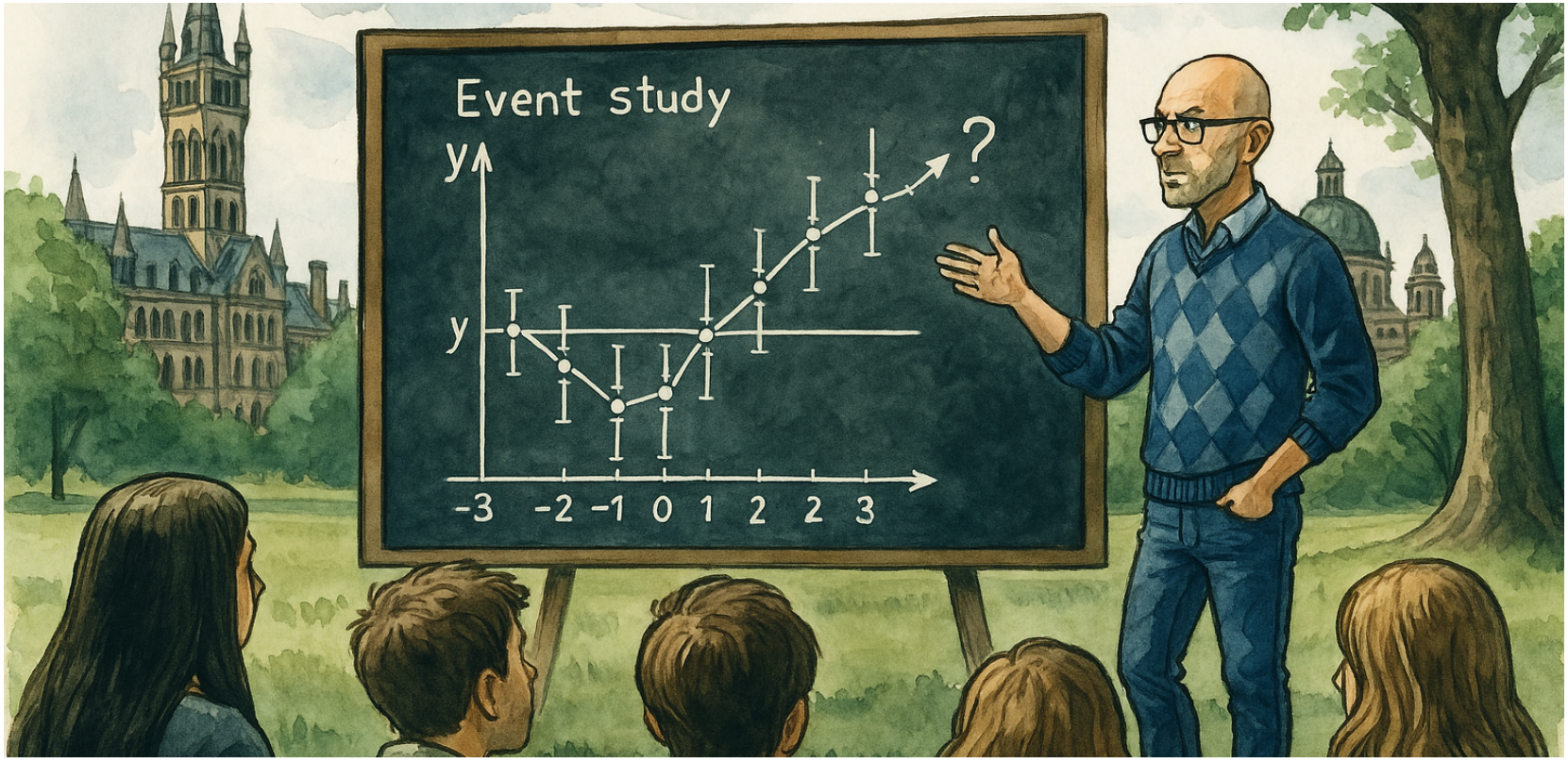
University of Glasgow

2025



Roadmap

Unconditional Parallel Trends

- Unconditional parallel trends is diff-in-diff on Y^0 in which the diff-in-diff equals zero

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

- There are two kinds of parallel trends: unconditional and conditional
- What is the difference between them?

Unconditional Parallel Trends

- Unconditional parallel trends is very close to assuming that the the treatment was as good as random
- It is not testable, but suspicion is raised when pre-treatment levels are very different as that implies non-random treatment assignment mechanisms
- Think of these are more like 'preponderance of evidence' than proof as parallel trends *cannot* be tested and graphs can be misleading

Different Falsifications

1. Falsifications on same outcomes, and similar, but untreated, groups
2. Falsifications on different impossible outcomes, but same treatment groups
3. Event study graphical plots of pre-trends and post-trends

Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends
- Parallel pre-trends are not the same as parallel counterfactual post-trends, but this is the smoking gun we typically look for nonetheless
- Pre-period falsifications are common in causal inference, even outside of diff-in-diff, because the pre-period probably has the same confounder structure, but no treatment

Event study regression

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

- 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

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
- **Falsifications** – 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

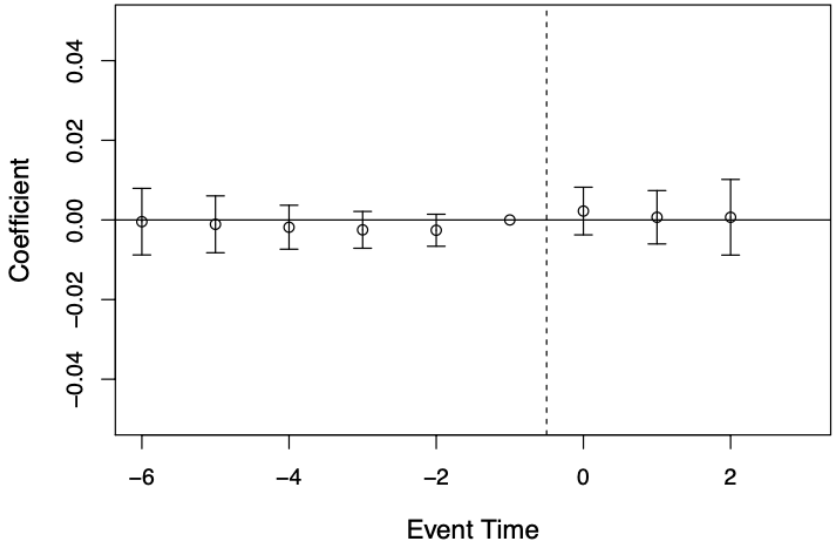
(c) Uninsured

Falsification

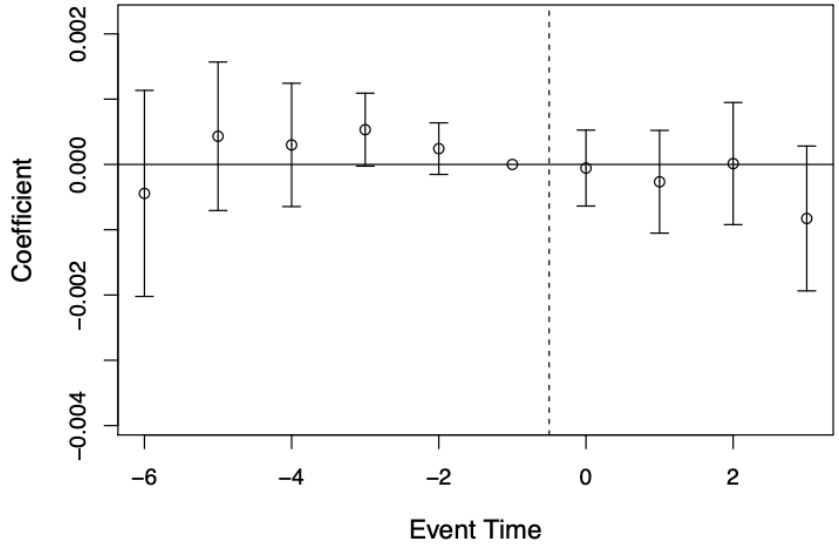
- Their study focuses on “near elderly”, which means just under 65
- They choose just under 65 because in the US, 65 and older are eligible for Medicare so more generous Medicaid is irrelevant
- *But* probably the near elderly and the elderly are equally susceptible to unobserved factors correlated with 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

