Causal Inference II

Mixtape Session

Roadmap

- Estimation
  - OLS specification
  - Inference
- Parallel trends violations
  - How parallel trends can get violated
  - DiD in Court
  - Triple difference
  - Placebo outcomes
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD
  - Lalonde lab

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

Replying to @BCAppelbaum

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

3:49 PM · Mar 18, 2019

179 Reply Share

Read 18 replies

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/MOtbuRX4eyQ?t=1882

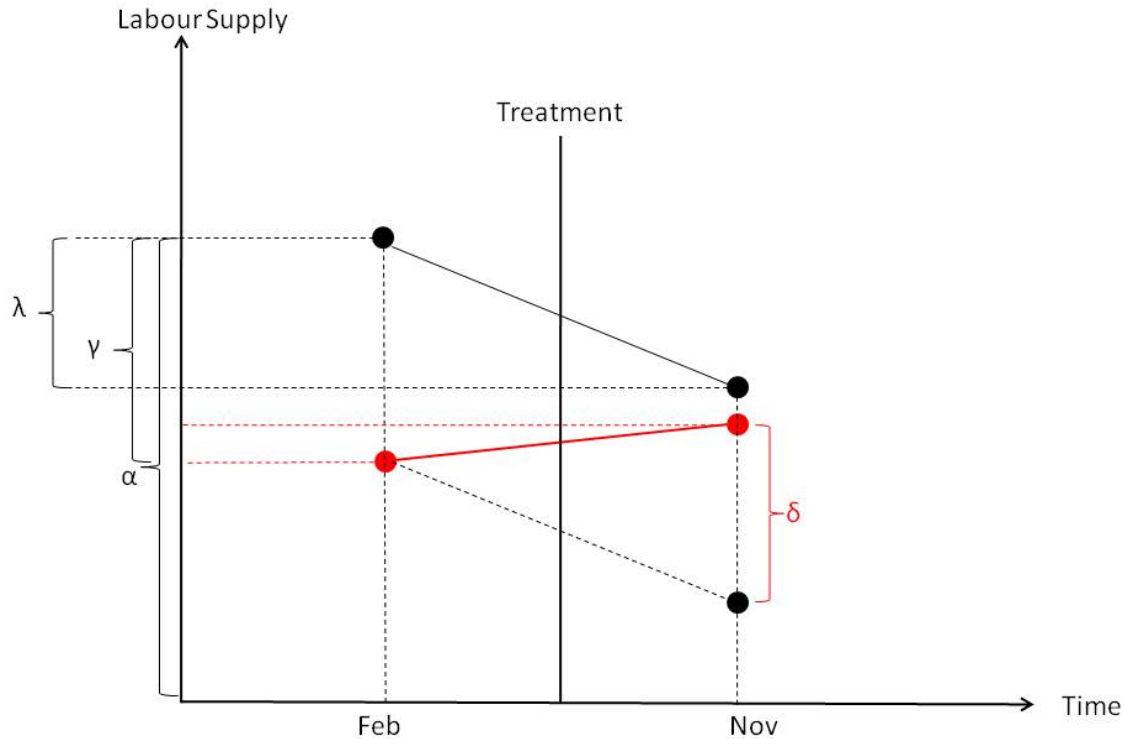
OLS specification of the DiD equation

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

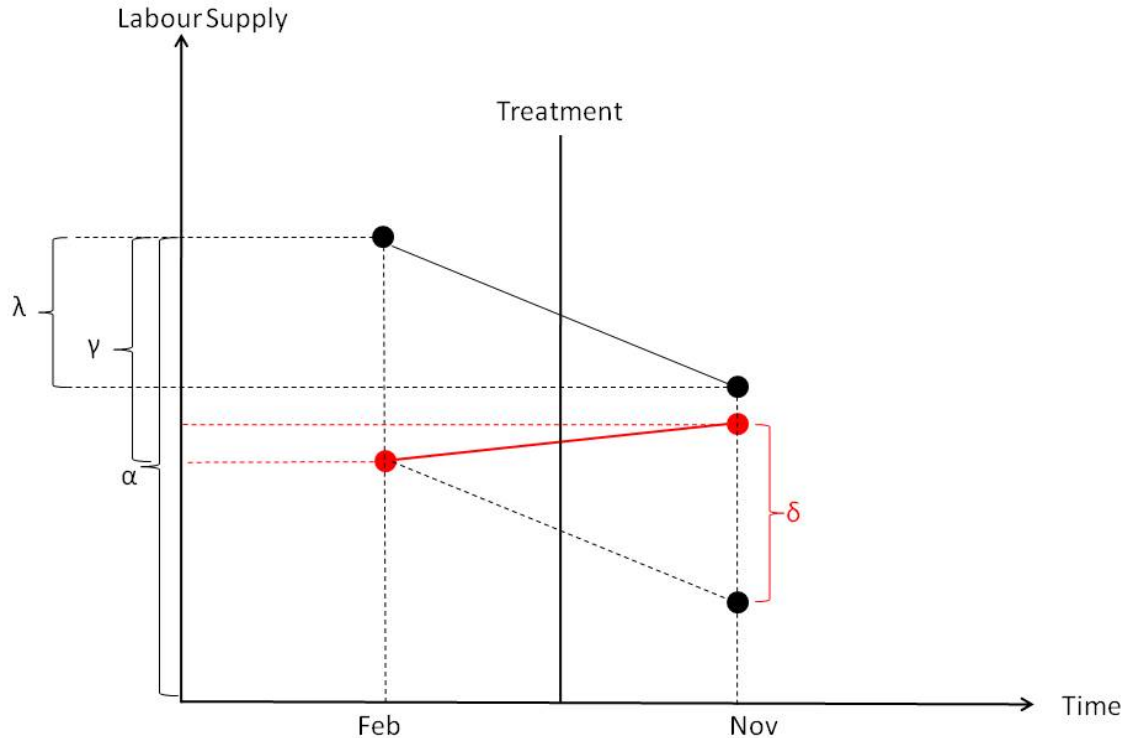
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

  - → 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 Post - NJ Pre) - (PA Post - PA Pre) $= \delta$

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



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



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

Inference

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely understate 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.

