Difference-in-Differences

Mixtape Session

Roadmap

- Introducing difference-in-differences
  - Numerical examples
  - Potential outcomes
  - Identification
- Estimation
  - OLS Specification
  - Event study
  - Triple difference
  - Falsifications
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD

Miasma Debates

- DiD is a 19th century tool used in health policy debates
- Dominant disease theory in 19th century was *miasma* – disease caused by smelly vapor
- Keep in mind – microorganisms would not be identified until much later, partly caused by poor resolution in microscopes (Freedman 2007)

Miasma I: Ignaz Semmelweis

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

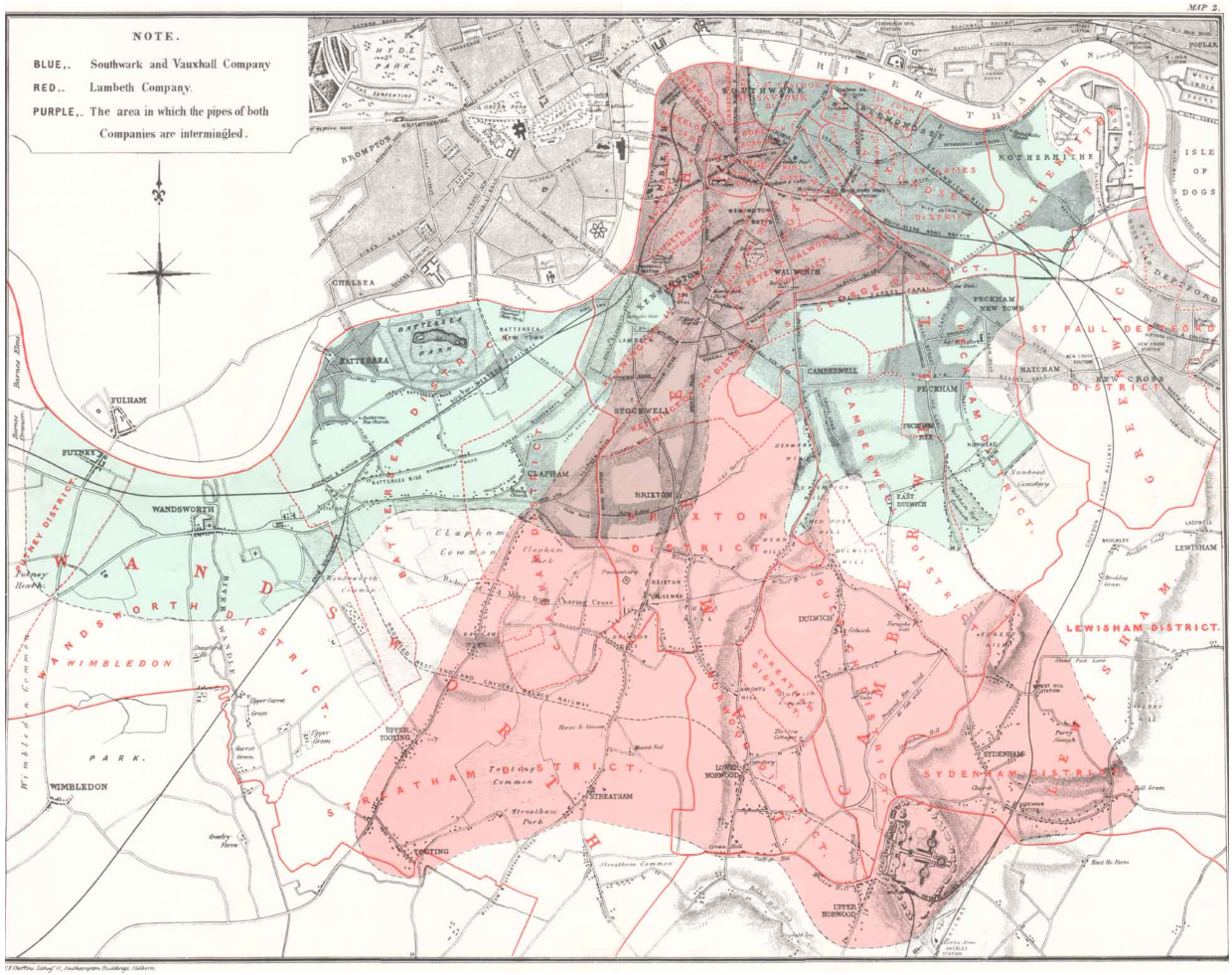
Miasma I: Ignaz Semmelweis

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

Miasma II: John Snow and cholera

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

Figure: Two water utility companies in London 1854



1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

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

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

What is L and SV?

1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

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

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

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

2) Interrupted time series design

Table: Lambeth, 1849 and 1854

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

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

What is required for this estimator to be unbiased?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

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

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

How do we calculate T_{SV}?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

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

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

How do we calculate T_L?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

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

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

This second term is called “parallel trends”

Potential outcomes review

- DiD really can't be understood without committing to some common causality language
- Standard language is the potential outcomes model, sometimes called the Rubin-Neyman model
- Don't go over potential outcomes too fast or you'll miss all the fun
- Potential outcomes are thought experiments about worlds that never existed, but which *could have*

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if pipe inlet is upstream at time } t \\ 0 \text{ if pipe inlet is downstream at time } t \end{cases}$$

  where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1\text{: health if drank from upstream at time } t \\ 0\text{: health if drank from downstream at time } t \end{cases}$$

  where j indexes a counterfactual state of the world
- I'll drop t subscript, but note – these are potential outcomes for the same person at the exact same moment in time

Potential vs observed

- A potential outcome Y^1 and an observed outcome Y are distinct
- Potential outcomes are *hypothetical* possibilities describing states of the world but historical outcomes actually occurred
- Potential outcomes become observed outcomes when treatments are assigned (the “switching equation”)

$$Y = DY^1 + (1 - D)Y^0$$

Treatment effect definitions

Individual treatment effect

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

Individual causal effects cannot be calculated because one of the two needed potential outcomes will always be missing. Epistemologically "unknowable" in some important but difficult to define way.