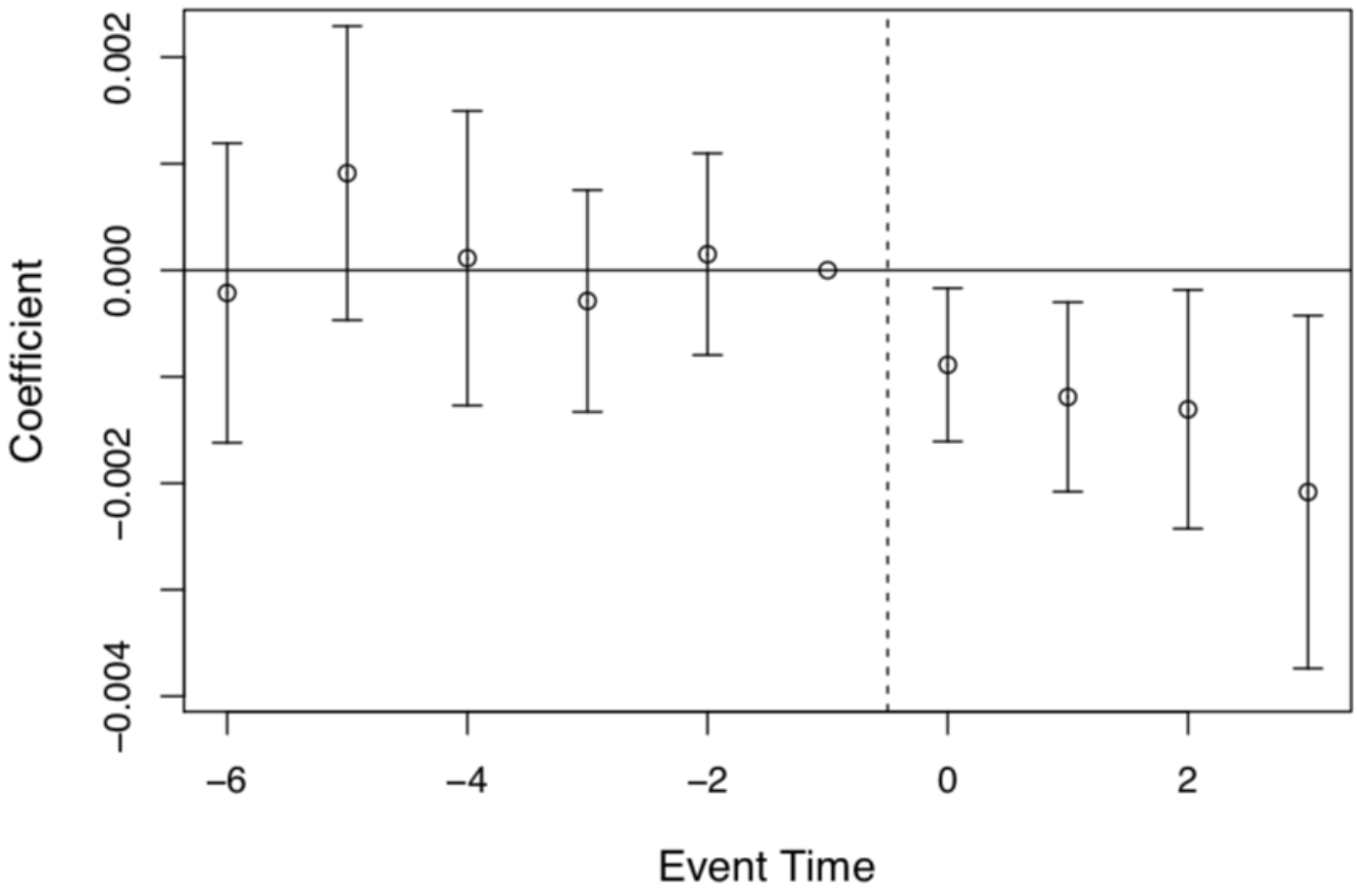


(d) Annual Mortality

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."

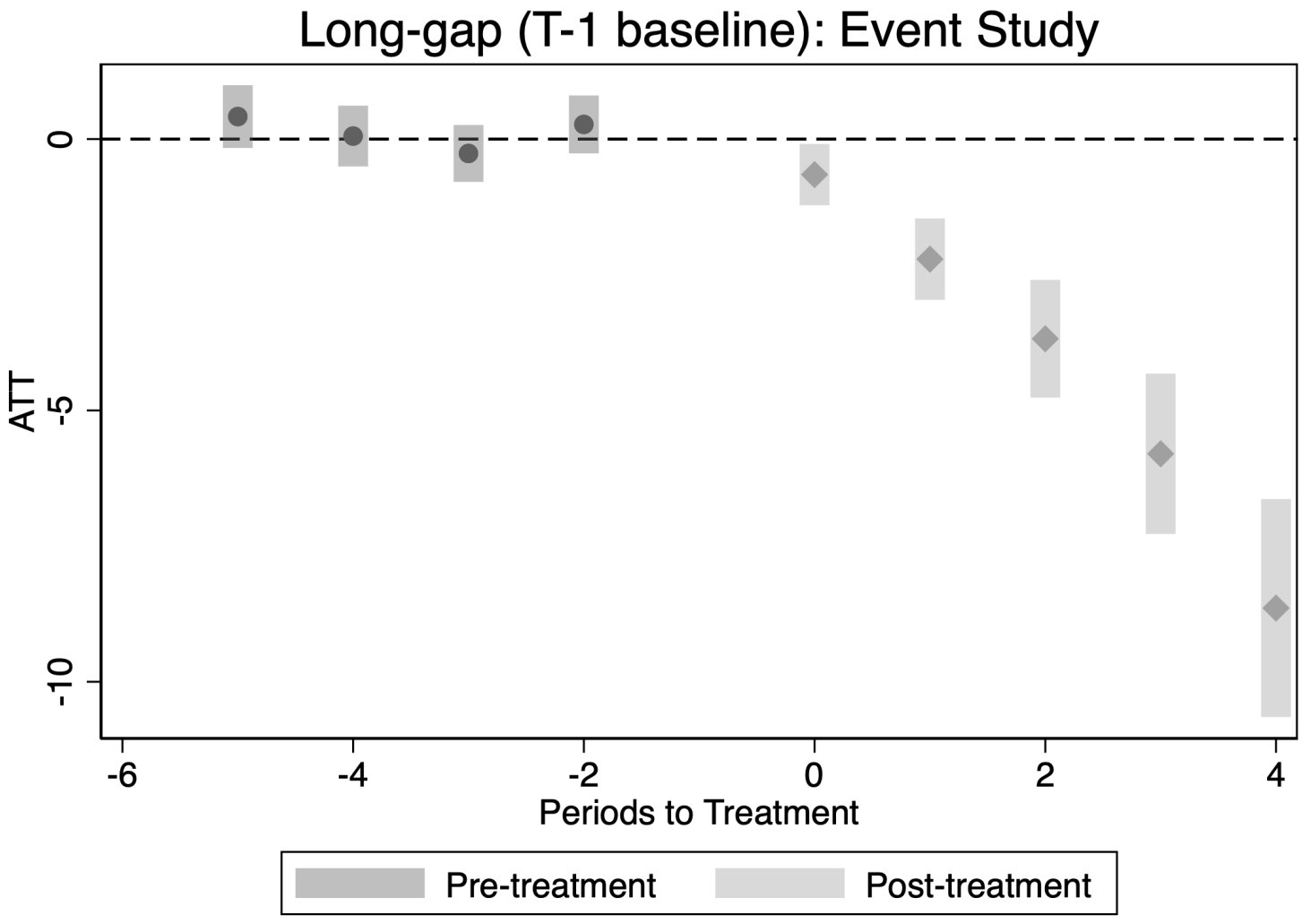
Making event study

- 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)
- When you drop $t - \tau$ as the baseline, then all estimates become ”long differences” compared to that point

