

Causal Inference

MIXTAPE SESSION



Roadmap

Difference-in-differences

Two group case

Estimation

Event study

Covariates

Differential timing

Revisiting event studies

Alternative DD estimators

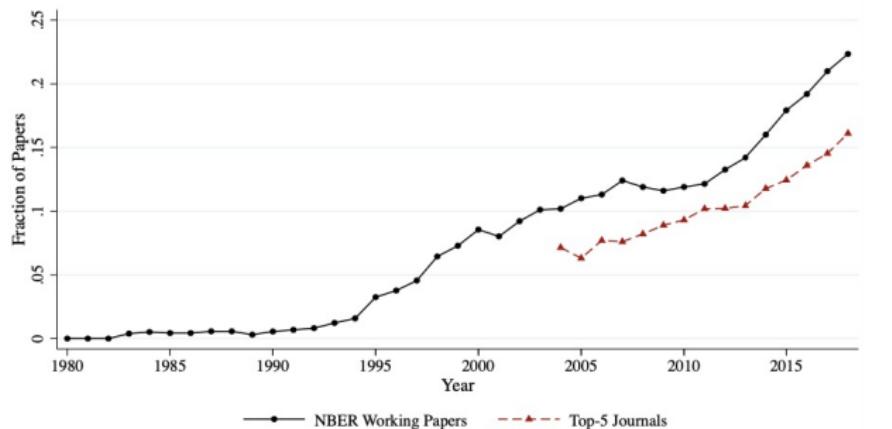
Conclusion

What is difference-in-differences (DiD)

- A group of units (treatment) are assigned some treatment and then compared to a group of units (control, or comparison) that weren't
- Historically used in 19th century health policy debates, re-introduced in 1970s and 1980s by Orley Ashenfelter and David Card
- It has become the most popular of all quasi-experimental research designs

Figure: Currie, et al. (2020)

A: Difference-in-Differences



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:

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

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”

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

k index people with PhDs, U index people without PhDs, $Post$ is after k individuals got their PhD, Pre before k group had gotten their PhDs (baseline), and $E[y]$ mean happiness.

“Pre” and “Post” refer to when our treatment group, k , was treated and thus is the same for both k and U groups

Potential outcomes and the switching equation

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

Parallel trends bias

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

Identification

Parallel trends

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

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

“The evolution of happiness for PhDs *had they not gotten their PhDs* is the same as the evolution of happiness for those who never got their PhDs”. Nontrivial assumption.

Contrast PT with independence

Recall identification of treatment effects with randomized treatment assignment

Independence assumption

Treatment is independent of potential outcomes

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

Allows us to write down conditional expected potential outcome equations like $E[Y^0|D = 1] = E[Y^0|D = 0]$ (no selection bias)

How the science works

- Randomization gives *near certainty* that selection bias will not exist in our contrasts of treatment and control
- Don Rubin commented once, “we know how the science works”
- But there is **no science of parallel trends** – it may or may not hold in observational data – so epistemological uncertainty seems greater than with the RCT

Roadmap

Difference-in-differences

Two group case

Estimation

Event study

Covariates

Differential timing

Revisiting event studies

Alternative DD estimators

Conclusion

Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

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

Interrupted time series design

Table: Lambeth, 1849 and 1854

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

Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$	$T + D$	D
	After	$Y = L + T + D$		
Southwark and Vauxhall	Before	$Y = SV$	T	
	After	$Y = SV + T$		

Sample averages

$$\widehat{\delta}_{kU}^{2x2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Population expectations

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

Potential outcomes and the switching equation

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

Parallel trends bias

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

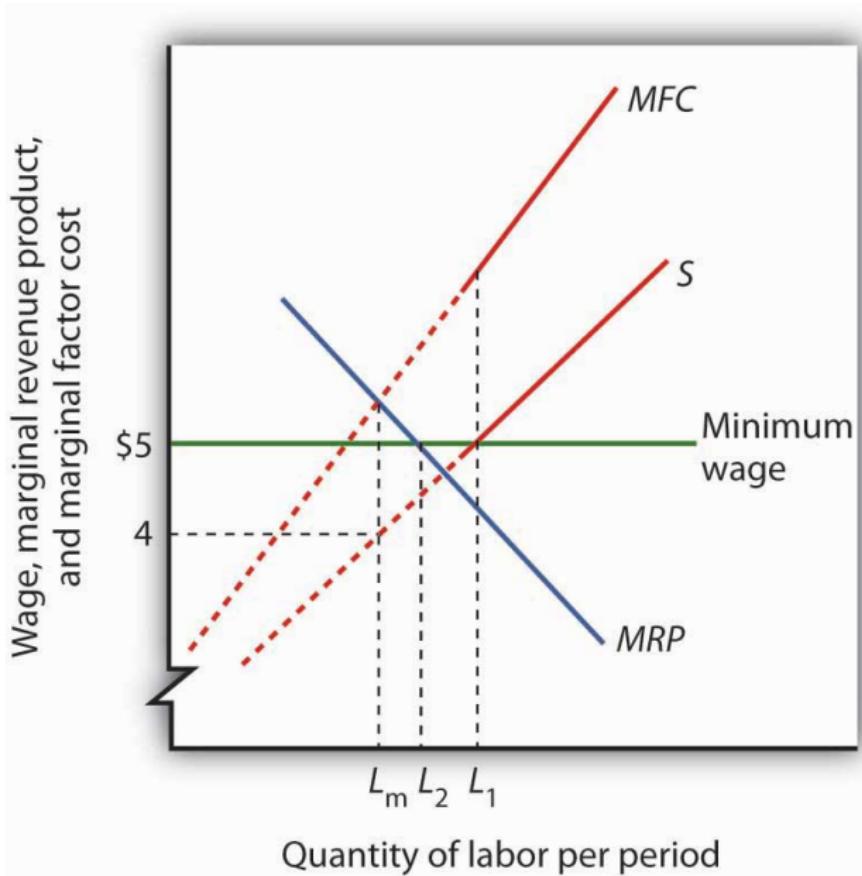
Another famous DD study

- Card and Krueger (1994) was a seminal study on the minimum wage both for the result and for the design
- Not the first time we saw DD in the modern period - there's Ashenfelter (1978) and Card (1991) - but got a lot of attention

Competitive vs noncompetitive markets

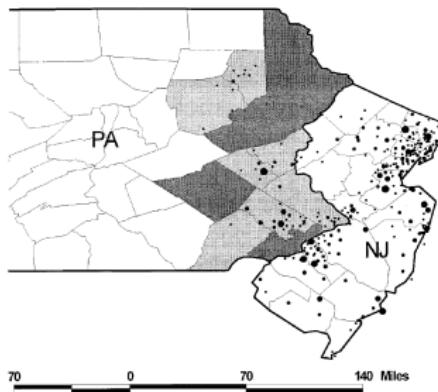
- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In a competitive input market, increases in the minimum wage would move us up a downward sloping labor demand curve → employment would fall
- Monopsony (imperfect labor markets) suggest the opposite effect whereby raising the minimum wage increases employment

Monopsony's minimum wage predictions



Card and Krueger (1994)

- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey - shoeleather!

