auto and find no evidence of an effect

Rational addiction as a placebo critique

Sometimes, an empirical literature may be criticized using nothing more than placebo analysis

"A majority of [our] respondents believe the literature is a success story that demonstrates the power of economic reasoning. At the same time, they also believe the empirical evidence is weak, and they disagree both on the type of evidence that would validate the theory and the policy implications. Taken together, this points to an interesting gap. On the one hand, most of the respondents claim that the theory has valuable real world implications. On the other hand, they do not believe the theory has received empirical support."

Placebo as critique of empirical rational addiction

- Auld and Grootendorst (2004) estimated standard “rational addiction” models (Becker and Murphy 1988) on data with milk, eggs, oranges and apples.
- They find these plausibly non-addictive goods are addictive, which casts doubt on the empirical rational addiction models.

Placebo as critique of peer effects

- Several studies found evidence for “peer effects” involving inter-peer transmission of smoking, alcohol use and happiness tendencies
- Christakis and Fowler (2007) found significant network effects on outcomes like obesity
- Cohen-Cole and Fletcher (2008) use similar models and data and find similar network “effects” for things that aren’t contagious like acne, height and headaches
- Ockham’s razor - given social interaction endogeneity (Manski 1993), homophily more likely explanation

Roadmap

Introduction

Managing expectations

Diff-in-diff background

Potential outcomes

Estimation

Types of parallel trends evidence

How parallel trends can get violated

Event studies

Falsifications

Conditional parallel trends with covariates

Inverse probability weighting

Outcome Regression and Double Robust

Lalonde lab

Covariates and violations

- There is an assumption called “unconfoundedness”

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

- It means that within the dimensions of X (e.g., Asian males aged 45), D is assigned to units independent of their potential outcomes or any combination of them (e.g., treatment effects)
- It's the basis for running regressions with covariates in order to recover aggregate causal parameters outside of the experiment but it claims that with the inclusion of the covariates, you have isolated a randomized experiment
- We usually motivate this assumption in diff-in-diff, too, but it is technically not what is going on

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

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

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:

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

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

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

Semiparametric DiD

Abadie (2005) proposed a model that simply reweights the control group in the DiD equation using a particular specification (“semiparametric”) of the propensity score on pretreatment covariates

1. Calculate each unit’s “after minus before” (DiD equation)
2. Estimate the conditional probability of treatment based on baseline covariates (propensity score estimation)
3. Weight the comparison group’s DiD equation with the propensity score

Remember – ATT is only missing Y^0 for treatment, so we only have to apply weights to the comparison group units

Novel elements of time in Abadie's model

- There is only one treatment group so therefore there is only one relevant treatment date, t
- The period prior to treatment is called the baseline, or b , period and it is when treated units were not treated
- X_b are “baseline” covariates meaning the value of X in the pre-treatment period for either the treated or comparison group units
- Propensity scores are estimated off the b period *only*
- Abadie “throws away” covariates after treatment because this is all about re-establishing parallel trends which is a *baseline* concept recall

Assumptions

Three main assumptions

1. Conditional parallel trends

$$E[Y_t^0 - Y_b^0 | D = 1, X_b] - E[Y_t^0 - Y_b^0 | D = 0, X_b]$$

2. Common support

$$Pr(D = 1) > 0; Pr(D = 1 | X) < 1$$

3. Propensity score model is properly specified

Propensity scores as dimension reduction

- Propensity scores are ways of dealing with a conditioning set X that has large dimensions
- Dimensions are not the same as covariates – if you have continuous X , then it has infinite dimensions
- Common support means that *within* all combinations of the covariates (e.g., white male 47yo versus whites, males, age) there are units in treatment and control

Common support example

Think of common support like “exact matches” but on the propensity score

I'm a white male 47 years old with a PhD; can I find a white male 47 years old without a PhD