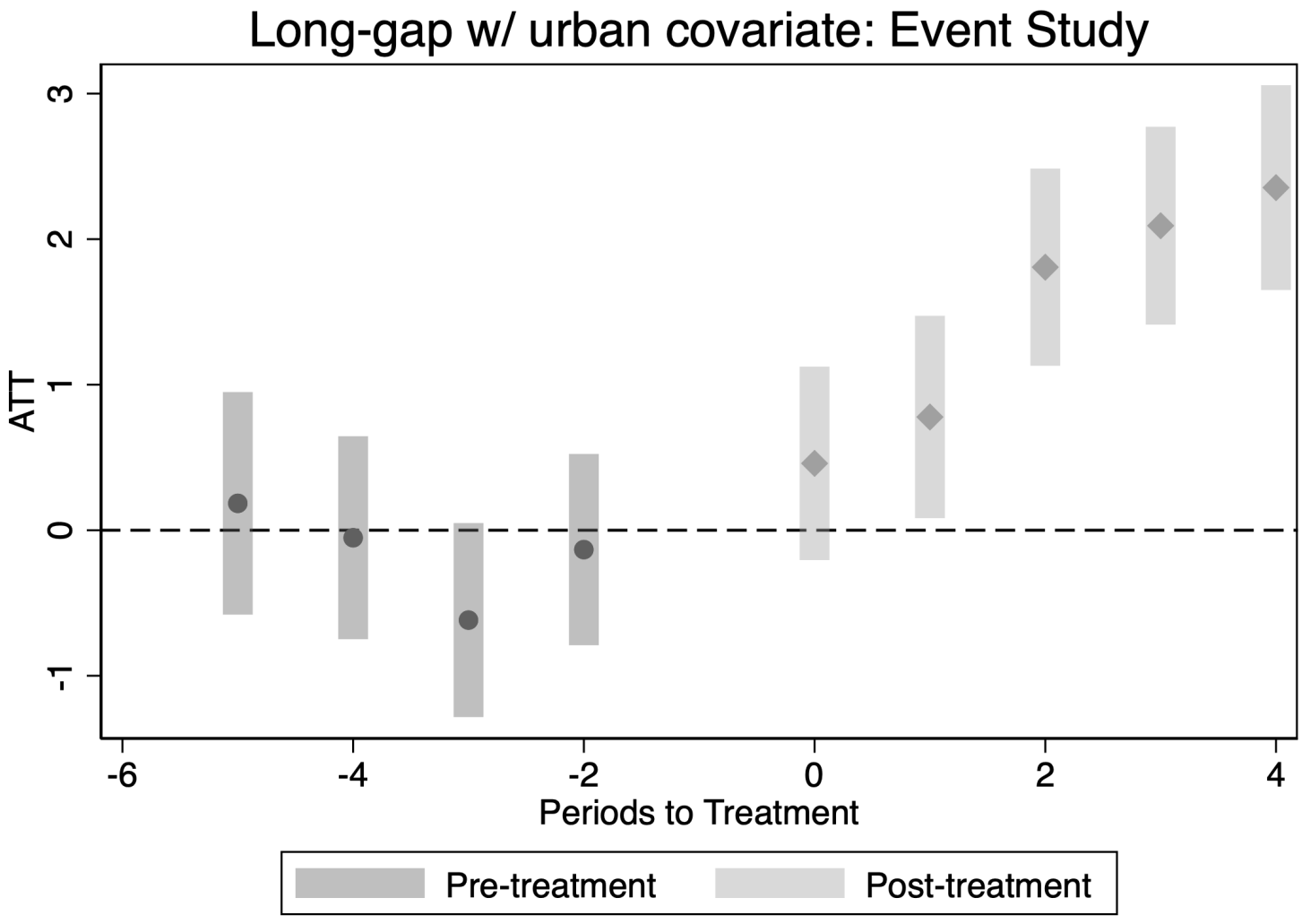
Event studies can mislead

- People mistakenly equate parallel trends with parallel pretrends, but you can have one without the other, both or neither
- One of the ways you can have parallel pre-trends but violate parallel trends is if the trends change differentially by covariate over time (called x-specific trends)
- Code is called `misleading_eventstudy.do` and `misleading_eventstudy.R`
- Note how well the unconditional parallel trends assumption appears to be when reviewing the graphs only