Roadmap

- Estimation
  - OLS specification
  - Inference
- Parallel trends violations
  - How parallel trends can get violated
  - DiD in Court
  - Triple difference
  - Placebo outcomes
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD
  - 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

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

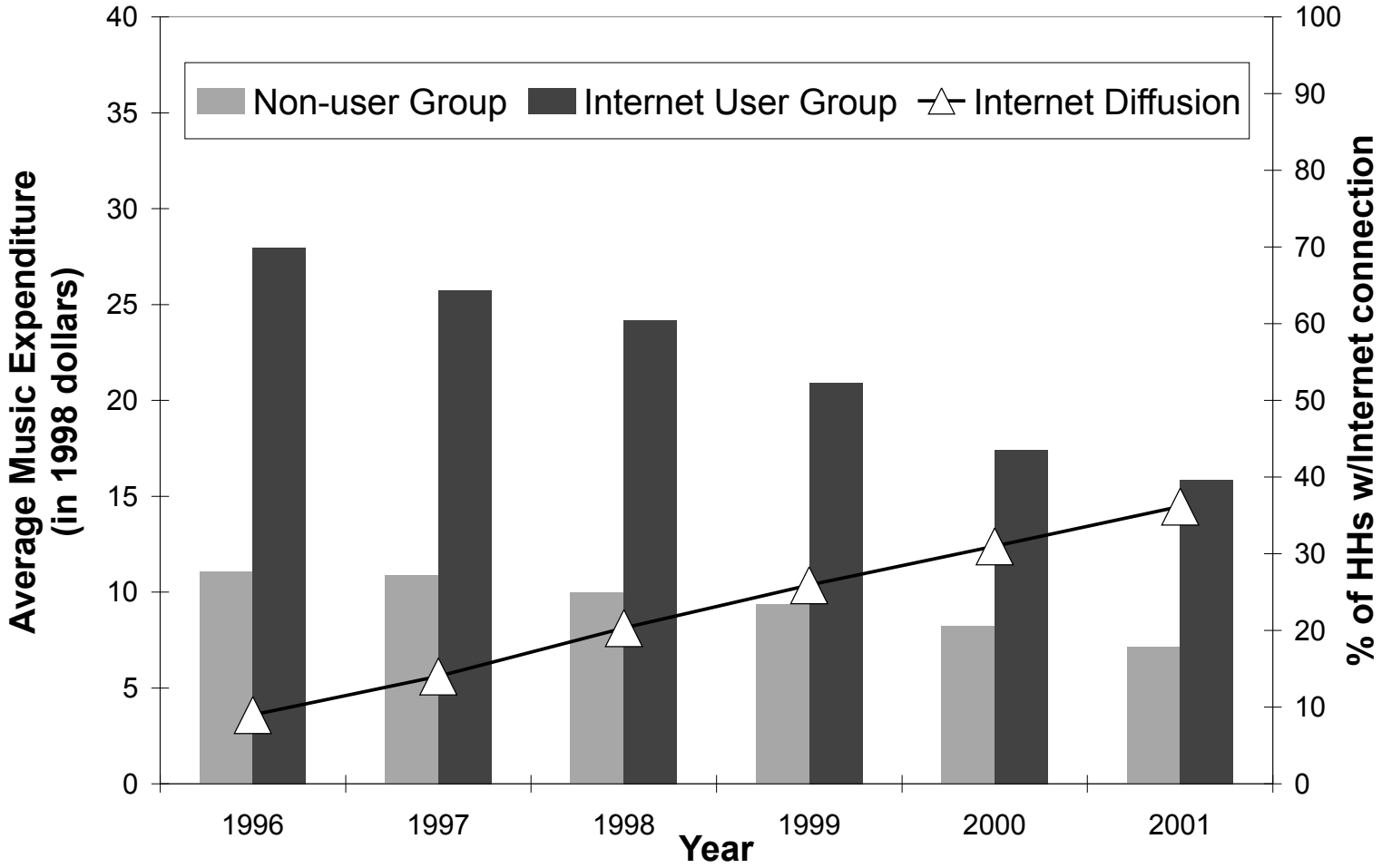


Table 1: Descriptive Statistics for Internet User and Non-user Groups[a]

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)

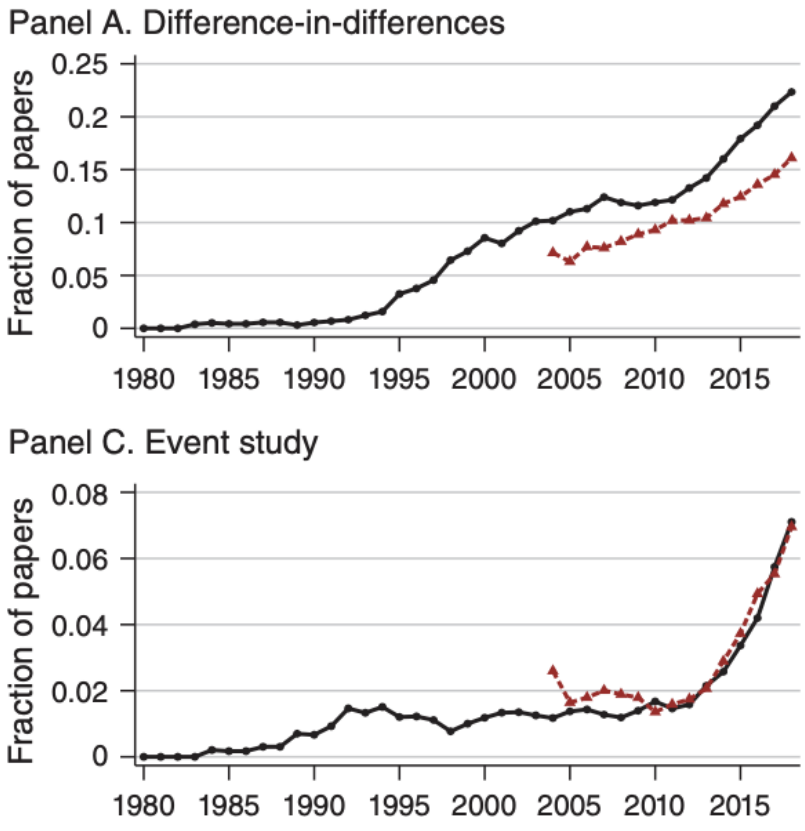
Repeated cross-sections

- 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

- 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 mixing tight logic based falsifications with particular kinds of data visualization, starting with the event study