Conditional Average Treatment Effects

Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

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

Again that “epistemological” uncertainty. We can estimate the ATT, but never be sure due to missing potential outcomes for the treated group

Identification without randomization

- We may be unable to randomize – not because we lack the imagination, but because we lack the permission
- If we cannot randomize, then how does DiD identify a treatment effect, and which treatment effect?
- DiD identifies the ATT, and since we are missing Y^0 for treated group, we will restrict counterfactual Y^0 in expectation

DiD equation

I call this the DiD equation, but Goodman-Bacon calls it the “2x2”; I’ll use his k and U notation for treated and untreated groups

$$\widehat{\delta}_{kU}^{2x2} = \Big(E[Y_k|Post] - E[Y_k|Pre] \Big) - \Big(E[Y_U|Post] - E[Y_U|Pre] \Big)$$

k index people with Lambeth, U index people with Southwark and Vauxhall, $Post$ is after Lambeth moved pipe upstream, Pre before Lambeth moved its pipe (baseline), and $E[y]$ mean cholera mortality.

DiD equation

"Pre" and "Post" refer to when Lambeth, k, was treated which is why it is the same for both k and U groups

If we had more than one treatment group, then "Pre" and "Post" no longer are defined for all units

Potential outcomes and the switching equation

$$\widehat{\delta}_{kU}^{2x2} = \underbrace{\Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)}_{\text{Switching equation}}$$

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

Parallel trends bias

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

Identification

Parallel trends

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

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

In words: "The evolution of cholera mortality for Lambeth *had it kept its pipe downstream* is the same as the evolution of cholera mortality for Southwark and Vauxhall".

It's in red so you know it's a nontrivial assumption. But why? Can't we just check?

How does the science work

- You've probably heard people say RCT is the gold standard for causal inference. But why?
- Because randomization gives *near certainty* that selection bias won't exist
- Don Rubin commented once, "we know how the science works"

Independence

Independence assumption

Treatment is independent of potential outcomes

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

RCTs use *independence* to estimate causal effects; DiD does not

Independence

Independence allows us to write down conditional expected potential outcome equations like

$$E[Y^0|D=1] = E[Y^0|D=0]$$

In the simple comparison in means, $\widehat{\delta}_{cs} = D + (L - SV)$, independence implies $L = SV$ (no selection bias).

Parallel trends

- Unlike selection bias under RCT being zero, **there is no science of parallel trends**
- DiD is a harder design because it doesn't rely on randomization (it relies on parallel trends)
- Ashenfelter's Dip – treatment and control group units prior to treatment seemed like they would've diverged anyway
- Now we do an exercise together. Open up the word document and excel spreadsheet.

Roadmap

- Introducing difference-in-differences
  - Numerical examples
  - Potential outcomes
  - Identification
- Estimation
  - OLS Specification
  - Event study
  - Triple difference
  - Falsifications
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD

OLS Specification

- Properly specified OLS model will also identify the ATT when there is only two groups and no covariates
- Often preferred because
  - → OLS estimates the ATT under parallel trends
  - → Easy to calculate the standard errors
  - → Easy to include multiple periods
- Some issues emerge with differential timing, time varying covariates and continuous treatments, so we take those separately

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

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

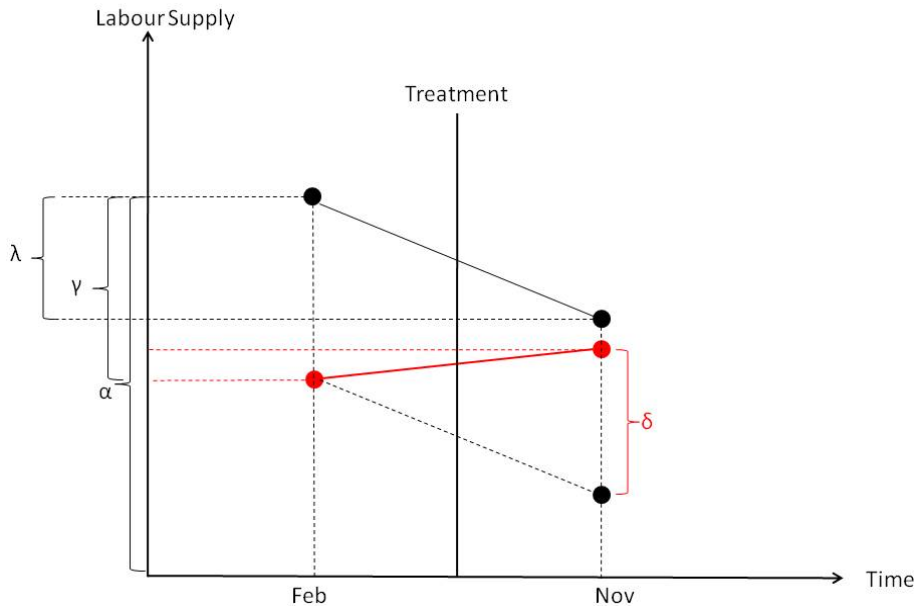
OLS specification of the DiD equation

- 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) $= \delta$

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



Compositional differences violate parallel trends

- 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
- Hong (2011) 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 Napter (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