Event studies can mislead



Event studies can mislead



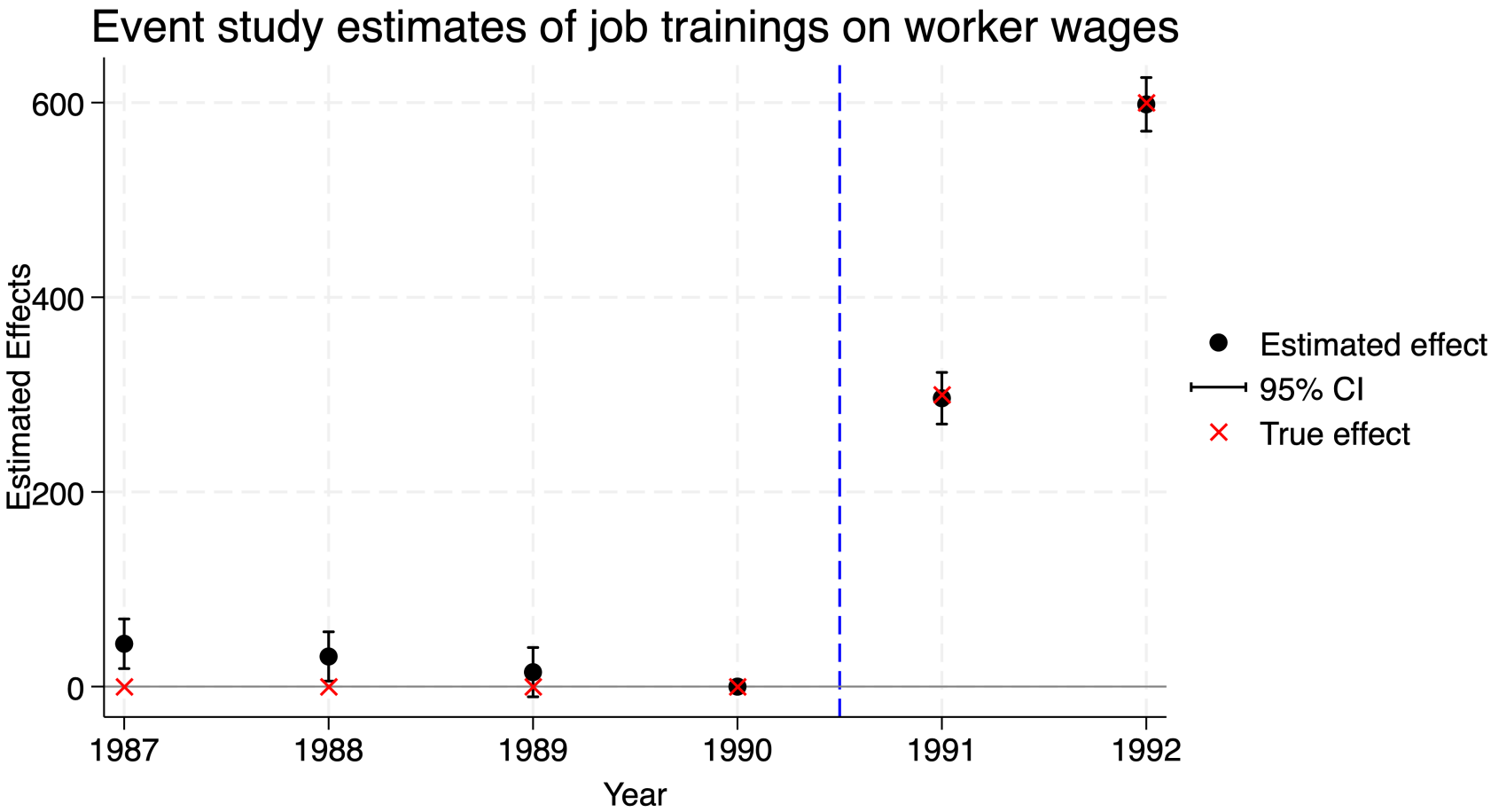
Time varying trends in Y^0 have *nothing* to do with parallel trends

```
* Step 2: Generate potential outcomes (Revised to shrink pre-trends)

* Smaller urban decline, nearly flat rural increase
bys county_id: gen trend = .
replace trend = -0.15*urban + 0.1*(1-urban) if year==1
replace trend = -0.25*urban + 0.1*(1-urban) if year==2
replace trend = -0.15*urban + 0.1*(1-urban) if year==3
replace trend = -0.1*urban + 0.1*(1-urban) if year==4
replace trend = -0.1*urban + 0.1*(1-urban) if year==5
replace trend = -0.5*urban + 0.1*(1-urban) if year==6
replace trend = -1*urban + 0.1*(1-urban) if year==7
replace trend = -1.5*urban + 0.1*(1-urban) if year==8
replace trend = -2*urban + 0.1*(1-urban) if year==9
replace trend = -2.5*urban + 0.1*(1-urban) if year==10

* Generate county-level fixed effects (increasing cross-sectional variance)
bys county_id: gen county_fe = rnormal(0, 2)

* Declare panel before generating serial correlation
xtset county_id year

* Generate serially correlated errors within counties
gen u = .
bys county_id: replace u = rnormal(0, 1) if year==1
bys county_id: replace u = 0.7*u[_n-1] + rnormal(0, 1) if year>1

* Potential outcome y0 now includes county FE and serially correlated errors
bys county_id: gen y0 = 15 + county_fe + trend*(year) + u

* Clearly positive treatment effect
gen y1 = y0 + post*treated*(year - 5)*0.6

* Observed outcome
gen birth_rate = treated*y1 + (1-treated)*y0
```

Selection can also make event studies misleading

- Remember from earlier: some forms of selection that satisfy parallel trends will nonetheless complicate event study plots
- Selection on baseline Y^0, for instance, can cause the pre-trends to break but not parallel trends

Selection on Y^0 and event studies



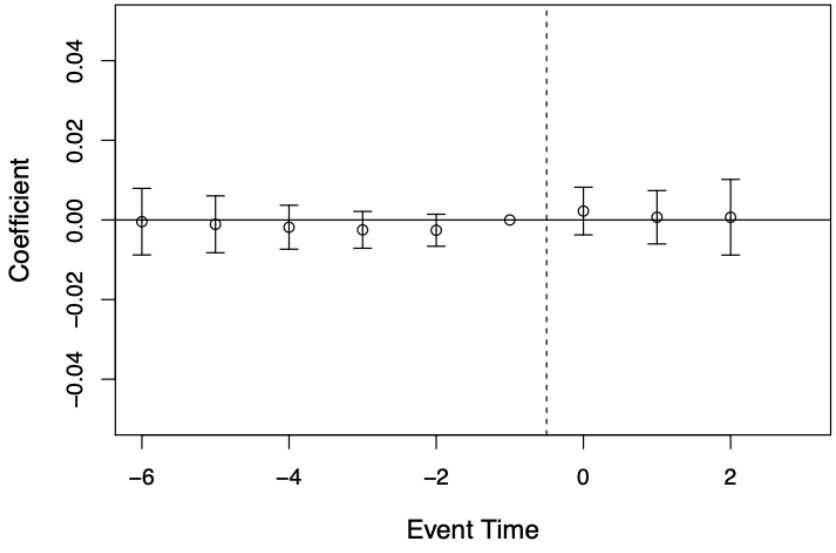
Worker fixed effects included. X marks true effects. 1990 is baseline year. Selection on baseline potential outcome.

Proof vs Evidence

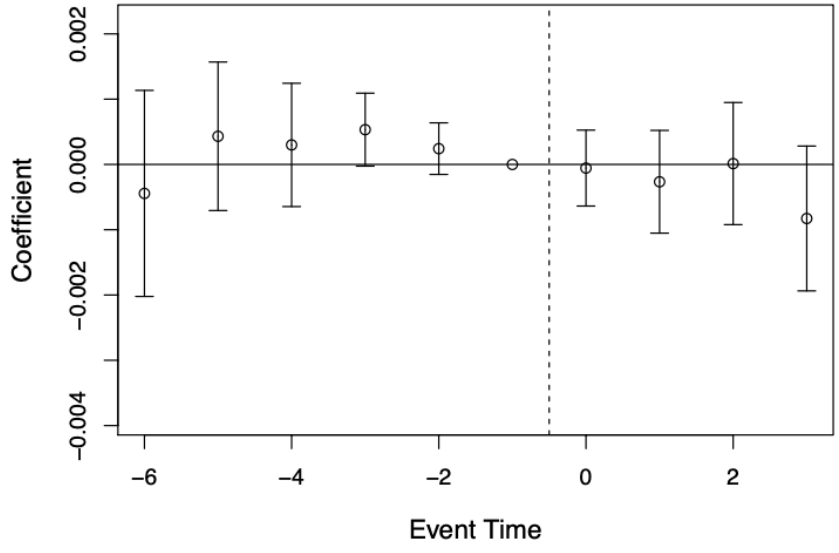
- Think of event studies as *evidence*, not *proof*
  - → Prosecutor based on police produces fingerprints, eyewitnesses and imprints in mud
- Event studies are evidence of pre-trends only, but they are consistent with various forms (but not all) forms of selection
- Falsifications are post-treatment heuristics based on logic and counter-evidence (Popperian reasoning)

Falsification 1: same outcome, but untreated groups

Age 65+ in 2014



(c) Medicaid Coverage



(d) Annual Mortality

Falsification 2: different outcomes, same treatment group

- Good falsifications make logical sense:
  - → They are a similar enough group (near elderly versus elderly)
  - → They are an outcome that would be susceptible to some omitted variable you're worried about
- Two examples
  - → Cheng and Hoekstra (2013) examine the effect of a gun laws on non-gun related offenses like grand theft auto and find no evidence of an effect
  - → Cunningham, DeAngelo and Tripp (2024) examine effect of a sex worker platform on irrelevant crime categories

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
- Maybe tall people have tall friends because basketball players are friends with one another?