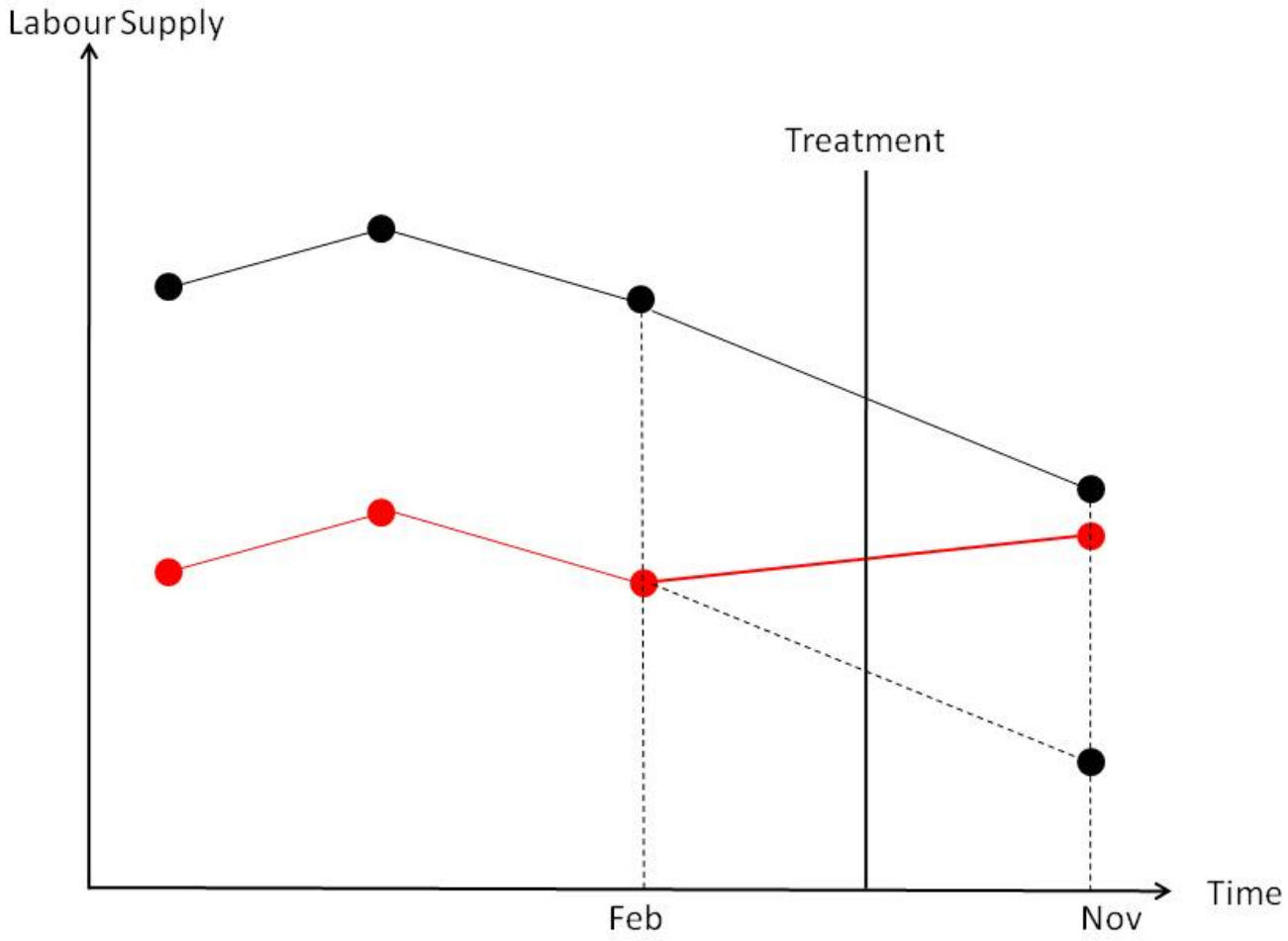
Event studies have become mandatory in DiD



Intuition behind event studies

- Princeton Industrial Relations Section seems to be behind this – this intense focus on research design but also verifying assumptions
- The identifying assumption for all DD designs is parallel trends , but since we cannot verify parallel trends, we often look at pre-trends
- It's a type of check for selection bias, but you must understand what it is and what it isn't to see its value but not be naive about it (it is not a silver bullet)
- 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)

Plot the raw data when there's only two groups



Evidence for parallel trends: pre-trends

Let's do the bonus questions on first and second tab now

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

Beware of naive reasoning

- Parallel pre-trends $\neq$ parallel trends – these are often thought to be the same thing, and they aren't
- Equating them is a kind of *post hoc ergo propter hoc* fallacy
- Parallel pre-trends is more like a smoking gun based on things "looking the same" before
- Similar to checking for covariate balance in RCTs
- But **cannot** substitute for domain knowledge!

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 group s with the calendar year τ
  - → Treatment occurs in year 0, no anticipation, drop baseline $t-1$
  - → 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 $\widehat{\mu}$ 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, $\widehat{\delta}$ 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