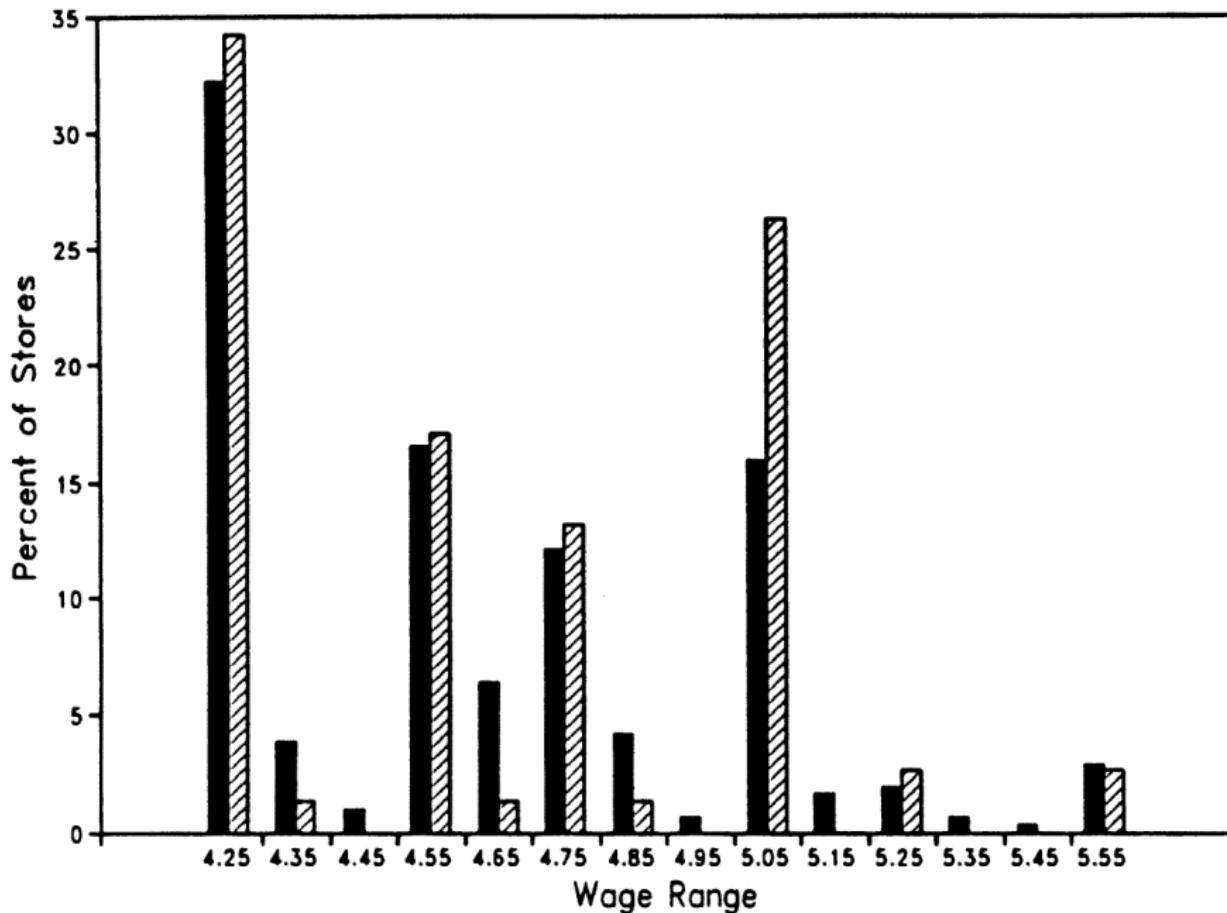
Parallel trends assumption

- Key identifying assumption is the “parallel trends” assumption

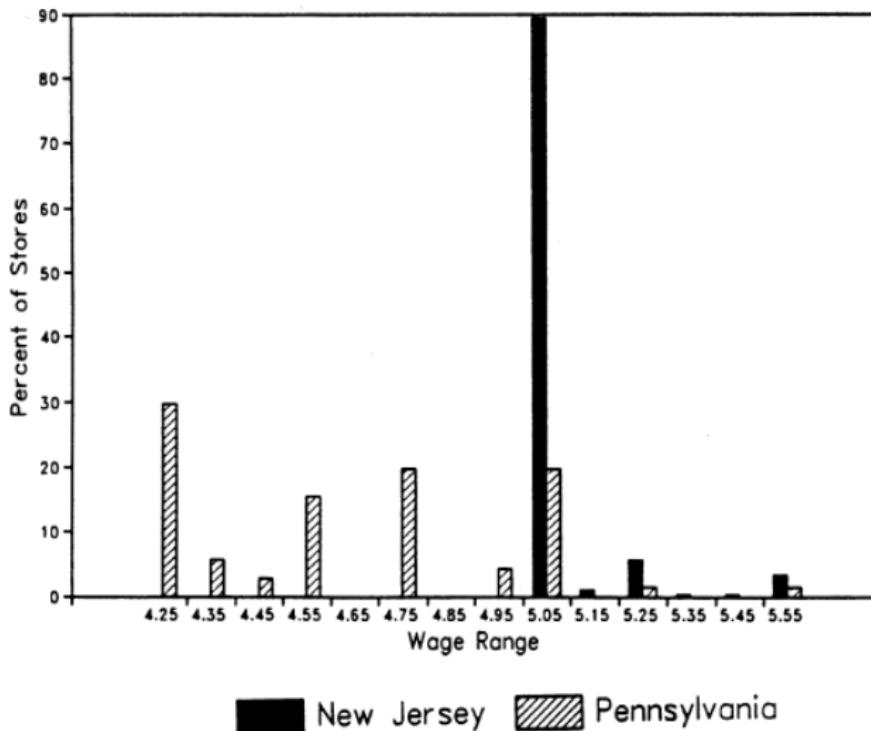
$$\underbrace{[E[Y_{NJ}^0 | Post] - E[Y_{NJ}^0 | Pre]] - [E[Y_{PA}^0 | Post] - E[Y_{PA}^0 | Pre]]}_{\text{Non-parallel trends bias}}$$