Triple differences is a research design

- Many people equate triple differences with falsification exercise, but actually it isn't that – it is it's own design
- You use triple differences when you have a parallel trends violation – that is, when your diff-if-diff is biased
- Triple differences may sound like a falsification, but it isn't – it's a research design you use when parallel trends is violated
- Miller, Johnson and Wherry (2021) didn't use triple differences with near elderly because they didn't think they had a parallel trends violation – they used a falsification

Biased diff-in-diff #1

Table: Biased diff-in-diff #1: comparing states

States	**Period**	**Outcomes**	D_1	D_2
Experimental states	Before	$Y = NJ$		
	After	$Y = NJ + NJ_t + D$	$NJ_t + D$	
				$D + (NJ_t - PA_t)$
Non-experimental states	Before	$Y = PA$		
	After	$Y = PA + PA_t$	PA_t	

$$\widehat{\delta}_{did}^{true} = D + (NJ_t - PA_t)$$

The ATT is D. Assume, though, that parallel trends does not hold, $(NJ_t \neq PA_t)$

Biased Placebo diff-in-diff

Table: Biased placebo diff-in-diff: comparing states but single men and older women

States	**Period**	**Outcomes**	D_1	D_2
Experimental states	Before	$Y = NJ$		
	After	$Y = NJ + NJ_t$	NJ_t	
				$(NJ_t - PA_t)$
Non-experimental states	Before	$Y = PA$		
	After	$Y = PA + PA_t$	PA_t	

$$\widehat{\delta}_{did}^{placebo} = (NJ_t - PA_t)$$

Assume that parallel trends does not hold, $(NJ_t \neq PA_t)$

Two biased diff-in-diffs

- Parallel trends does not hold, ($NJ_t \neq PA_t$), but what if that's the same bias in our placebo DiD?
- Then we can subtract the second from the first:

$$\widehat{\delta}_{ddd} = \widehat{\delta}_{did}^{true} - \widehat{\delta}_{did}^{placebo}$$

- Triple differences is a "real design" with one parallel trends assumption:

$$(NJ_t{}^{true} - PA_t^{true}) = (NJ_t{}^{placebo} - PA_t^{placebo})$$

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)**		

Triple differences commentary

- Some people think that it requires that the placebo DiD be zero, but that's incorrect
- In Gruber's 1995 article, it isn't clear why he needed triple differences in the first place – his triple differences yielded -0.054 which is almost the same as what he found with his first diff-in-diff (-0.062)
- The main value of triple differences is that you use it when you believe the parallel trends assumption doesn't hold

Table: Difference-in-Difference-in-Differences (Gruber version)

Groups	States	Period	Outcomes	D_1	D_2	D_3
Married women 20-40	Experimental states	After	$NJ + MW + NJ_t + MW_t + D$	$NJ_t + MW_t + D$	$D + NJ_t - PA_t$	
		Before	$NJ + MW$			
	Non-experimental states	After	$PA + MW + PA_t + MW_t$	$PA_t + MW_t$		
		Before	$PA + MW$			
Single men Older women	Experimental states	After	$NJ + SO + NJ_t + SO_t$	$NJ_t + SO_t$	$NJ_t - PA_t$	D
		Before	$NJ + SO$			
	Non-experimental states	After	$PA + SO + PA_t + SO_t$	$PA_t + SO_t$		
		Before	$PA + SO$			

Triple diff assumption

$$\widehat{\delta}_{DDD} = D + \underbrace{[(NJ_t^{MW} - PA_t^{MW}) - (NJ_t^{SO} - PA_t^{SO})]}_{\text{Equally biased DiD \#1 and \# 2}}$$

Triple differences requires two diff-in-diff, from different groups, with the same bias. Parallel bias

DDD in Regression

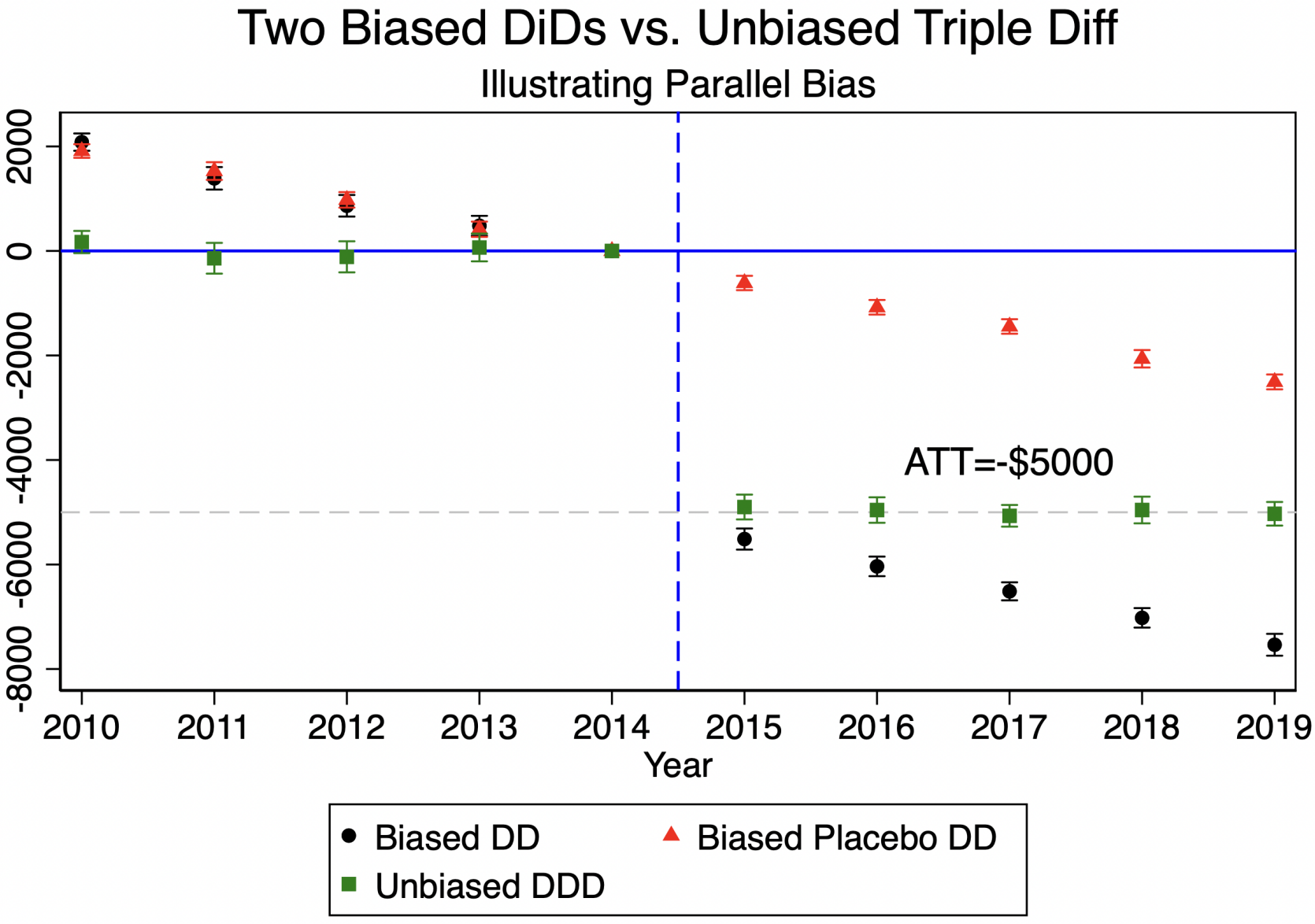
$$\begin{aligned} Y_{ijt} &= \alpha + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ &\quad + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt} \end{aligned}$$

- Your dataset will be stacked by group j and state i
- $\widehat{\beta_8}$ estimates the ATT
- Parallel bias, NA and SUTVA necessary and sufficient for identification

Simulation

In /Labs/DDD I have a simulation to illustrate this for us called ddd2.do. The ATT is -$5,000 but the biased DiD is -$7487. The non-parallel trends bias is -$2,487. So I replicate Gruber (with simulated data) where the placebo DiD is close (-$2,507). I then present a triple differences which gives us -$4,972. Let's look at the final product.