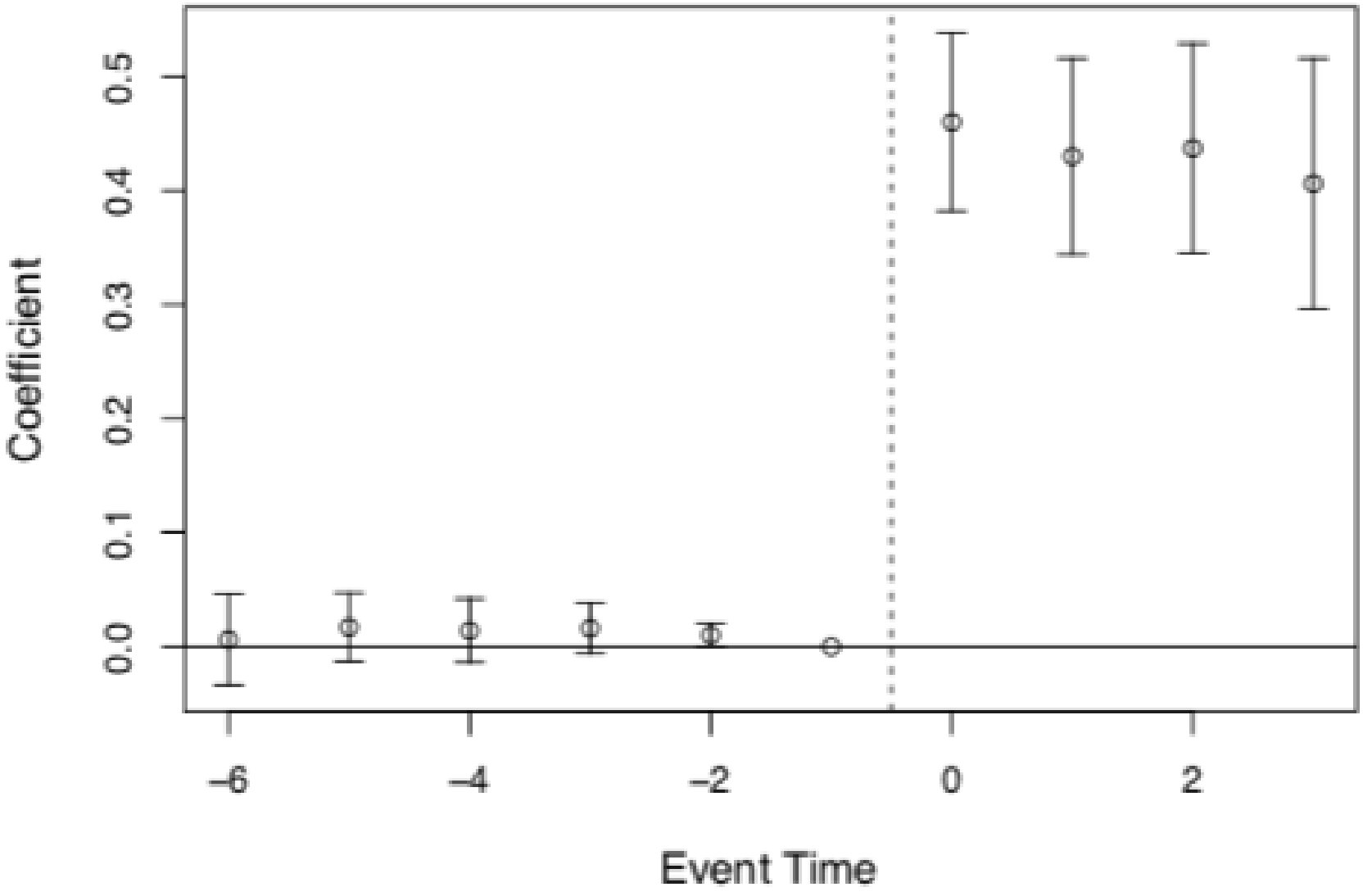
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults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