- Note the counterfactual - it is *not testable* no matter what someone tells you, bc New Jersey's post period potential employment in a world with a lower minimum wage is unobserved
- Let's look at this a couple of different ways, including a graphic showing the binding minimum wage

February 1992



November 1992



Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment rose in NJ relative to PA after the minimum wage change - consistent with monopsony theory

Regression DD

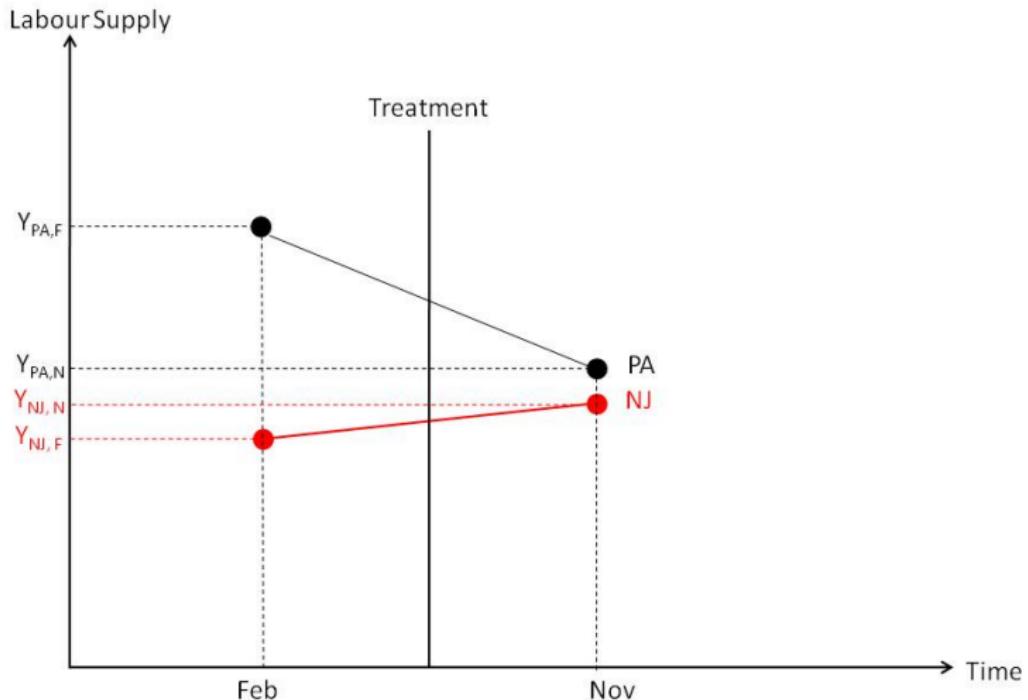
- There are several good reasons to use TWFE
 - It estimates the ATT under parallel trends
 - It's easy to calculate the standard errors
 - It's easy to include multiple periods
 - We can study treatments with different treatment intensity. (e.g., varying increases in the minimum wage for different states)
- But there are bad reasons, too, which I'll discuss under differential timing and covariates

Regression DD - Card and Krueger

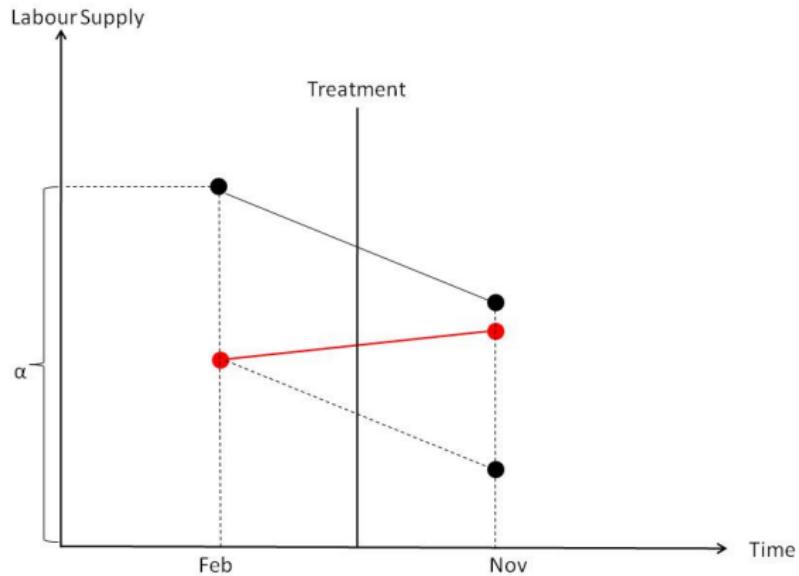
- In the Card and Krueger case, the equivalent regression would be:

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

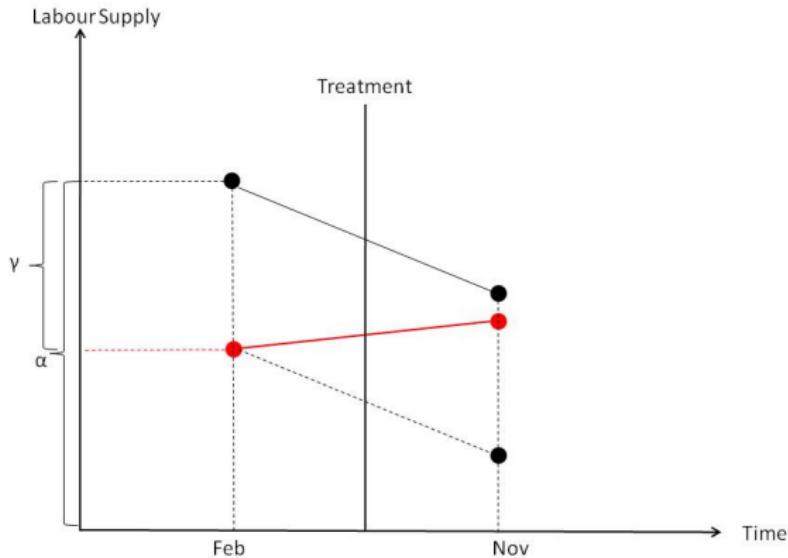
- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$