Triple differences event study



Great new paper to learn more

The Econometrics Journal

Econometrics Journal (2022), volume **00**, pp. 1–23.
https://doi.org/10.1093/ectj/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'.

Summarizing DDD

- Used to be people thought DDD required two parallel trends assumptions but it does not – it is a real design and requires one parallel trends assumption
- Parallel trends assumption is “parallel bias” – that the bias of the true DiD is the same as the bias of the placebo DiD
- The ladder of evidence still holds – you’ll want to present the event study plot, and my code provides it for you, because you need to evaluate the parallel bias assumption
- Given the lack of triple diff literacy, you may have to write this anticipating reader and maybe editor confusion and so ”educate” as you go – overlaying all three plots could be help

Roadmap

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

Introduction of Napster and spending on music

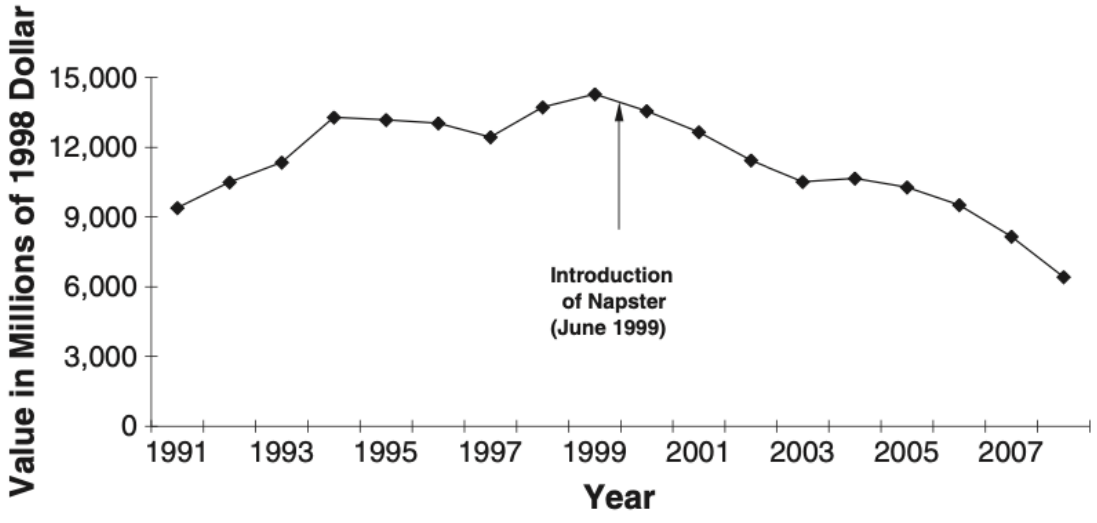
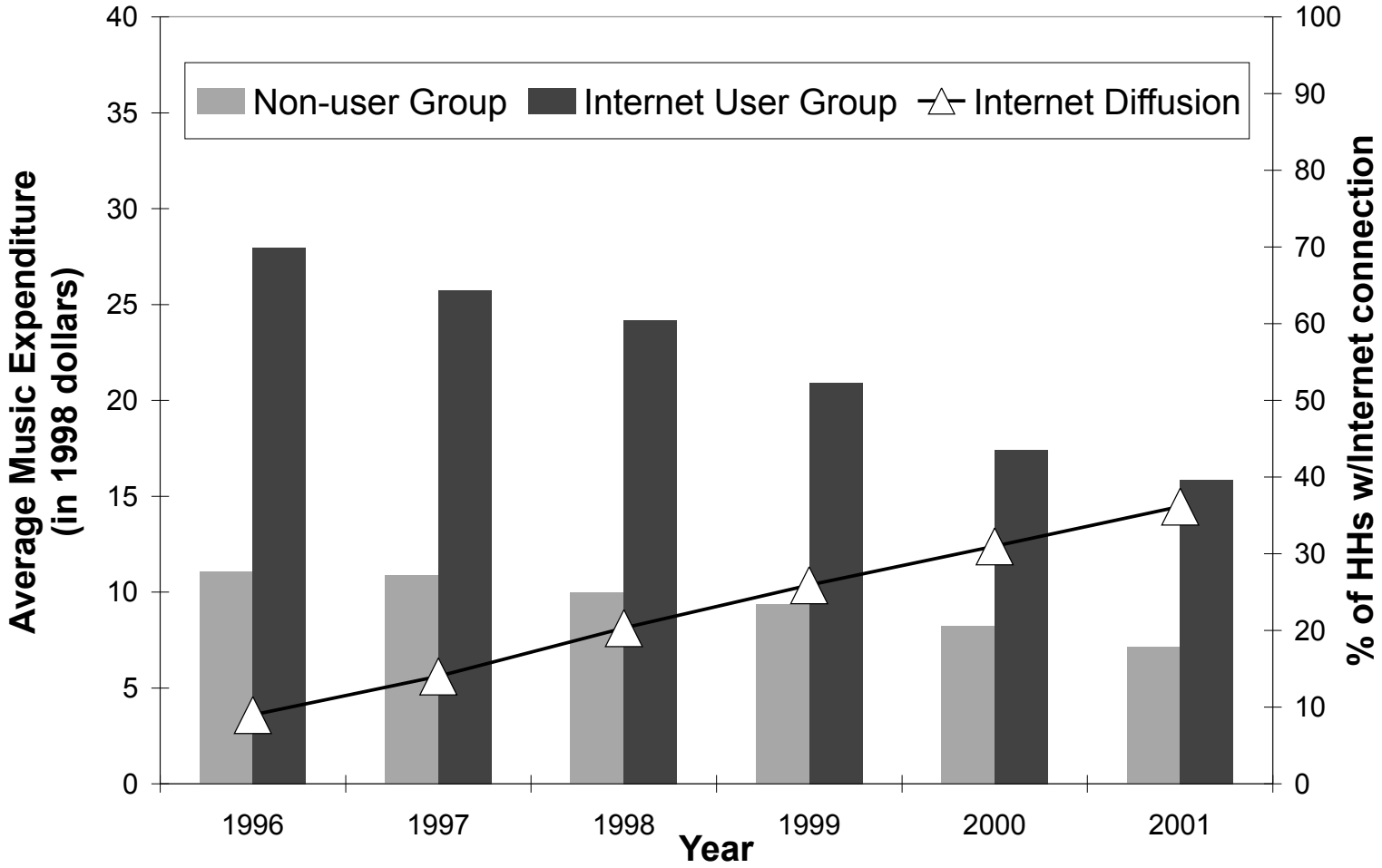


Figure 2. Total real value of record shipments in the USA. Refer to the RIAA's year-end statistics. Total sales include CDs, cassettes, LPs, and music videos. Starting from 2004, total sales also include digital formats such as legitimate download

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



Repeated Cross Section Risks

- Repeated cross sections have their own challenges that panels don't in that the group could be shifting compositionally
- Detect using a balance table with covariates highly predictive of the missing $E[Y^0|D=1]$ for this exercise
  - → Percent of cat owners is probably irrelevant to trends in potential outcomes
  - → But age and income is probably relevant for spending habits
  - → We'll discuss covariates more later, but for now just consider what characteristics are relevant to your outcome
- Documenting covariates that are cannot be affected by the treatment like this table is a way to check for *compositional changes* in the sample

Changes Between Internet and Non-Internet Users Over Time

Table 77: Changes between Internet and non-Internet users over time

Year	1997		1998		1999		2000	
	Internet user	Non-user	Internet user	Non-user	Internet user	Non-user	Internet user	Non-user
Demographics								
Age	40.2	49.0	42.3	49.0	44.1	49.4	44.3	49.9
Income	$52,887	$30,459	$51,995	$26,189	$49,970	$26,649	$47,510	$26,336
High school graduate	0.18	0.31	0.17	0.32	0.21	0.32	0.22	0.33
Some college	0.37	0.28	0.35	0.27	0.34	0.27	0.36	0.27
College grad	0.43	0.21	0.45	0.21	0.42	0.20	0.37	0.20
Manager	0.16	0.08	0.16	0.08	0.14	0.08	0.14	0.07

Sample means from the Consumer Expenditure Survey from Hong [2013].