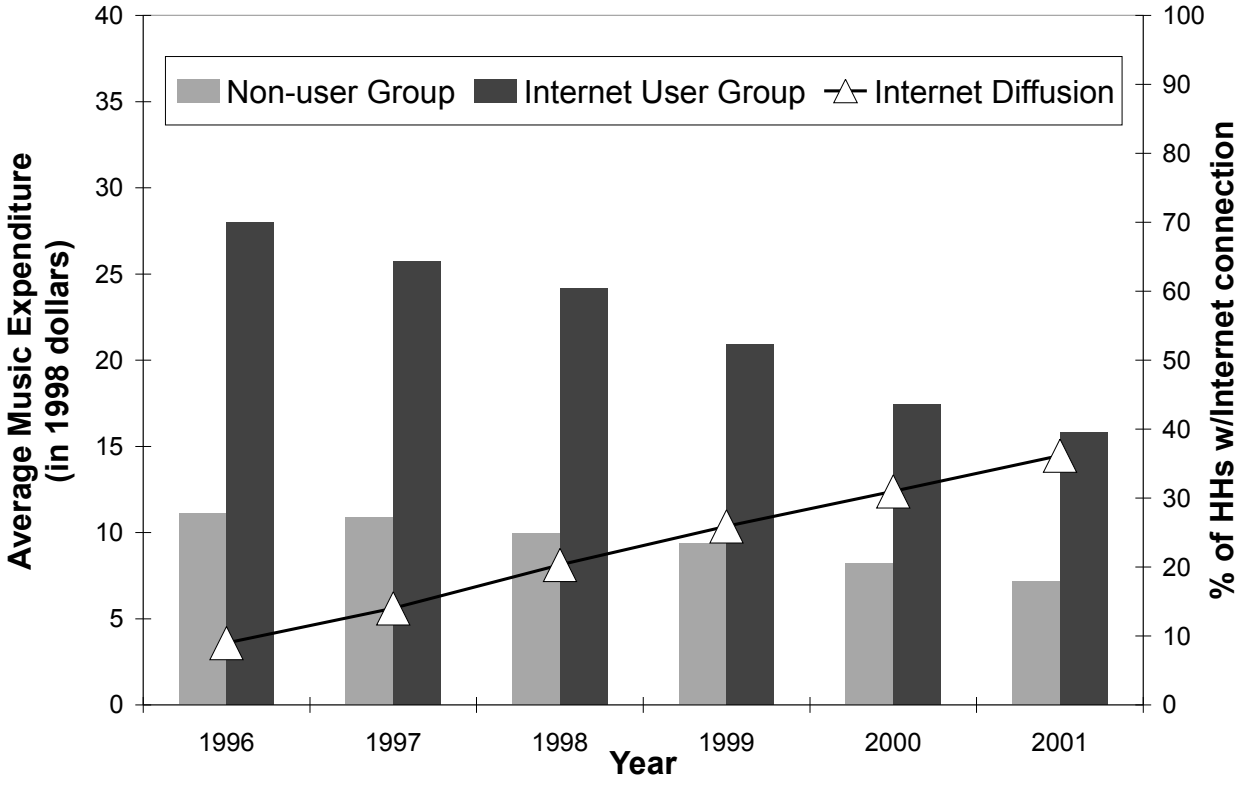


Table 1: Descriptive Statistics for Internet User and Non-user Groups[a]

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)

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)