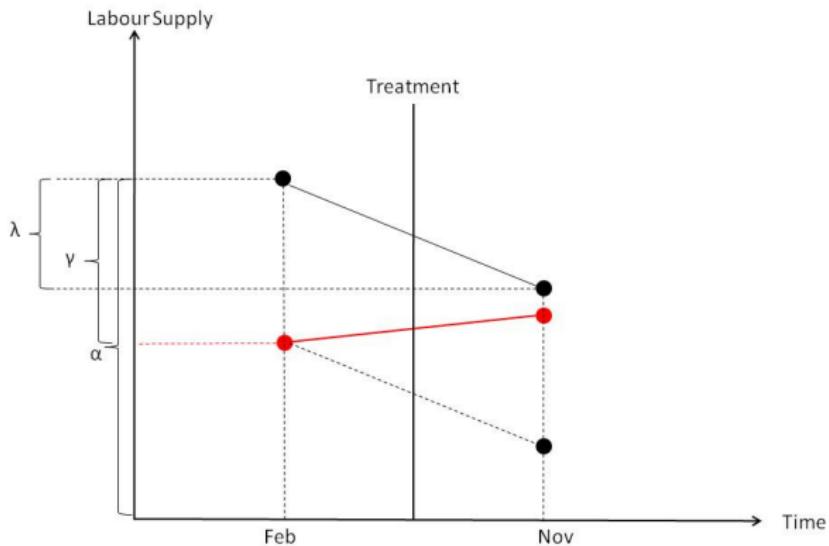
$$Y_{ist} = \alpha + \gamma N J_s + \lambda d_t + \delta (N J \times d)_{st} + \varepsilon_{ist}$$



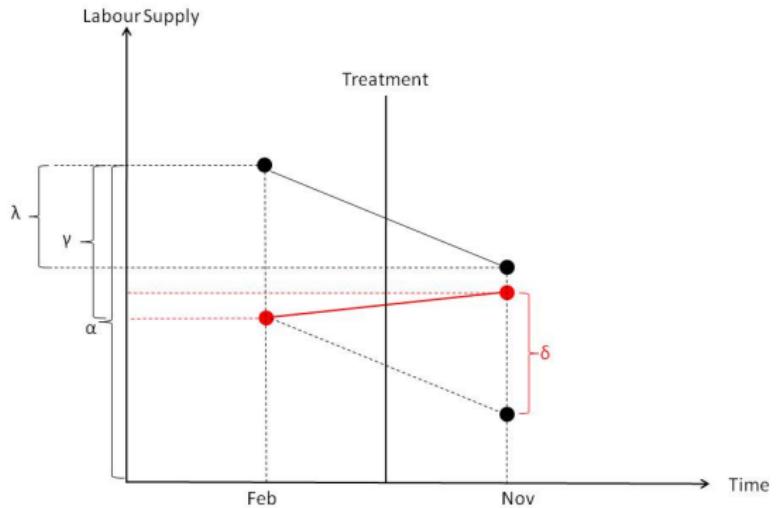
$$Y_{ist} = \alpha + \gamma N J_s + \lambda d_t + \delta (N J \times d)_{st} + \varepsilon_{ist}$$



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



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



Losing parallel trends

- If parallel trends doesn't hold, then ATT is not identified
- But, regardless of whether ATT is identified, OLS always estimates the same thing
- That's because OLS uses the slope of the control group to estimate the DD parameter, which is only unbiased if that slope is the correct counterfactual trend for the treatment group

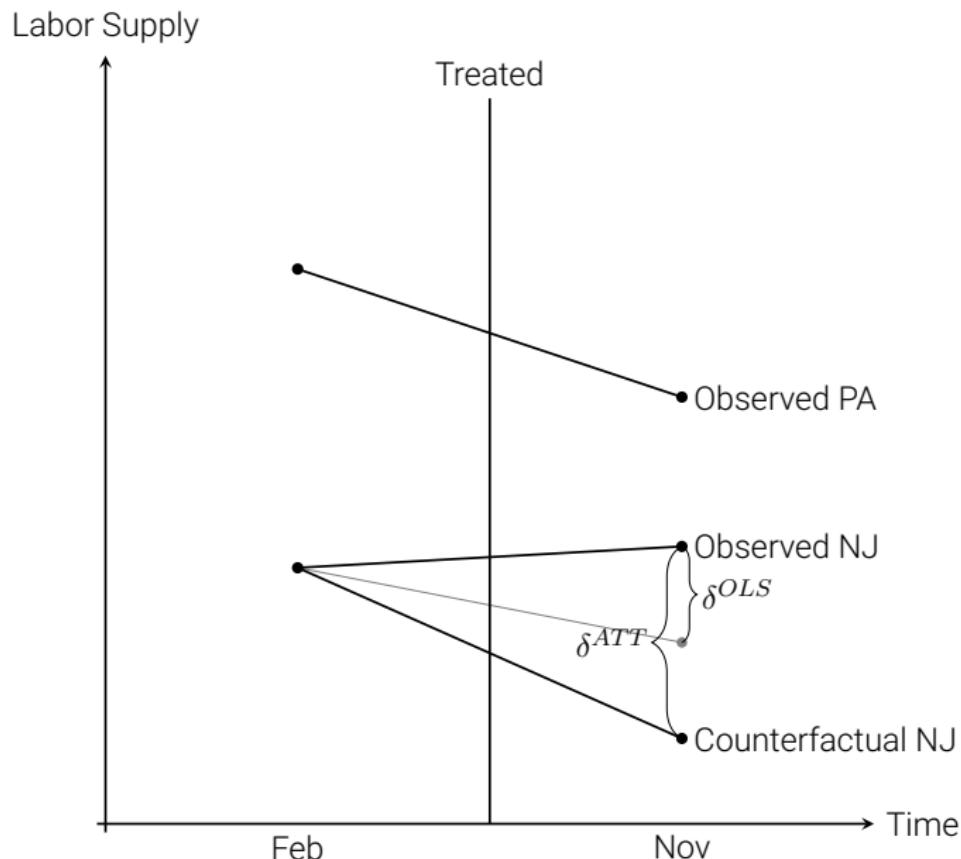


Figure: DD regression diagram without parallel trends

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 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