What was this table about?

- Notice that users are getting older and poorer – both of which predict spending less money on music
- If these covariates are themselves predictive of trends, then it is suggestive parallel trends could be *mechanically* breaking, so estimate

$$X_{it} = \alpha + \beta_1 D_i + \beta_t Post_t + \delta(D_i \times Post_t) + \varepsilon_{it}$$

- Or consider the normalized difference in means table we discuss later
- If violated, then consider the following fix by Hong (2013) which adjusts for propensity scores in both periods

Step 1. Estimate Propensity Scores

- Estimate two separate propensity scores – one for each time period
- Pre-treatment period ($T = 0$) uses both treated and control units (as $Y = Y^0$ for both)

$$P_b = \Pr(D = 1 \mid T = 0, X)$$

- Post-treatment period ($T = 1$) using only control group units (as $Y = Y^0$ for control only)

$$P_a = \Pr(D = 1 \mid T = 1, X)$$

Step 2. Weight Observations Using Propensity Scores

- Use inverse probability weighting (IPW) to balance samples:
- For the pre-treatment period ($T = 0$):

$$E_w[Y \mid D = 1, T = 0] = \frac{\sum_{i \in T=0, D=1} Y_i / P_b(X_i)}{\sum_{i \in T=0, D=1} 1 / P_b(X_i)}$$

$$E_w[Y \mid D = 0, T = 0] = \frac{\sum_{i \in T=0, D=0} Y_i / (1 - P_b(X_i))}{\sum_{i \in T=0, D=0} 1 / (1 - P_b(X_i))}$$

- For the post-treatment period ($T = 1$):

$$E_w[Y \mid D = 1, T = 1] = \frac{\sum_{i \in T=1, D=1} Y_i / P_a(X_i)}{\sum_{i \in T=1, D=1} 1 / P_a(X_i)}$$

$$E_w[Y \mid D = 0, T = 1] = \frac{\sum_{i \in T=1, D=0} Y_i / (1 - P_a(X_i))}{\sum_{i \in T=1, D=0} 1 / (1 - P_a(X_i))}$$

Step 3. Calculate the Adjusted Difference-in-Differences

- Use the weighted averages to compute the adjusted DiD estimator:

$$\begin{aligned}\text{Adjusted DiD} = &(E_w[Y \mid D=1, T=1] - E_w[Y \mid D=0, T=1]) \\ &- (E_w[Y \mid D=1, T=0] - E_w[Y \mid D=0, T=0])\end{aligned}$$

- This accounts for compositional change and enhances credibility of ATT estimation in repeated cross-sections.

Reweighted regressions

- It's a reweighted 2x2 estimator which we know how as an OLS equivalence
- You're going to use as analytical weights the IPW formulas and one of the regressions with WLS
- But make sure you are using *both* IPW expressions – it is two propensity score formulas, not one

Value of the method

- Panel datasets are expensive because following the same people over time is expensive
  - → Tracking respondents, maintaining contact, managing attrition
  - → Many countries, particularly in the Global South, have fewer of them than in highly developed countries
- Repeated cross sections are often more common – rich, nationally representative datasets collected in waves, but with different individuals each time
- Hong's estimator was developed precisely for this context – an estimator that can account for the changing compositional changes (a problem that happens often, but is rarely addressed)
- Diagnostic tools are also very intuitive – balance tables, seemingly unrelated regressions on multiple covariates

Is Unconditional Parallel Trends Plausible?

- We have been working with a parallel trends assumption that requires the two *groups* have the same trend in average Y^0
- But what if these two groups were so different from one another that their potential outcomes didn't add up to be the same?
- What if the workers in the job training program were poor and the control group was rich – do we think the earnings of those two groups would have been evolved similarly? Why/why not?

You may need a *different* PT assumption

- Parallel trends for sub-populations, but not the whole population:
  1. Unconditional parallel trends: parallel trends holds for the overall average treatment and control groups
  2. Conditional parallel trends: parallel trends only has to hold for observable sub-groups, but not necessarily the whole group
- Example: Male versus female earnings growth
  - → Assume male earnings grow by $+2$, female by $+1$,
  - → Treatment is 75% male, control is 25% is male
  - → $E[\Delta Y^0|D=1] = 1.75$ but $E[\Delta Y^0|D=0] = 1.25$
- Unconditional parallel trends won't hold because the groups aren't balanced on the characteristics that cause trends

Conditional parallel trends

Conditional parallel trends (CPT) is a weakened version of the parallel trends assumption requiring that it holds *within* the dimensions of the selected covariates:

$$CPT: E[Y_k^0|Post, X] - E[Y_k^0|Pre, X] = E[Y_U^0|Post, X] - E[Y_U^0|Pre, X]$$

First time it shows up is Heckman, Ichimura and Todd (1997). Idea can be abstract so let's look at a graph to help us decipher what it means and what we can do. And it's still not directly testable due to still missing the first potential outcome (i.e., counterfactual).

Which covariates do you need?

- Conditional parallel trends is tricky because it first assumes you know the covariates you need to satisfy it
- But how do we decide which variables do this and which ones don't?
- Econometrics is no help here – you need common sense, theory, logic, and expertise
- When selecting covariates, use "confounders" logic – but confounders of the *trend*?

Graphs Can Help Pick Covariates

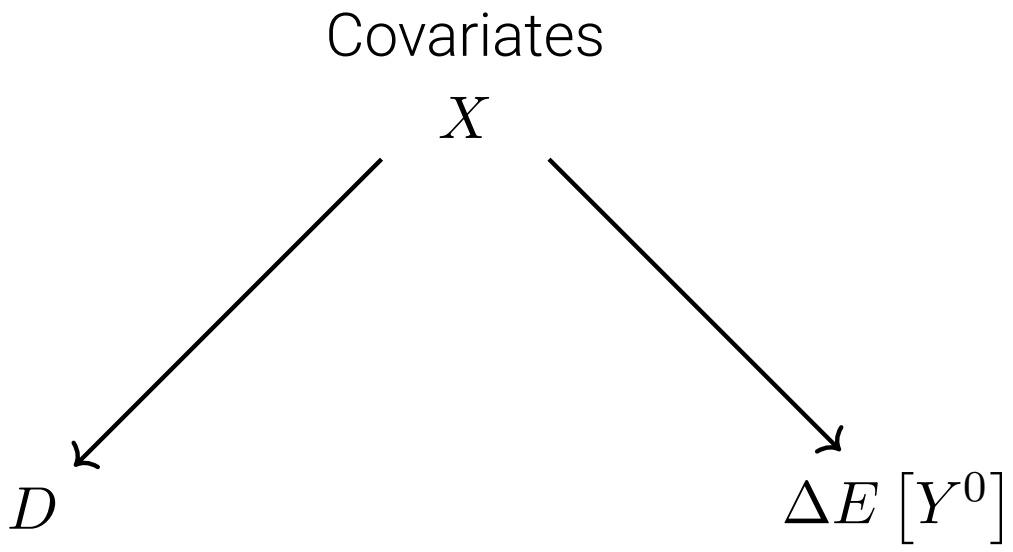


Figure: DAG representing differences in county-level covariate composition (X) across treatment and control groups (D) and their determination of the untreated potential outcome trends ($\Delta E\left[Y^0\right]$).