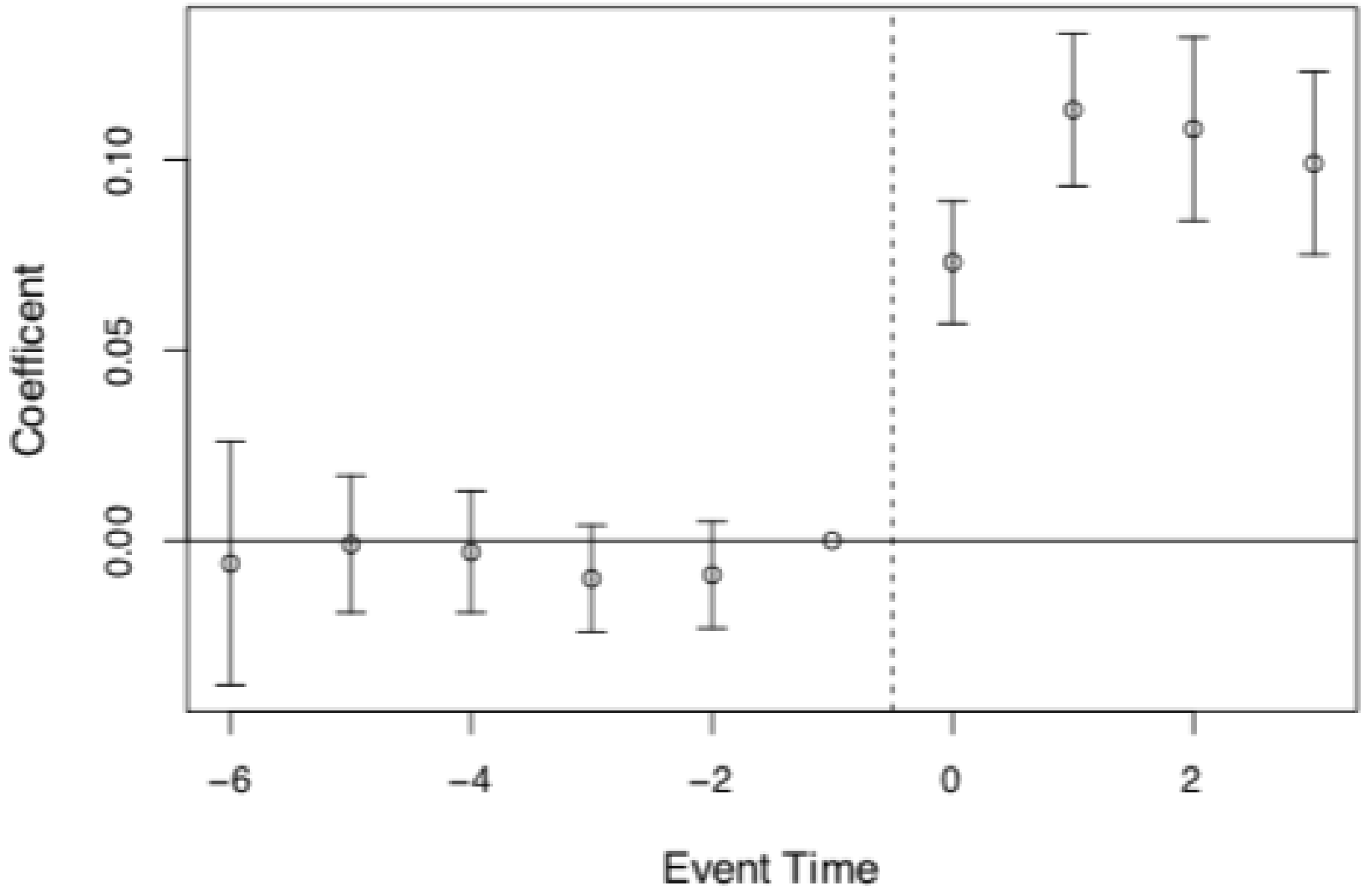
Issue Section: Article

Types of evidence

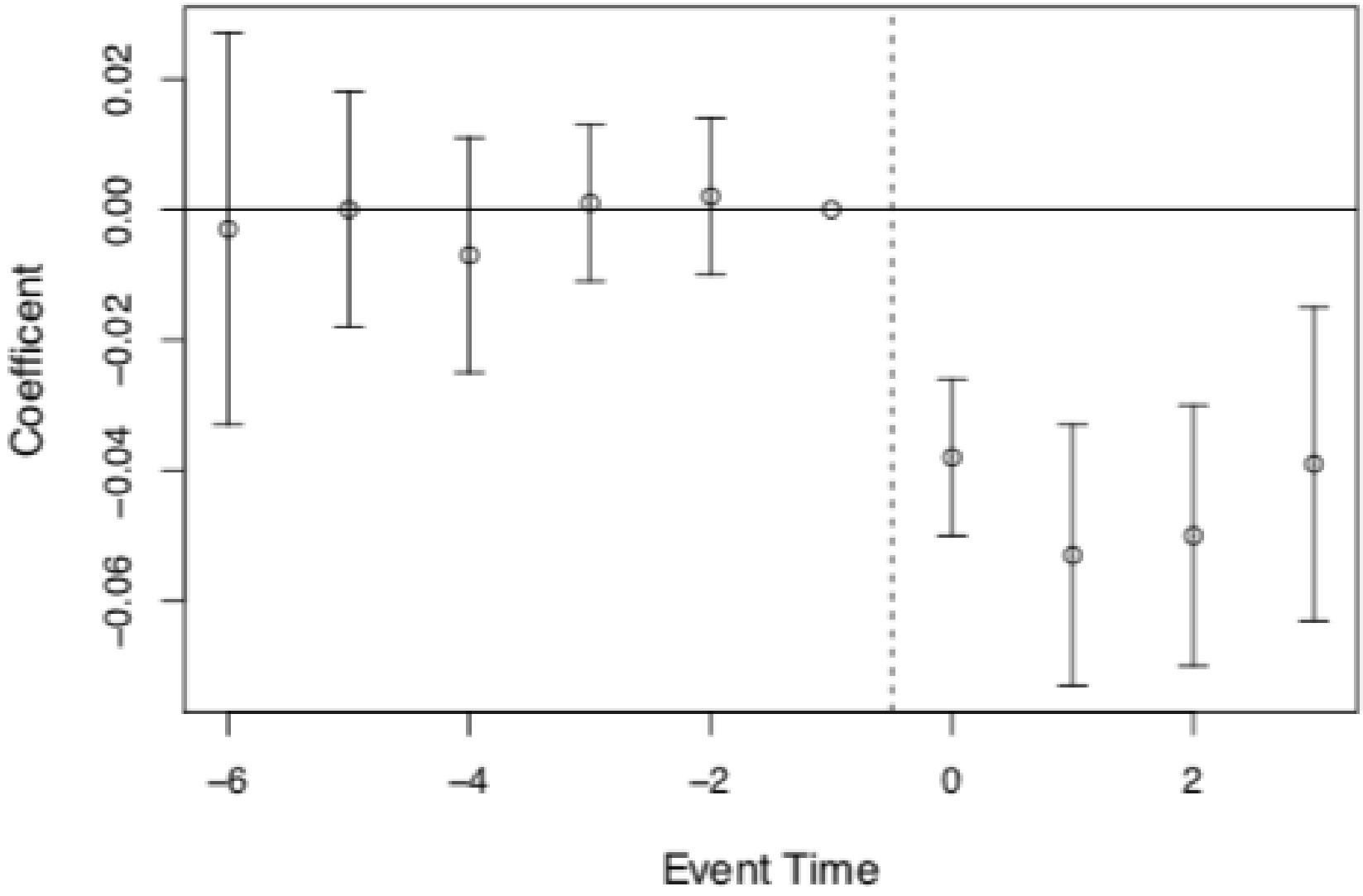
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Event study** – Show leads and lags on mortality