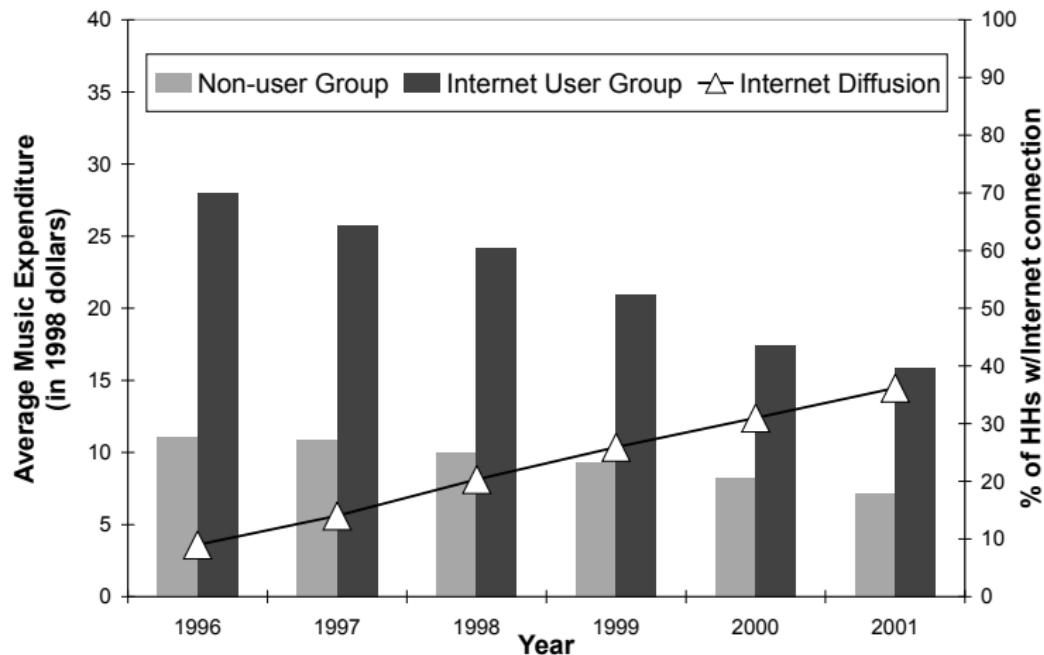


Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

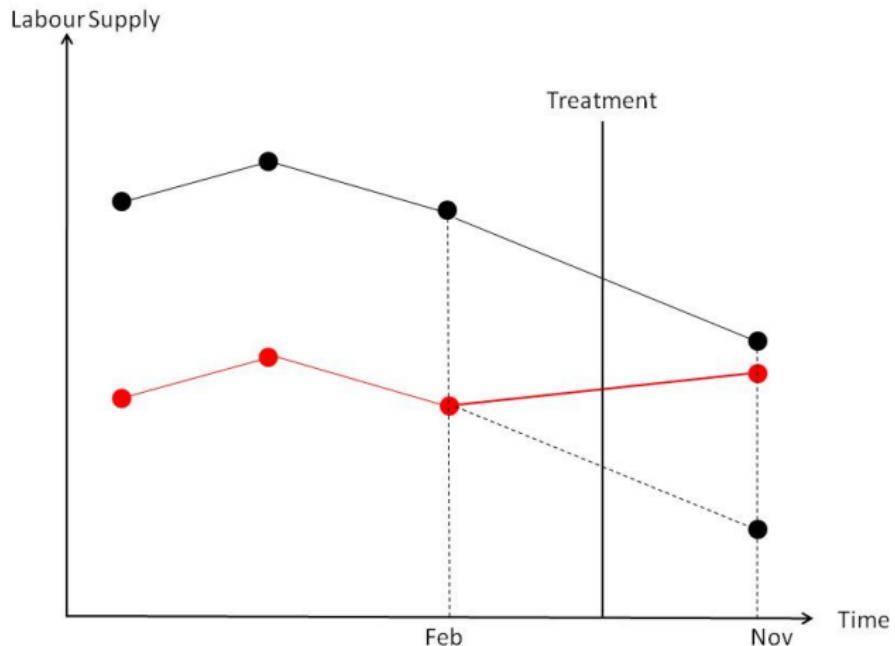
Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.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)

Evidence for parallel trends: pre-trends

- The identifying assumption for all DD designs is parallel trends
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check a hunch for parallel trends using 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

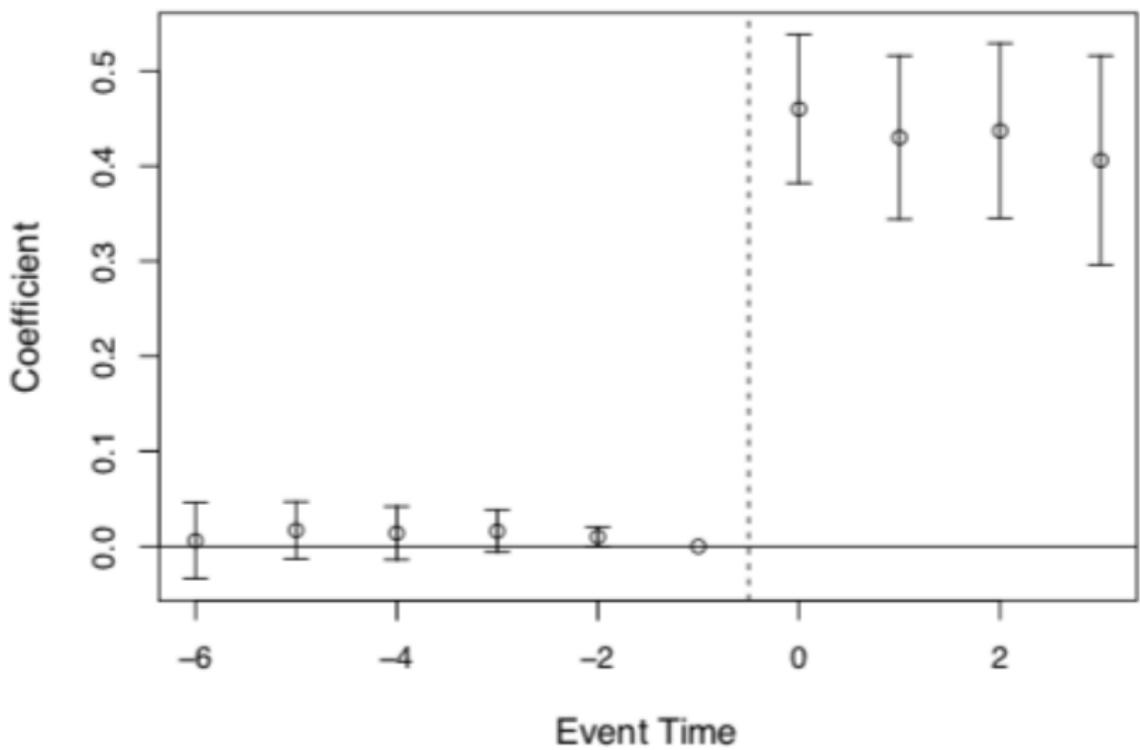
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