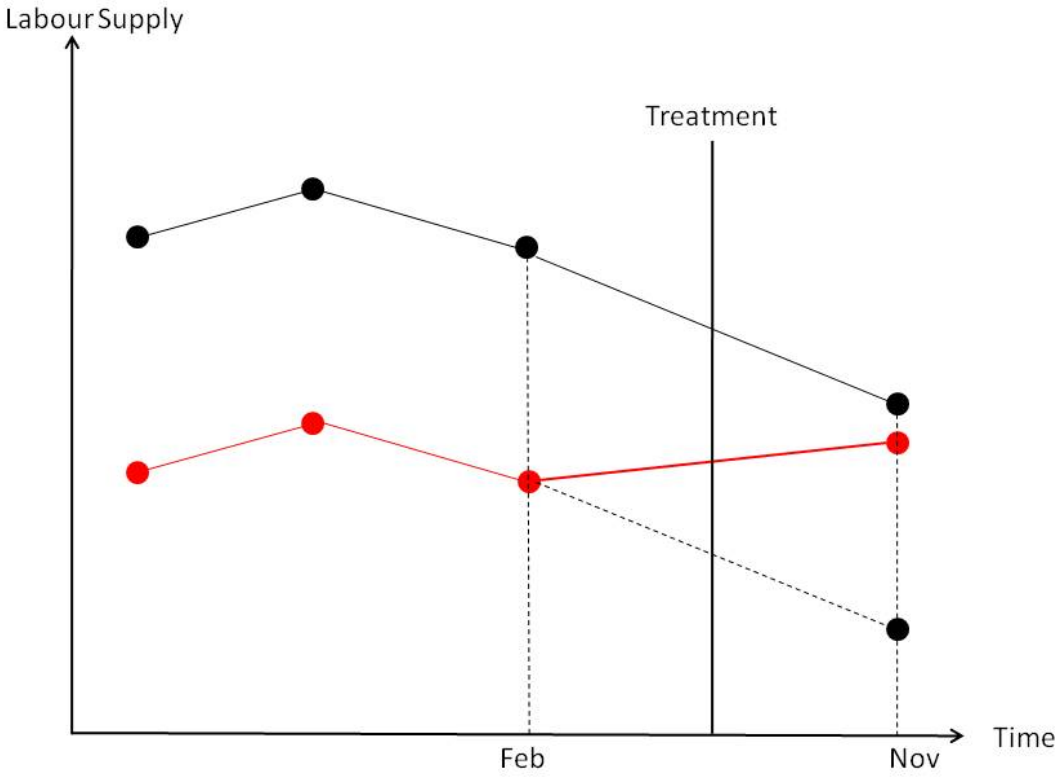
Inference

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)
3. Aggregating the data into one pre and one post period. Literally works if there is only one treatment data. With staggered treatment dates one should adopt the following procedure:
   - → Regress Y_{st} onto state FE, year FE and relevant covariates
   - → Obtain residuals from the treatment states only and divide them into 2 groups: pre and post treatment
   - → Then regress the two groups of residuals onto a post dummy

Evidence for parallel trends: pre-trends

- The identifying assumption for all DD designs is parallel trends, but since we cannot verify parallel trends, we often look at pre-trends
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

Plot the raw data when there's only two groups



Event study regression

- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

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

  - → D equals the treatment group interacted with the calendar year
  - → Treatment occurs in year 0, no anticipation, drop baseline $t-1$
  - → Includes q leads or anticipatory effects
  - → Includes m lags or post treatment effects

Event study regression

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

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

Under parallel trends and no anticipation, then you expect $\widehat{\mu}$ coefficients to be zero, which gives you confidence that parallel trends holds (but is not a guarantee)

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

Abstract

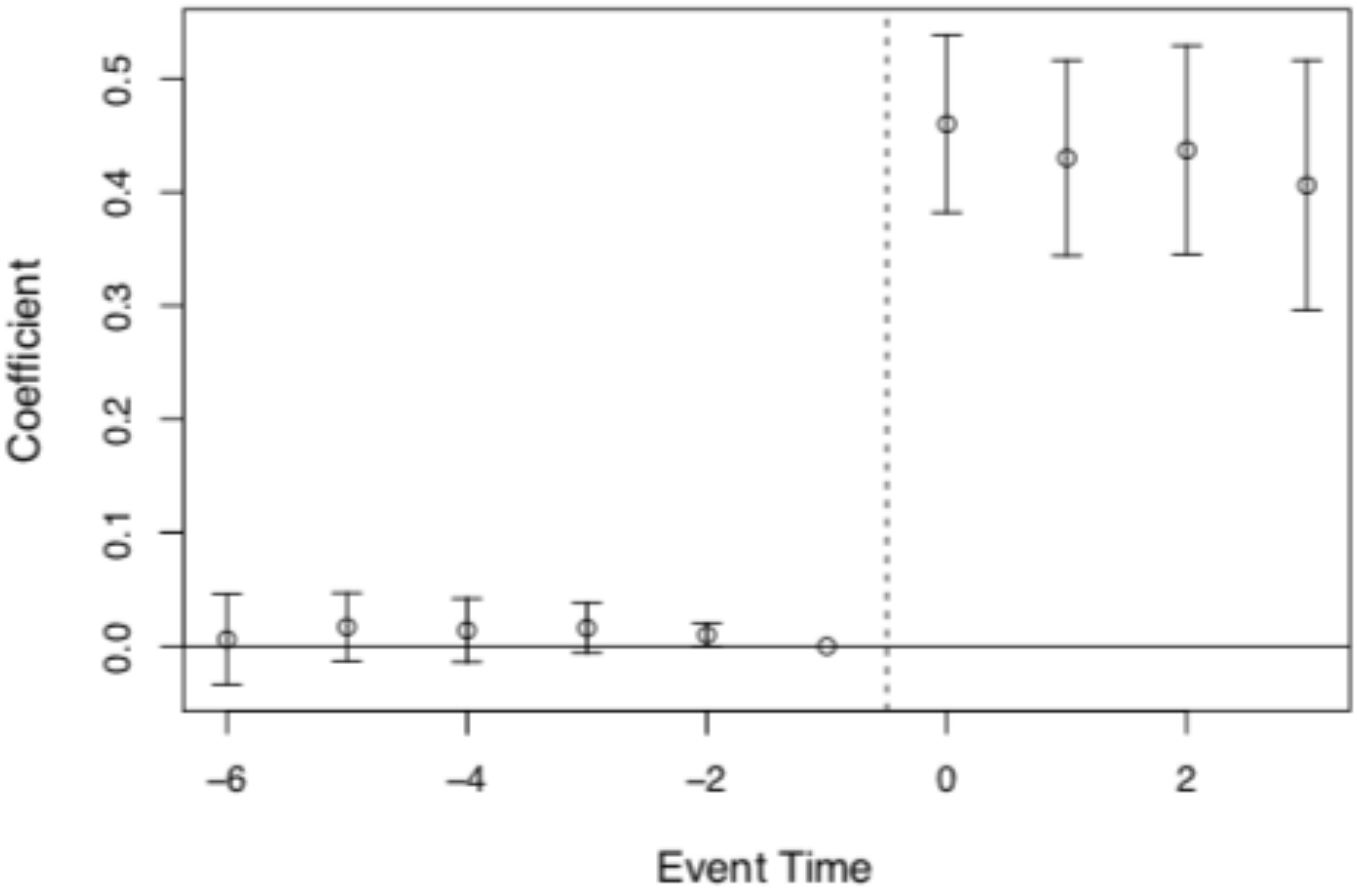
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults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