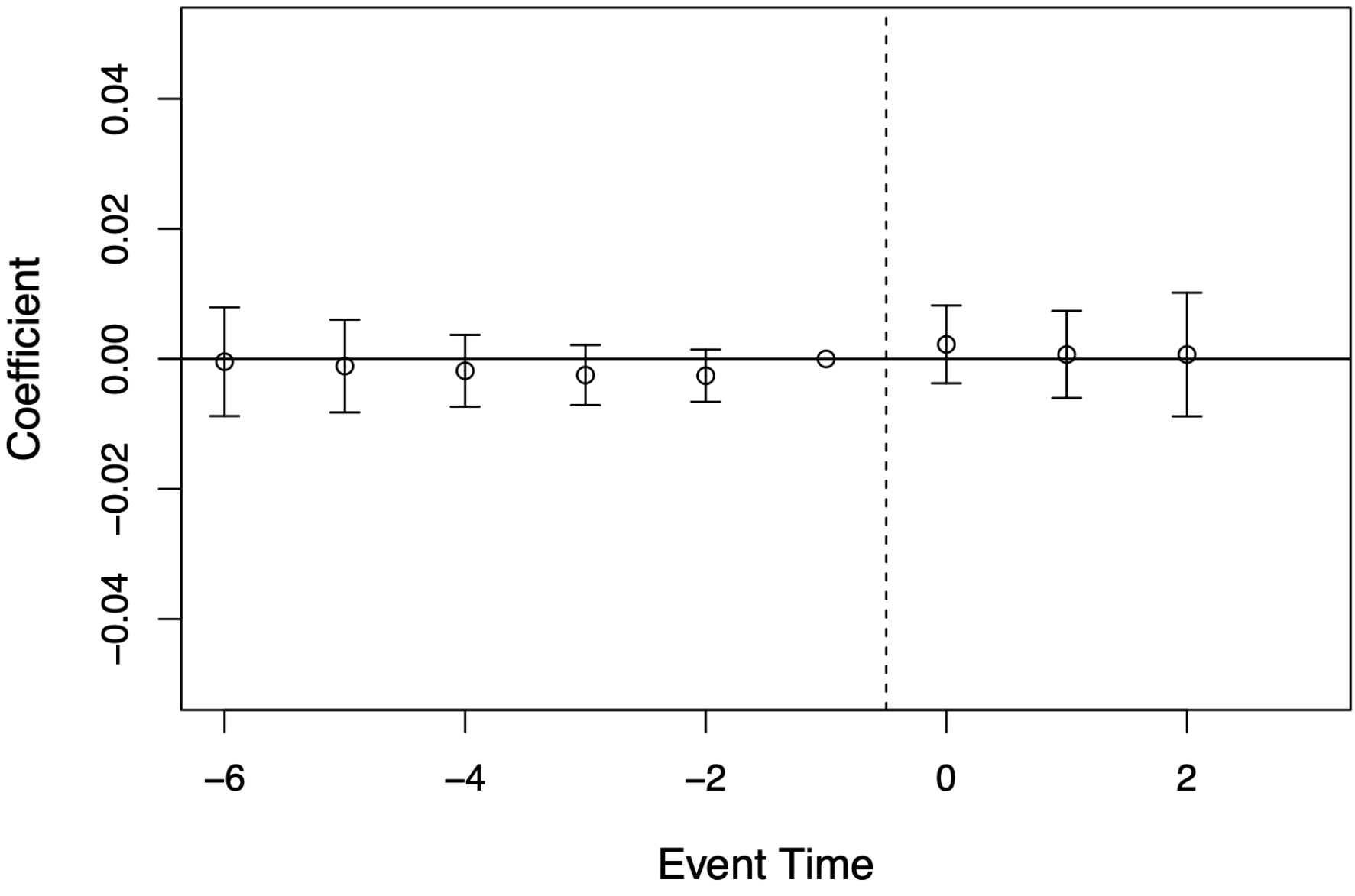
(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured


Coefficient
0.04
0.02
0.00
−0.02
−0.04
−6
−4
−2
0
2
Event Time

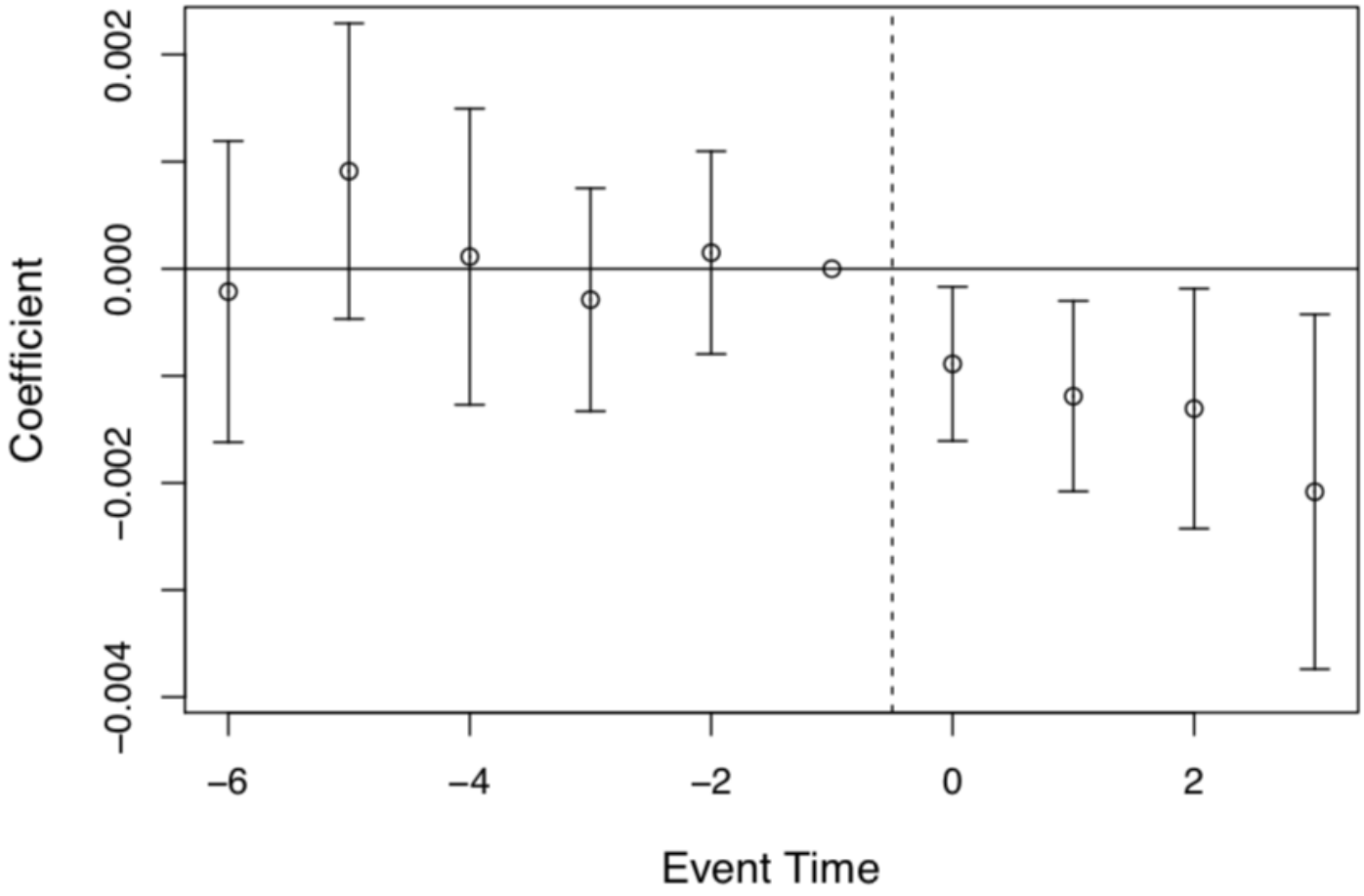


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

Additional identification things

- 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 uses a within-state untreated group; little trickier, so let's use the table again

Table: Difference-in-Difference-in-differences

States	Group	Period	Outcomes	D_1	D_2	D_3
NJ	Low wage employment	After	$NJ + T + NJ_t + l_t + D$	$T + NJ_t + l_t + D$	$D + l_t - s_t$	
		Before	NJ			
	High wage employment	After	$NJ + T + NJ_t + s_t$	$T + NJ_t + s_t$		
		Before	NJ			
PA	Low wage employment	After	$PA + T + PA_t + l_t$	$T + PA_t + l_t$	$l_t - s_t$	D
		Before	PA			
	High wage employment	After	$PA + T + PA_t + s_t$	$T + PA_t + s_t$		
		Before	PA			

What is our identifying assumption?