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

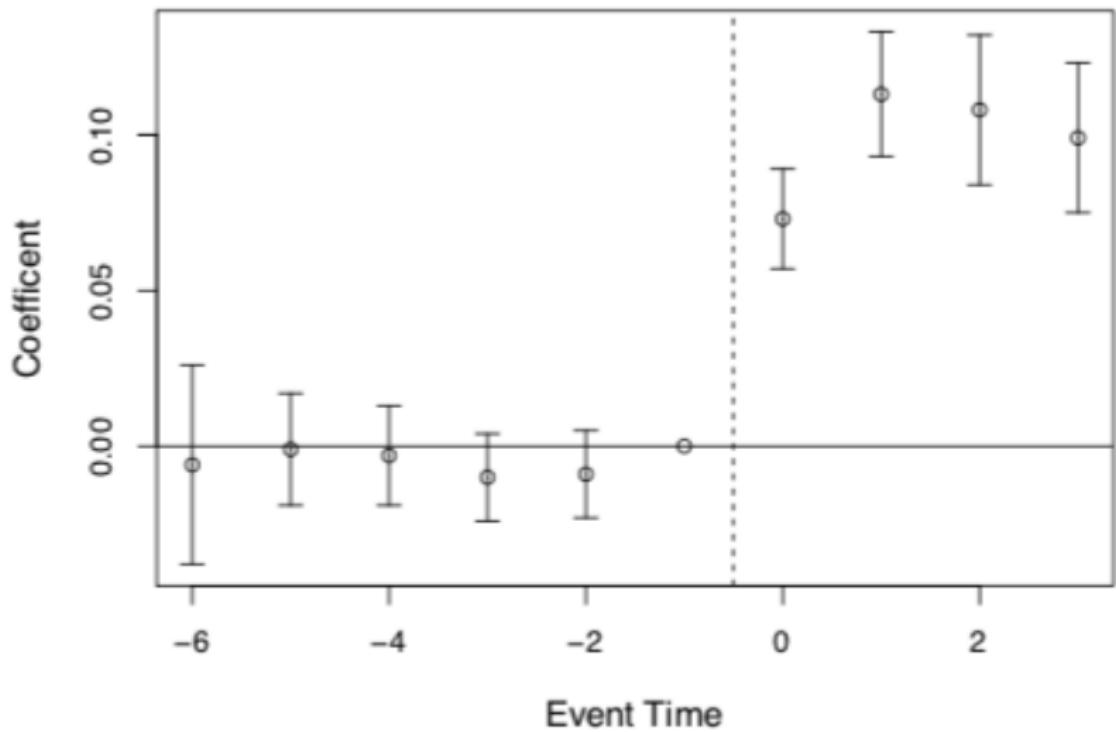
- Treatment occurs in year 0
- Includes q leads or anticipatory effects
- Includes m leads or post treatment effects

Medicaid and Affordable Care Act example

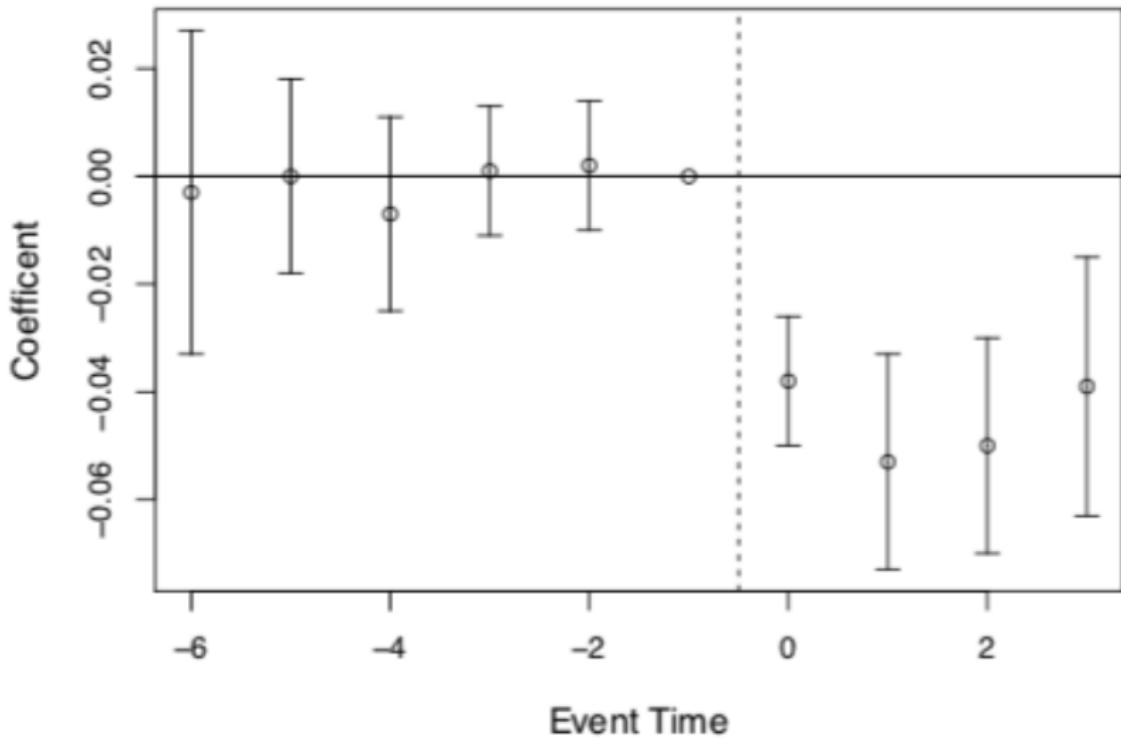
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

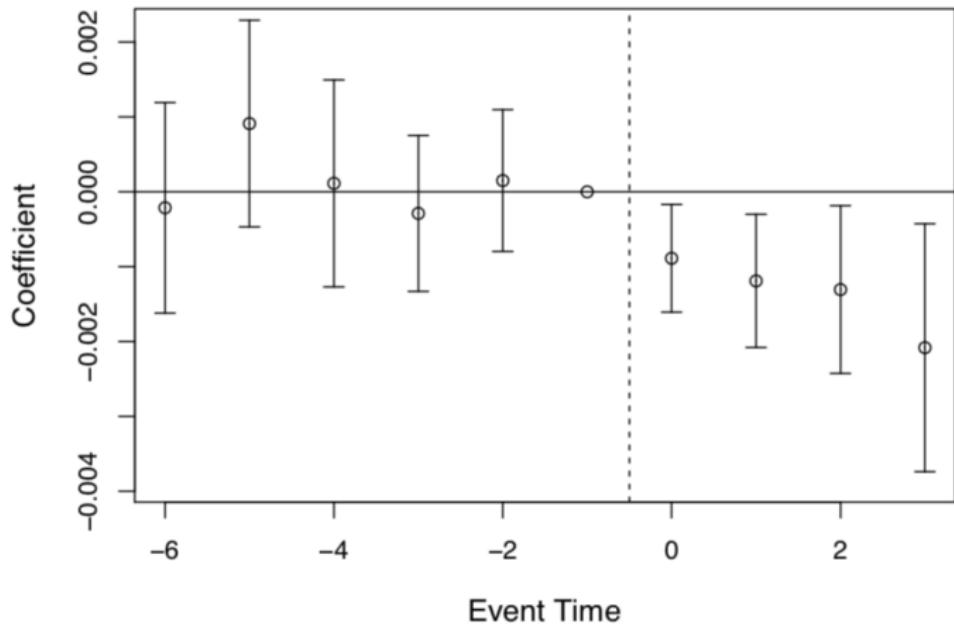


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on annual mortality

Standard errors in DD strategies

- Many papers using DD strategies use data from many years – not just 1 pre and 1 post period
- The variables of interest in many of these setups only vary at a group level (say a state level) and outcome variables are often serially correlated
- As Bertrand, Duflo and Mullainathan (2004) point out, conventional standard errors often severely underestimate the standard deviation of the estimators – standard errors are biased downward (i.e., too small, over reject)