Example: Concealed Carry Gun Laws and Murder

- Become an expert on your left-hand-side variable because you need to know $X \rightarrow \Delta Y^0$
  - $\rightarrow$ Drugs and gangs (i.e., crack cocaine epidemic)
  - $\rightarrow$ Cities had different homicide rates than rural counties
  - $\rightarrow$ Police, incarceration
  - $\rightarrow$ Demographics (e.g., race shares, age shares) and economic things (e.g., poverty, per capital income, AFDC rolls)
  - $\rightarrow$ Consider LASSO in the pre-periods only to select covariates predictive of pre-treatment *untreated potential outcome* trends (i.e., ΔY^0)

Covariate Imbalance

- Kahn-Lang and Lang (2020) say that if your two groups are different at baseline on outcomes, treatment was not random, is likely caused by covariates, and you need to explain why that matters or does not
- Randomized treatment will balance covariates, hence why we don't need to include them as controls in RCT *even though they cause the outcome*
- But if they aren't balanced in your diff-in-diff, and they cause *trends*, then parallel trends won't hold
- So, next you focus on $X \rightarrow D$, which is to say, check the imbalance in covariates you need for conditional parallel trends

Create a balance table

Table 4: Covariate Balance Statistics

	Unweighted			Weighted		
Variable	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	49.43	49.33	-0.03	50.48	50.07	-0.24
% White	81.64	90.48	0.59	77.91	79.54	0.11
% Hispanic	9.64	8.23	-0.10	17.01	18.86	0.11
Unemployment Rate	7.61	8.01	0.16	7.00	8.01	0.50
Poverty Rate	19.28	16.53	-0.42	17.24	15.29	-0.37
Median Income	43.04	47.97	0.43	49.31	57.86	0.68
2014 - 2013 Covariate Differences						
% Female	-0.02	-0.02	0.00	0.02	0.01	-0.09
% White	-0.21	-0.21	0.01	-0.32	-0.33	-0.04
% Hispanic	0.20	0.21	0.04	0.25	0.33	0.29
Unemployment Rate	-1.16	-1.30	-0.21	-1.08	-1.36	-0.55
Poverty Rate	-0.55	-0.28	0.14	-0.41	-0.35	0.05
Median Income	0.98	1.11	0.06	1.10	1.74	0.32

Notes: 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. All variables are measured in percentage values, except for median household income, which is measured in thousands of U.S. dollars. 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 2×2 setting.

Baseline Covariates and Normalized Difference

- Baseline covariates are measured before treatment $(t-1)$.
- Report the averages of covariates for both groups in a table and the "normalized difference in means" 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 in absolute value to avoid problematic imbalance (Imbens and Rubin 2015).

Summarizing

- So, when choosing covariates, remember that there are two steps involved
  1. $X \rightarrow \Delta E[Y^0]$. Pick covariates that are needed to satisfying conditional parallel trends
  2. $X \rightarrow D$. Check for imbalance using the normalized difference in means equation to determine if you have "problematic imbalance"
- And the heuristic I am suggesting is to select covariates that are the "ordinary determinants of Y^0" (i.e., become an expert on your left-hand-side variable)

Estimators for Covariates

- But, let's say that we feel we have to assume conditional parallel trends
- What estimators can accomodate that assumption with the least amount of assumptions?
- We will review four now:
    1. Inverse probability weighting
    2. Outcome regression
    3. Double robust
    4. Regressions

Three key covariate estimators in diff-in-diff

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 Zhao (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

Inverse probability weighting

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

Four main assumptions

1. No anticipation
2. Conditional parallel trends

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

3. Common support

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

4. 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)

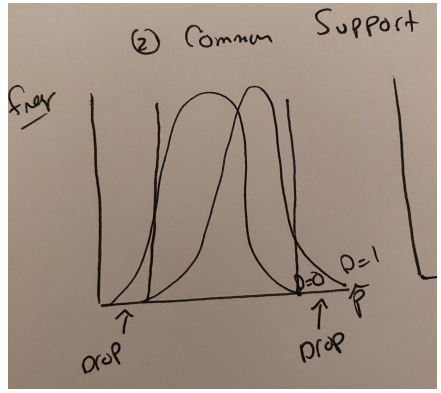
How do these work together?

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

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

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")

Outcome regression and double robust

- 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*

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

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

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

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 Zhao (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

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.

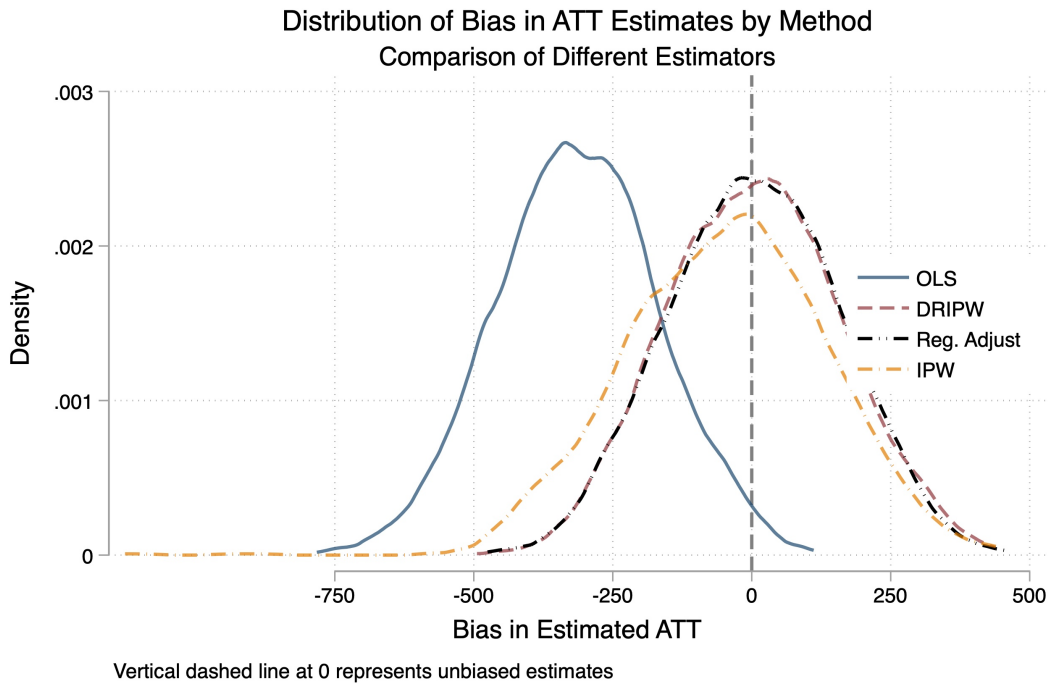
Why not both?

- 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 couple of Monte Carlos – one by Pedro Sant'Anna, and then one by me

Monte Carlo Simulations

- 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

Temporary page!

LaTeX was unable to guess the total number of pages correctly.
was some unprocessed data that should have been added to
page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus pag
away, because LaTeX now knows how many pages to expect fo
document.