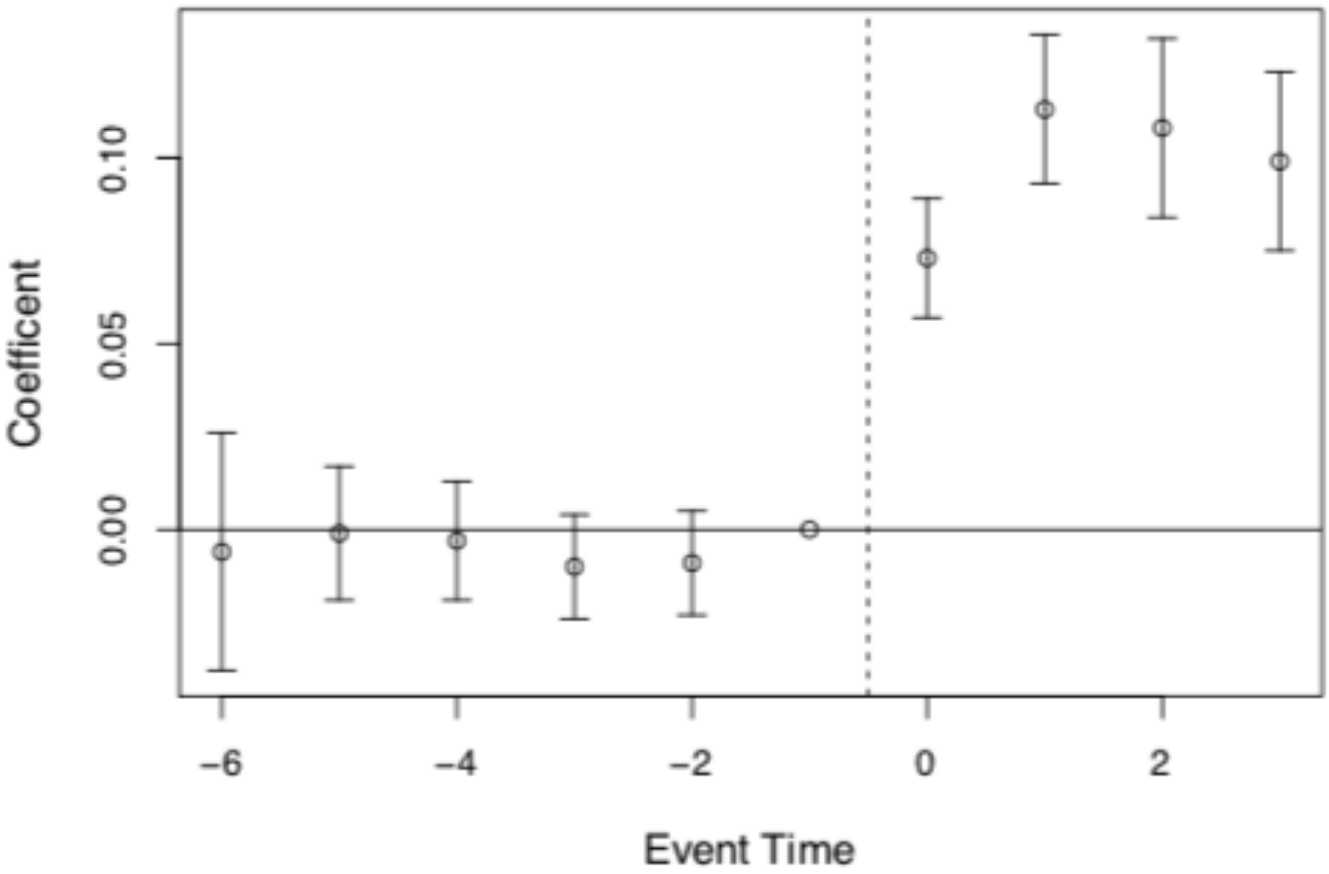
Issue Section: Article

Evidence

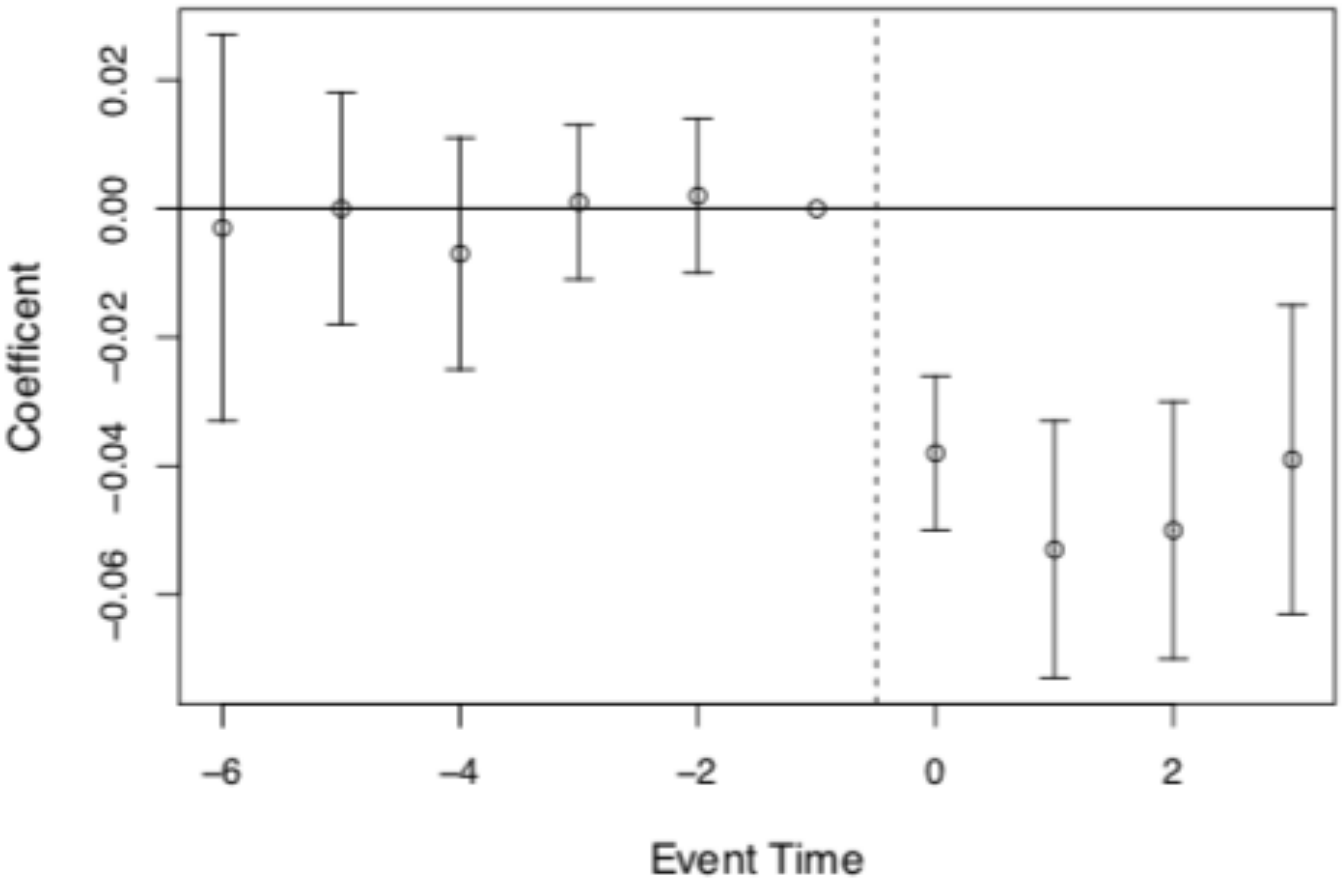
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Event