Standard errors in DD – practical solutions

- Bertrand, Duflo and Mullainathan (2004) propose the following solutions:
 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

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
 - 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	$NJ + T + NJ_t + s_t$	$T + NJ_t + s_t$	$l_t - s_t$	D
		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	$PA + T + PA_t + s_t$	$T + PA_t + s_t$	$l_t - s_t$	D
		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

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

- 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

High level

"A good way to do econometrics is to look good for natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us. – Daniel McFadden

Conditional parallel trends

- We will discuss two papers now: Abadie (2005) and Sant'Anna and Zhao (2020)
- In some ways you can think of Abadie (2005) as the father of Callaway and Sant'Anna (which we discuss under the differential timing section) only here we don't use differential timing
- The Abadie (2005) is really used best for longitudinal data or repeated cross sections where treatment occurs at one point in time
- But like CS, it's used for modeling the differential selection based on what you think are covariates, which means you need to think carefully about what those might be

DD method at its core

- Abadie (2005) proposed a method to estimate the ATT
- The method is a DD type estimator, but isn't using TWFE
- You need treatment and comparison group, before and after treatment
- But you also need conditional parallel trends (based on X)
- Kind of neat but it's a lot like Callaway and Sant'Anna, only not for differential timing interestingly

Why do this anyway?

- In a DD, we may need to control for X because treatment is only conditional on X
- But in TWFE, when you controlling for baseline X, it gets absorbed by the unit fixed effects
- And when you use time-varying controls, you can get even stranger weights than we had already seen from Bacon
- To get around this, he won't be proposing an OLS model with fixed effects, but he will be proposing a simpler difference in means in a DD framework by a specific form of weight called the propensity score which has been estimated with polynomial series

Not a critique, but an estimator

- Goodman-Bacon was a critique of TWFE, not a proposed estimator
- CS was a proposed estimator
- Abadie is a proposed estimator
- Let's look at the steps involved

Three step method

1. Compute each unit's "after minus before" which is the DD part
2. Then estimate a propensity score which you'll use to weight each unit
3. Finally, compare weighted changes in "after minus before" for treatment versus comparison groups

Inference will take into account step two, which is often the sticky part (see Abadie and Imbens matching paper which shows you can't use the bootstrap for matching, but you can for propensity scores)

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
- b is the “baseline” which is similar to CS using g as the one year pre-treatment
- X 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

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

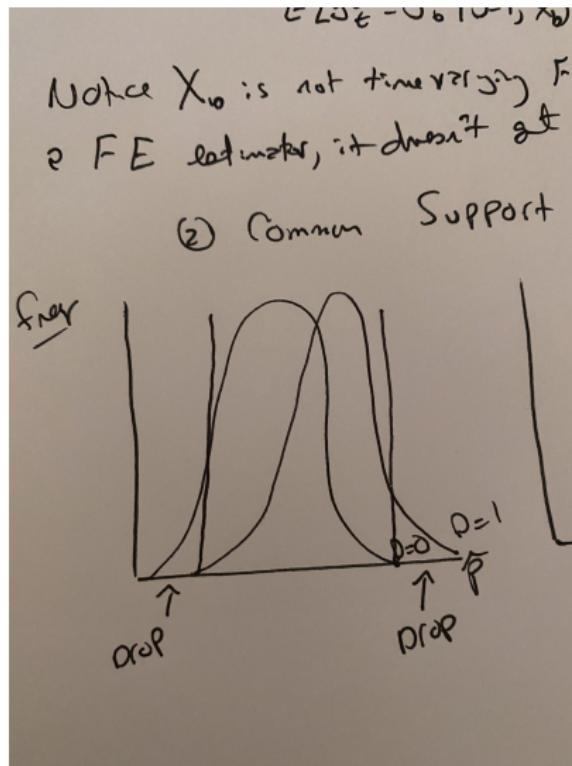
(Notice the b subscript. What is that you think?)

2. Common support

$$Pr(D = 1) > 0 \text{ & } Pr(D = 1 | X) < 1$$

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

Trimming the propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

$$ATT = E[Y_t^1 - Y_t^0 | D_t = 1] \quad (1)$$

Abadie's estimator

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

Propensity scores

- Usually there's almost no guidance that I've seen in how to estimate the propensity score except to say use logit or probit
- Dehejia and Wahba (2002) anyway
- Not so here – this is semi-parametric in the sense that you have to use a series of polynomials based on the X controls
- Weirdly, you can use OLS linear probability models (which I've never seen) or something called series logit estimation

Estimating 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

$$\widehat{Pr}(X_b) = \widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots \varepsilon \quad (3)$$

$$\widehat{Pr}(X_b) = F(\widehat{\gamma}_0 + \widehat{\gamma}_1 X + \widehat{\gamma}_2 X^2 + \dots) \quad (4)$$

Doubly Robust

- DR literature can be found in the older matching literature (Hirano and Imbens 2001; etc.)
- They combine regression and weighting estimators into one specification and are consistent so long as:
 - The regression specification for the outcome is correctly specified
 - The propensity score specification is correctly specified
- DR is a class of estimators that possess this property
- You're basically controlling for X twice: with a linear regression, with a propensity score, to cover your bases

DR DD

- Sant'Anna and Zhao (2020) incorporate DR into DD
- Think of it as a way of incorporating X into our new DD models (I'll show you why)
- It's in the engine of Callaway and Sant'Anna (2020) so we badly need to understand it
- Dense paper with hairy notation; I'll do my best

Literature

- Pedro is excellent at bridging gaps while simultaneously moving the ball forward – this is a good example
- The outcome regression part of DR goes back to Heckman, et al. (1997) and I use this in section 5.3.2 (“Bias correction”) of the mixtape
- The propensity score part goes back to Abadie (2005) which we’ve discussed
- New work on machine learning fits into this

Organization

- Basic assumptions for DD with covariates
- TWFE assumptions for DD with covariates
- Estimation alternative to TWFE with covariates
- Efficiency and semiparametric bounds

Insurance

- We covered covariates with Abadie (2005); why again?
- Maybe you're unsure whether the propensity score was properly specified
- How about some insurance?
- Two strikes instead of one

ATT

- DD *always* estimates the ATT because it's only the treatment effect for the treatment group in the post-treatment period
- It is not the ATE, or the LATE

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

Basic assumptions of DD

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

Handling repeated cross-sectional data is hairy, and so I've chosen to focus on the panel data for this talk, but results are similar for repeated cross sections

Basic assumptions of DD

Assumption 2: Conditional parallel trends

If you were putting covariates into your DD regression, then you were assuming conditional parallel trends

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

Basic assumptions of DD

Assumption 3: Common support or overlap

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 .