$l_t - s_t$ is the same for PA and NJ in the post period. That's the gap between high and low wage employment in PA (observed) is the same as NJ (counterfactual).

DDD Example by Gruber

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

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

DDD in Regression

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

- Your panel is now a group j state i (e.g., AR high wage worker 1991, AR high wage worker 1992, etc.)
- Assume we drop τ_t but I just want to show it to you for now.
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias

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'.

Falsification test with alternative outcome

- The within-group control group (DDD) is a form of placebo analysis using the same *outcome*
- But there are also placebos using a *different* outcome – but you need a hypothesis of mechanisms to figure out what is in fact a *different outcome*
- Figure out what those are, and test them – finding no effect 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

- Estimation
  - OLS specification
  - Inference
- Parallel trends violations
  - How parallel trends can get violated
  - DiD in Court
  - Triple difference
  - Placebo outcomes
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD
  - Lalonde lab

Controls

- Controls can address omitted variable bias (backdoor criterion), and they can improve precision
- OLS can accommodate controls, and so we tend to include them so long as they are time varying
- But unfortunately, time varying covariates can create problems, especially if the treatment causes the covariates (bad controls, colliders)

Inverse probability weighting DiD

Abadie (2005) incorporates baseline covariates into the propensity score which are then used as weights to estimate the ATT in a simple 3-step process

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 IPW

Terms

- t is year of treatment which doesn't vary across units (so no differential timing)
- Y^1 and Y^0 are potential outcomes (counterfactual versus actual)
- D is 1 or 0 based on group and time
- X_b are "baseline" covariates **only** – they do not vary over time, which means propensity scores are estimated off the b period **only**

Assumptions

Kind of common for this propensity score literature to only have two assumptions. But usually the first conditional independence. Now it is parallel trends because this is DD

1. Conditional parallel trends

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

   (Notice the b subscript. What is that you think?)

2. Common support

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

   Let's see a picture of common support that I drew. Apologies it's horrible

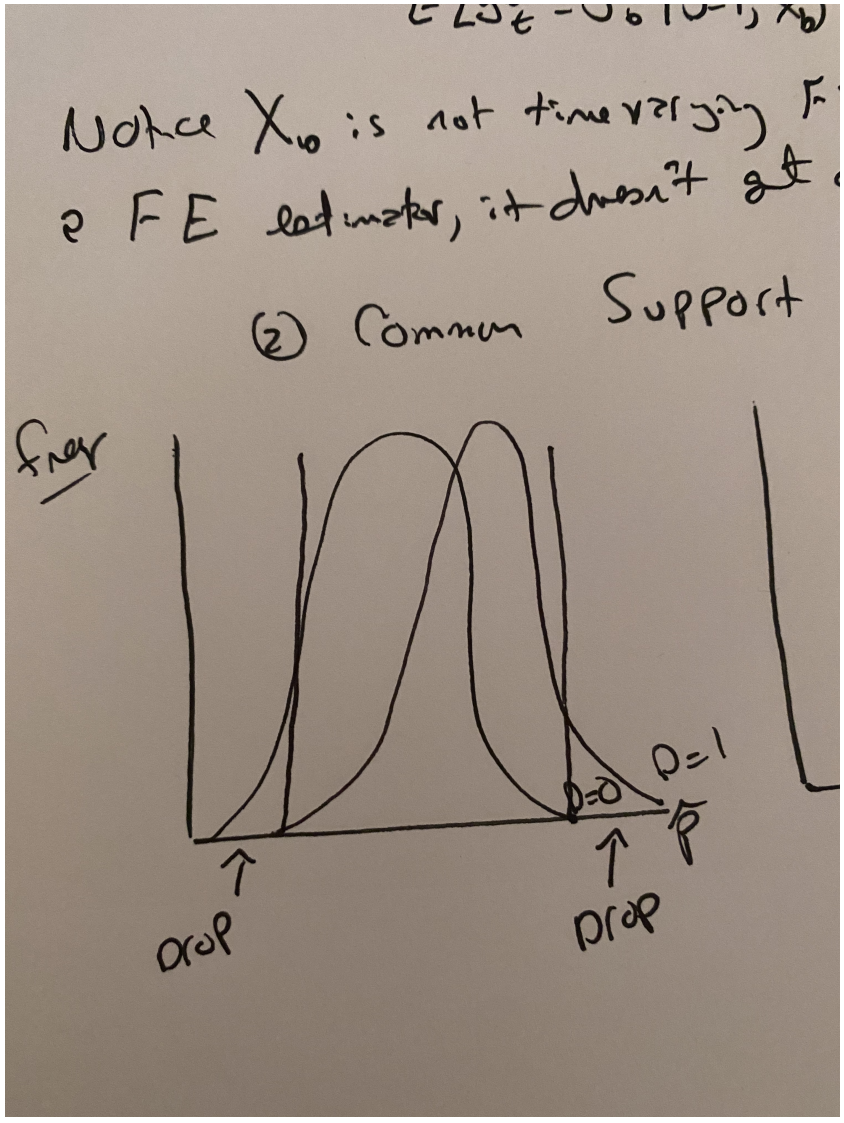
Common support

As we are identifying the ATT, we only need common support with respect to treated units

Your identify assumptions are always with respect to the missing covariates in other words and for the ATT, you are missing Y^0 for the treatment group

If we were estimating ATU, we'd be missing Y^1 for controls and need common support (Y in treatment for all ranges of control), and for ATE we'd need both

Visualizing propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

$$ATT = E[Y_t^1 - Y_t^0 | D_t = 1] \tag{1}$$

Abadie's estimator

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right] \tag{2}$$

Propensity scores

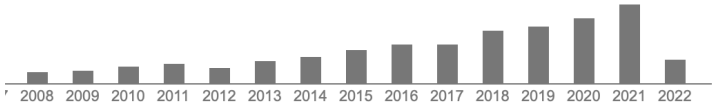
- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

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

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
- One of my favorite lesser known of the new DiD papers

Patterns in econometrician reasoning

1. Define the target parameter first (as opposed to writing down a regression specification first)
2. Identification (e.g., parallel trends)
3. Estimation
4. Aggregation
5. Inference

Defining the target parameter

Major part of the new econometrics is to always start with the target parameter and build to it using estimation and identification that “works”

$$\delta = E[Y_{it}^1 - Y_{it}^0 | D_i = 1]$$

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 $\leq 1 - e$.