study** – Show leads and lags on mortality



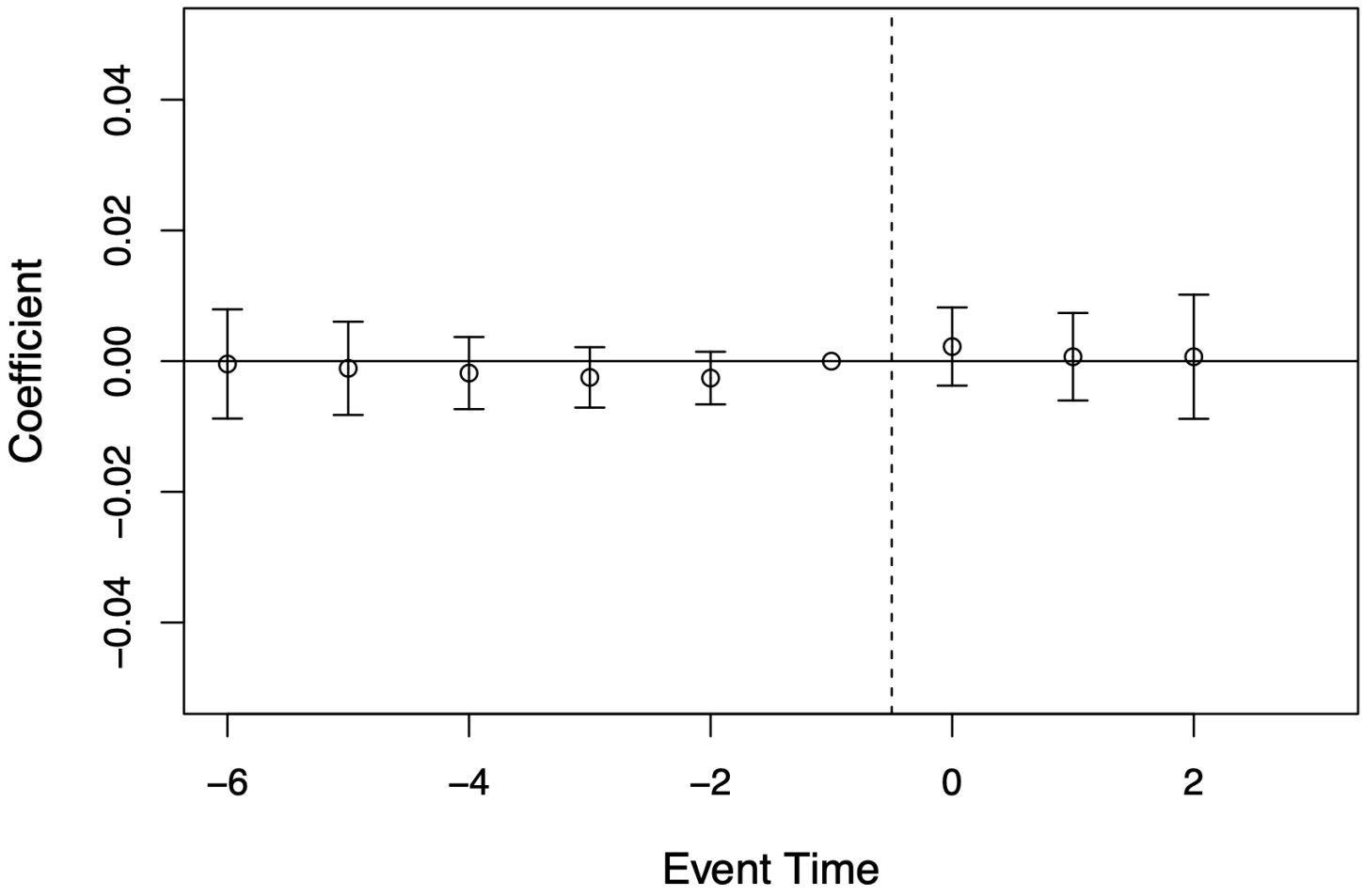
(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured


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

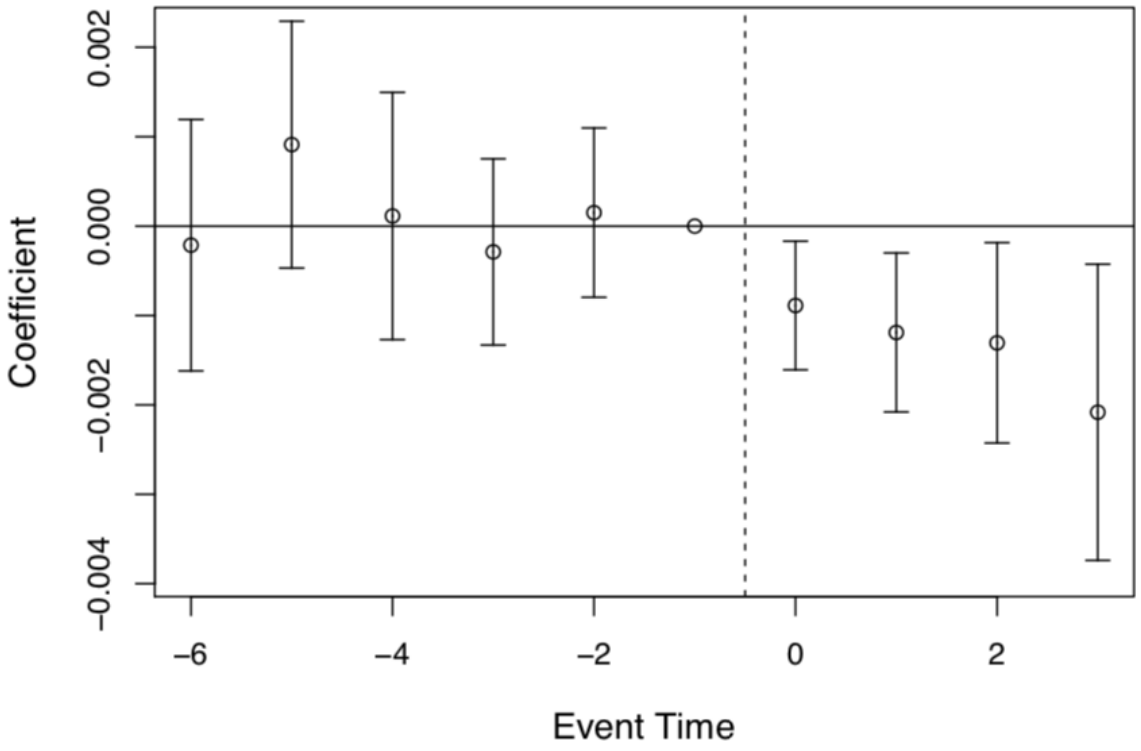


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

Additional identification things

- Very common for readers and others to request a variety of "robustness checks" from a DD design
- Think of these as along the same lines as the event study we just discussed
  - → Falsification test using data for alternative control group (e.g., Medicare population)
  - → Falsification test using alternative "placebo" outcome that should not be affected by the treatment

Table: Difference-in-Difference-in-differences

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

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 raises the epistemological credibility of the first result, interestingly
- 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

- Introducing difference-in-differences
  - Numerical examples
  - Potential outcomes
  - Identification
- Estimation
  - OLS Specification
  - Event study
  - Triple difference
  - Falsifications
- Including Covariates
  - Inverse probability weighting
  - Double Robust DiD