If I can, that's common support; if I cannot that's off support

Propensity scores as dimension reduction

- Propensity score theorem (Rosenbaum and Rubin 1983) showed that if you need X to satisfy some assumption, the propensity score will satisfy too
- Propensity scores essentially transform your large dimensional problem into a single scalar called the propensity score, which is the conditional probability of treatment (conditional on X)
- But we need to estimate the propensity score because we don't usually know it (only an experimentalist "knows" the true propensity score)

Common support and the propensity score

- Exact matches mean you have people who are identical on covariate values in both treatment and control
- Common support and the propensity score means you have people nearly identical on their probability of treatment
- I am 47yo white male with a PhD with a propensity score of 0.75, but you are an Asian female 27yo without a PhD and have a propensity score of 0.75
- Same idea, but for this to work, we need to have “matches” like that (just on the propensity score)

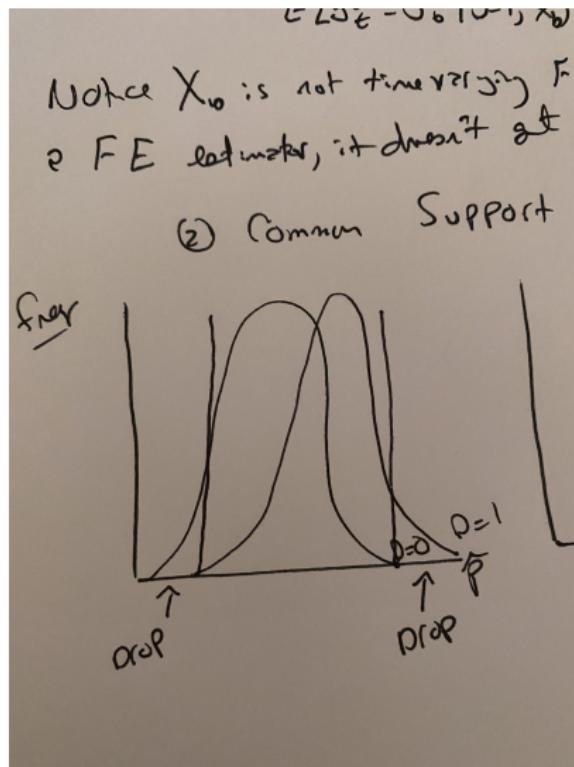
How do these work together?

Since we are identifying the ATT, and the ATT is missing Y^0 for the treated group, we are using the control group Y^0 in its place

Under conditional parallel trends and common support, some of the comparison group units are recovering the parallel trends because of their X values creating projections that in their differences perfectly aligned in expectation with the missing $\Delta E[Y^0|D = 1]$

But we have to have all three for it to work

Visualizing propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

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

Abadie's inverse probability weighting (IPW) estimator

$$E \left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)} \right]$$

The first is our causal parameter; the second is our reweighted DiD equation that estimates our causal parameter, but we need to estimate that propensity score

Abadie's IPW estimator

Look closely; what happens mathematically when you substitute $D = 1$ vs $D = 0$?

$$E \left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)} \right]$$

The reweighting with the propensity only happens to the comparison group's first differences – not the treatment groups! Why? Because it's the Y^0 that is missing, not the Y^1

Propensity scores

- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

Abadie 2005 influence



Alberto Abadie

Semiparametric difference-in-differences estimators

Authors Alberto Abadie

Publication date 2005/1/1

Journal The Review of Economic Studies

Volume 72

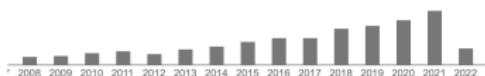
Issue 1

Pages 1-19

Publisher Wiley-Blackwell

Description The difference-in-differences (DID) estimator is one of the most popular tools for applied research in economics to evaluate the effects of public interventions and other treatments of interest on some relevant outcome variables. However, it is well known that the DID estimator is based on strong identifying assumptions. In particular, the conventional DID estimator requires that, in the absence of the treatment, the average outcomes for the treated and control groups would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between the treated and the untreated. That would be the case, for example, if selection for treatment is influenced by individual-transitory shocks on past outcomes (Ashenfelter's dip). This article considers the case in which differences in observed ...