Intuition of assumption 3: Called overlap or common support. Means there is at least a small fraction of the population that is treated and that for every value of the covariates X there is at least a small chance that the unit is not treated. It's called common support when it's a propensity score but it's just about the distribution of treatment and control across values of X. Very common when dealing with covariate comparisons as otherwise you're extrapolating (curse of dimensionality)

Estimating DD with Assumptions 1-3

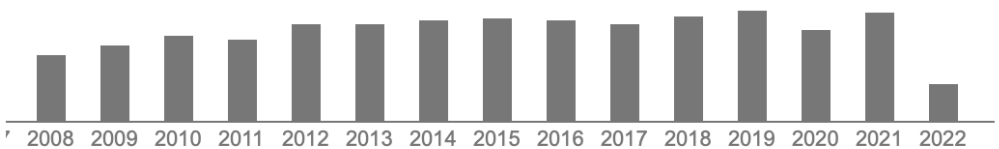
- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005)

Heckman, et al. 1997

Petra Todd

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 8751



Outcome regression

This is the Heckman, et al. (1997) approach where the outcome evolution is modeled with a regression

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

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

Outcome regression

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

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

Inverse probability weighting

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

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

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

These models cannot be ranked

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

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.

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

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
- 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

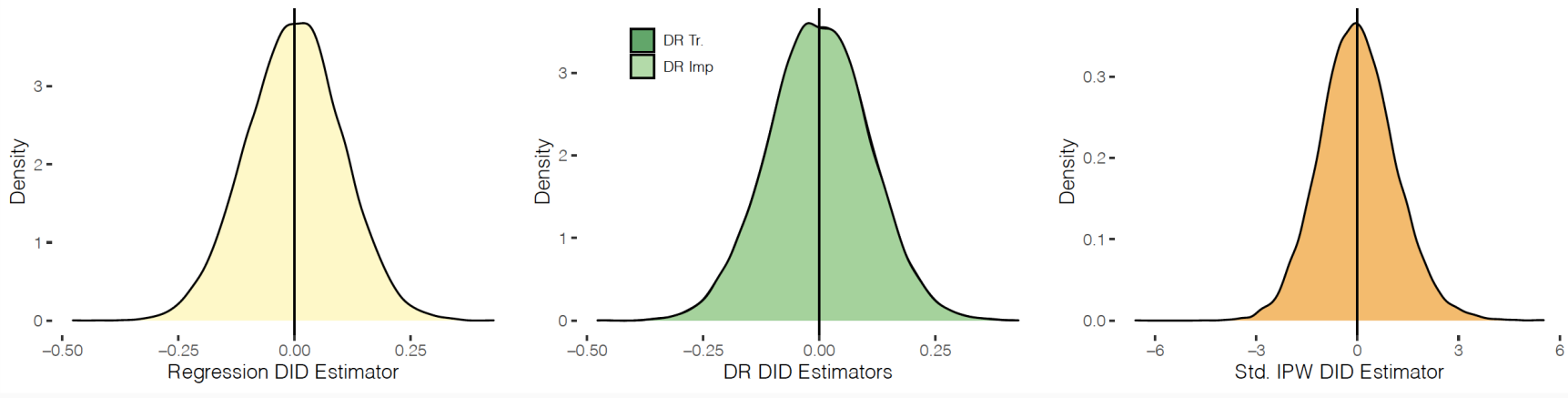
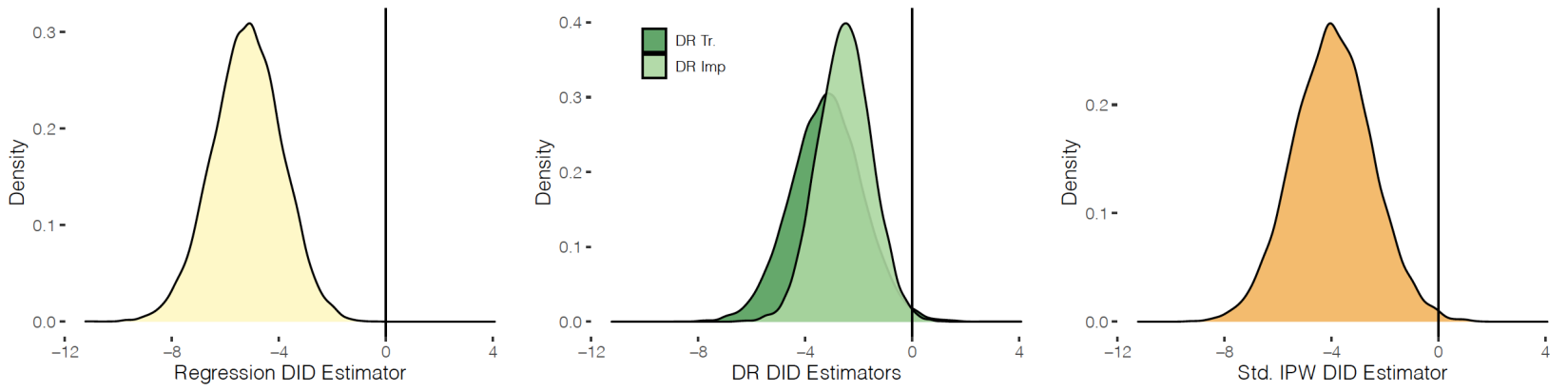


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

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

	All		CPS Controls $N_c = 15,992$	NSW Trainees $N_t = 297$		
covariate	mean	(s.d.)	mean	mean	t-stat	diff
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

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