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

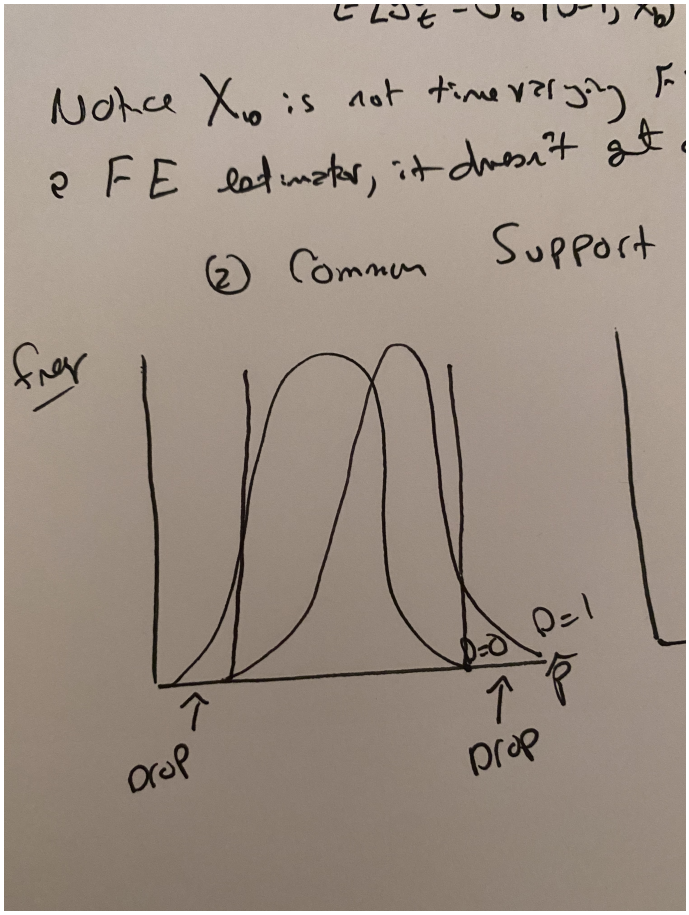
Common support

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

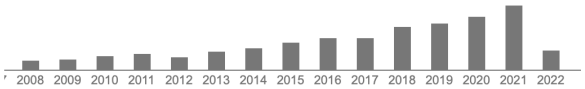
- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

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 modularity 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

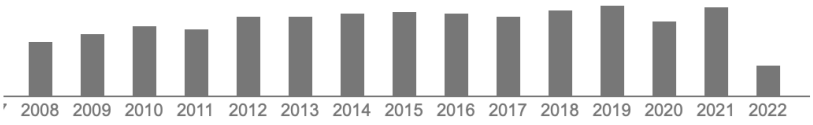
- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005)

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

$$\begin{aligned}
&E[Y_1^0 - Y_0^0|D=1,X] = E[Y_1^0 - Y_0^0|D=0,X] \\
&E[Y_1^0|D=1,X] - E[Y_0^0|D=1,X] = E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X] \\
&E[Y_1^0|D=1,X] = E[Y_0^0|D=1,X] + E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X] \\
&E[Y_1^0|D=1,X] = E[Y_0|D=1,X] + E[Y_1|D=0,X] - E[Y_0|D=0,X]
\end{aligned}$$

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:

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

Eliminating terms, we get:

$$\begin{aligned}\delta^{DD} = \quad & \delta + \\ & (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})\end{aligned}$$

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

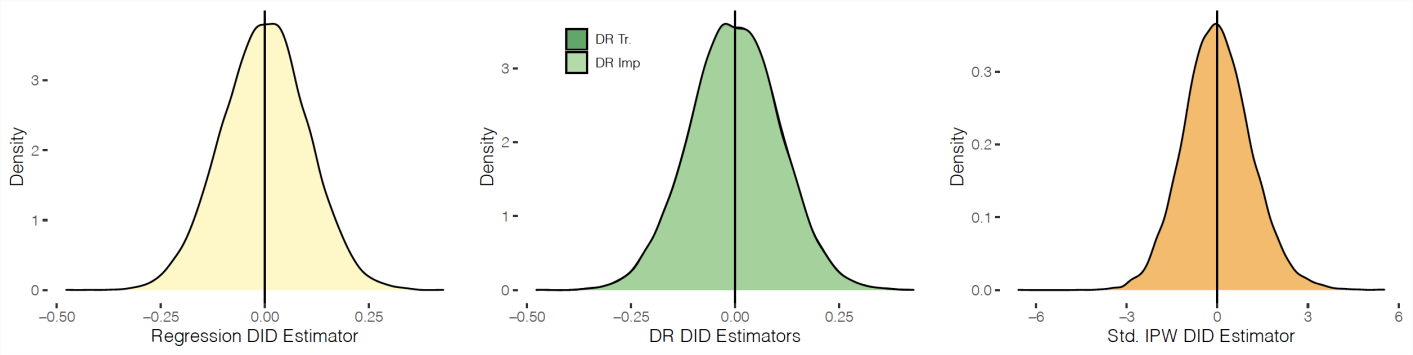
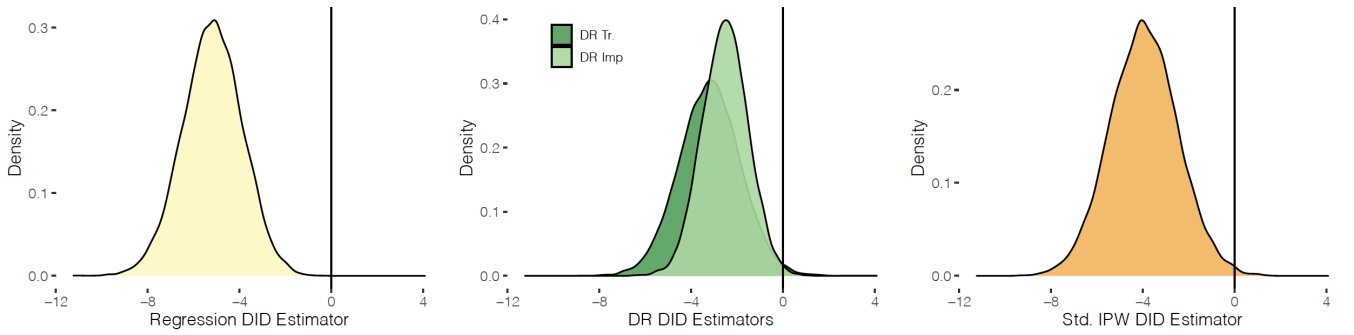


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	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



Code

There is code in R and Stata

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

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

Concluding remarks

- So we hopefully see a few of the key elements of DiD
  - → DiD equation and ATT equation are distinct
  - → Pre-trends is a placebo for evaluating parallel trends
  - → DiD is a design; OLS is a model
- Including covariates posed problems for the canonical OLS specification
- Problem with that OLS specification was *the specification*, not OLS – it assumed constant treatment effects