Total citations Cited by 2330



Scholar articles Semiparametric difference-in-differences estimators

A Abadie - The Review of Economic Studies, 2005

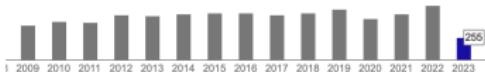
Cited by 2330 Related articles All 12 versions

Abadie (2005) is his fourth most cited paper

Outcome Regression Paper

Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme

Authors James J Heckman, Hidehiko Ichimura, Petra E Todd
Publication date 1997/10/1
Journal The review of economic studies
Volume 64
Issue 4
Pages 605-654
Publisher Wiley-Blackwell
Description This paper considers whether it is possible to devise a nonexperimental procedure for evaluating a prototypical job training programme. Using rich nonexperimental data, we examine the performance of a two-stage evaluation methodology that (a) estimates the probability that a person participates in a programme and (b) uses the estimated probability in extensions of the classical method of matching. We decompose the conventional measure of programme evaluation bias into several components and find that bias due to selection on unobservables, commonly called selection bias in econometrics, is empirically less important than other components, although it is still a sizeable fraction of the estimated programme impact. Matching methods applied to comparison groups located in the same labour markets as participants and administered the same questionnaire eliminate much of the bias as conventionally ...
Total citations Cited by 9508



Heckman, Ichimura and Todd (1997) is Petra and Hide's most cited paper and Heckman's second most cited!

Doubly Robust Paper

Doubly robust difference-in-differences estimators

Authors Pedro HC Sant'Anna, Jun Zhao

Publication date 2020/11/1

Journal Journal of Econometrics

Volume 219

Issue 1

Pages 101-122

Publisher North-Holland

Description This article proposes doubly robust estimators for the average treatment effect on the treated (ATT) in difference-in-differences (DID) research designs. In contrast to alternative DID estimators, the proposed estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified. We also derive the semiparametric efficiency bound for the ATT in DID designs when either panel or repeated cross-section data are available, and show that our proposed estimators attain the semiparametric efficiency bound when the working models are correctly specified. Furthermore, we quantify the potential efficiency gains of having access to panel data instead of repeated cross-section data. Finally, by paying particular attention to the estimation method used to estimate the nuisance parameters, we show that one can sometimes construct doubly robust DID ...

Total citations Cited by 398



Sant'Anna and Zhao (2020) is Pedro's second most cited paper

Doubly Robust Difference-in-differences

- DR models control for covariates twice – once using the propensity score, once using outcomes adjusted by regression – and are unbiased so long as:
 - The regression specification for the outcome is correctly specified
 - The propensity score specification is correctly specified
- Sant'Anna and Zhao (2020) incorporated DR into DiD by combining inverse probability weighting and outcome regression into a single DiD model
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so it merits close study

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

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997 (will require correct model too)
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005) (will require correct model too)

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*

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

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

Outcome regression

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

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

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

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

These models cannot be ranked

- Outcome regression needs $\hat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\hat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified

TWFE

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.

$$E[Y_1|D = 1] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}$$

$$E[Y_0|D = 1] = \alpha_1 + \alpha_3 + \theta X_{10}$$

$$E[Y_1|D = 0] = \alpha_1 + \alpha_2 + \theta X_{01}$$

$$E[Y_0|D = 0] = \alpha_1 + \theta X_{00}$$

X-specific trends

Now take the DiD formula:

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

Eliminating terms, we get:

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

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

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
- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one

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
- Let's skip to Monte Carlos

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

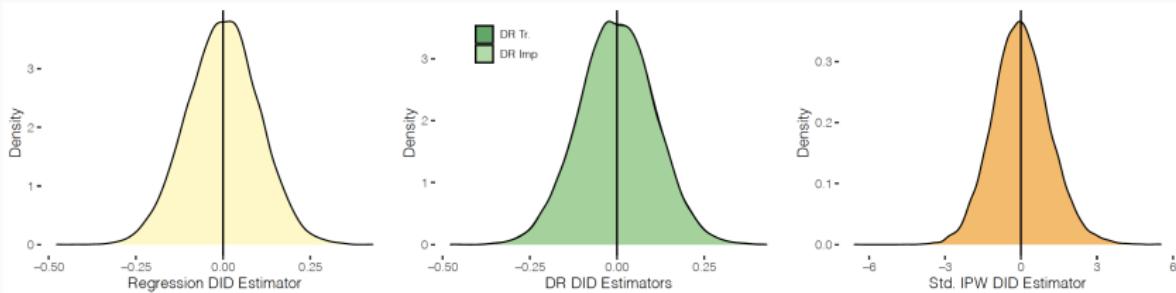
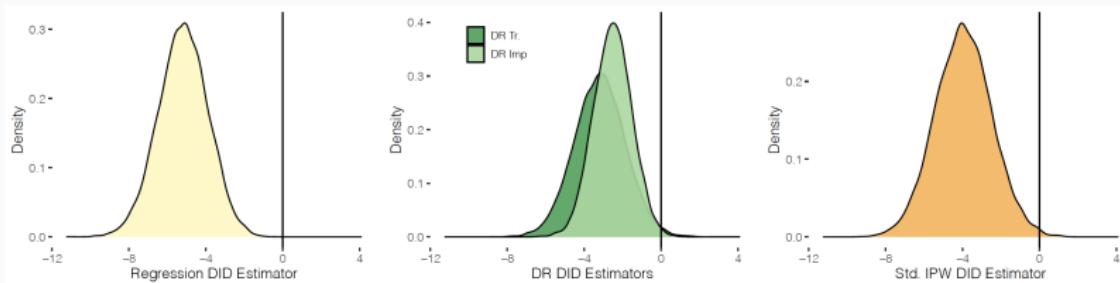


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	RMSE	SE	Coverage	CI length
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



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).

Application using real data

- Let's now use a real example with real data and see how well this does
- Famous paper in AER by Lalonde (1986), an Orley and Card student at Princeton
- Found that most program evaluation did badly, but let's revisit it with diff-in-diff

Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
 - **Treatment group**: received all the benefits of NSW program
 - **Control group**: left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

NSW Program

- Treatment group members were:
 - guaranteed a job for 9-18 months depending on the target group and site
 - divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
 - paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

NSW Program

- Other details about the NSW program:
 - Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
 - Post-treatment: after their term expired, they were forced to find regular employment
 - Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

NSW Data

- NSW data collection:
 - MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
 - Conducted up to 4 post-baseline interviews
 - Different sample sizes from study to study can be confusing, but has simple explanations

NSW Data

- Estimation:
 - NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$\frac{1}{N_t} \sum_{D_i=1} Y_i - \frac{1}{N_c} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e., $(Y^0, Y^1) \perp\!\!\!\perp D$.

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful – $\approx \$900$ (LaLonde 1986) to $\$1,800$ (Dehejia and Wahba 2002) depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of \$900

LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
 - His estimates were way off and usually the wrong sign
 - Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^a The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^b Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^c The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^d See Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	(\$455 (704))	(\$697 (760))	(\$509 (967))	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^a The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^b Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^c The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^d See Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

covariate	All		CPS	NSW	t-stat	diff
			Controls	Trainees		
	N _c	= 15,992	N _t	= 297		
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Lab

[https://github.com/Mixtape-Sessions/Causal-Inference-2/
tree/main/Lab/Lalonde](https://github.com/Mixtape-Sessions/Causal-Inference-2/tree/main/Lab/Lalonde)

Together let's do questions 1 and 2a-c

Concluding remarks

- So we hopefully see a few of the key elements of DiD
 - Remember: the DiD equation and ATT equation are distinct concepts and definitions
 - DiD designs can be implemented with OLS specifications that calculate differences in means
 - Parallel pre-trends and parallel trends are not the same thing – the first is testable, the latter is not testable
 - Event studies are mandatory but pre-trends are smoking guns, but can mislead nonetheless
- Including *time-varying* covariates in the canonical OLS specification requires additional assumptions
- Doubly robust and IPW incorporate covariates through propensity scores and outcome regressions (or both) using baseline covariate means only