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997
- Or we can use the propensity score approach of Abadie (2005)
- What about TWFE? Hold off on that question for a second until we look at the estimators based on Assumptions 1-3

Outcome regression

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

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

where \bar{Y} is the sample average of Y among units in the treatment group at time t and $\hat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$. See my Section 5.3.2 for more about this.

Inverse probability weighting

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

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

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

Caveat

- Outcome regression needs $\hat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\hat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified
- Well why don't we just use TWFE? I've never heard anyone complain about including covariates in TWFE and I've been doing it my entire adult life, so we're good right?
- Depends on if you want to assume three more things. (Mixtape didn't know about this...)

TWFE

Here's the 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 the next three assumptions (let's first look at estimators based on .

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

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

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

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

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

Last line from the switching equation. This gives us:

Collecting terms

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

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

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

$$= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X)$$

$$= \delta + (\theta_1 X - \theta_2 X)$$

By allowing for the possibility that $\theta_1 X \neq \theta_2 X$, we open up the possibility of bias from TWFE which is zero under three additional assumptions.

Assumption 4

The implications of that TWFE regression with assumptions 1-3 gave us those previous expressions which then require placing further restrictions on treatment effects and trends when estimating with TWFE.

TWFE Assumption 4: Homogenous treatment effects in X

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

This is because when you difference out those previous equations, you need θX to cancel to leave you with δ which implies homogeneity in X .

X-specific trends

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

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

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

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

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

X-specific trends

Now take DD:

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

Eliminating terms, we get:

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

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

Assumption 5 and 6

For $D = 0, 1$, we need “no X-specific trends in both groups”:

$$E[Y_1 - Y_0 | D = d, X] = E[Y_1 - Y_0 | D = d]$$

Intuition: Sant'Anna and Zhao (2020) say in footnote 4 “[this] follows from analogous arguments” which is the previous slides’ manipulation of terms. Key is to remember these are time-varying covariates so they don’t cancel out within treatment category, so you need the trends in X to cancel out.

Without these six, in general TWFE will not identify ATT. Unclear how off it’ll be, but it will be biased is the point.

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. That's the basic idea. Gives you two chances to be wrong

Next step

- Introduction to the three prior covariate DD models
- Assumptions – check
- Hints about combining OR and IPW
- Now we move into *estimation* phase
- Let's see what doubly robust estimator looks like
- As before, I'm going to only stick to the panel data expressions bc all repeated cross-section does is add in some terms

Estimation

Some terms

$p(x)$: 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 $\mu_{1,\Delta}$ is just the treatment group's change in average Y for each $X = x$

We're off to see the (DR) wizard!

Population DR DD model for panel data

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

Notice the propensity score modifying the control group (second term inside parentheses) *and* the $\mu(X)$ term modifying the long difference. This is the idea of the doubly robust – you only need one of these models to be correctly specified, not both.

Sidebar: This is also one of the options in the Callaway and Sant'Anna (2020) DD estimator. It lets you pick IPW, regression (OR) or DR. Pedro usually recommends DR because of its advantages.

Efficiency

- Last step is inference
- 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

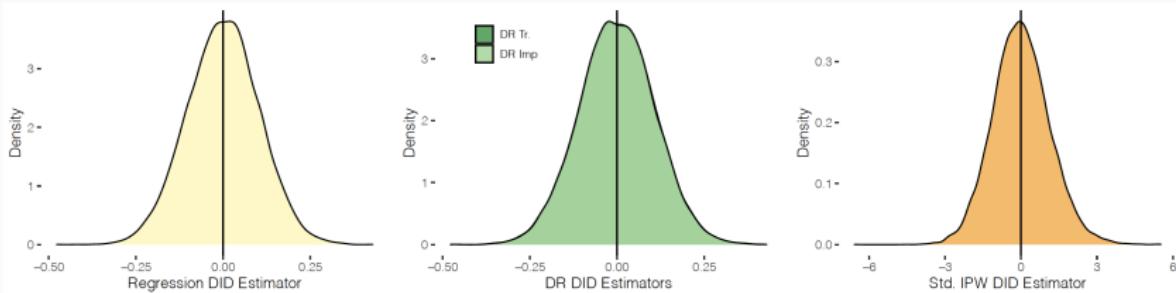
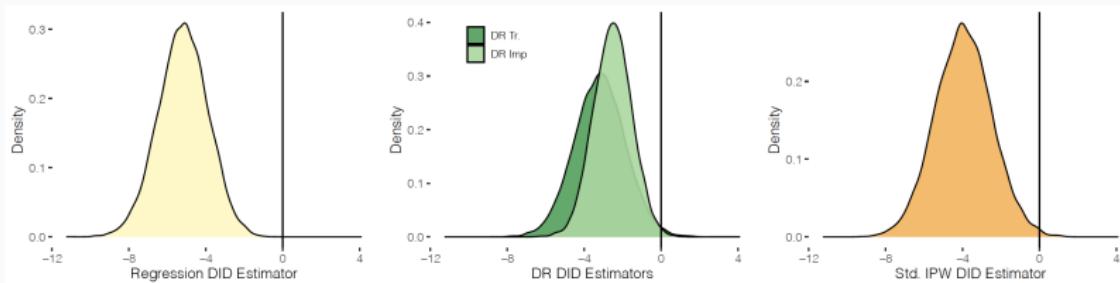


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	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



To a kid with a hammer, everything is a nail

- Use the right tool (oven) for the job (making lasagna), not the same tool (hammer) regardless of the job (making lasagna)
- One of the main things I learned from this paper was again biases in TWFE with covariates – Mixtape and MHE don't cover this
- This method only needed three assumptions not the six for TWFE
- Like everything Pedro does, there is code for this but it's only in R – DRDID
- But it's one of the main options in Callaway and Sant'anna under differential timing, and therefore it's crucial we understand this
- But you still have to have specified correctly either at least the outcome model or propensity score model



I ❤️ federalism
(for the natural experiments)

Tweets
30.4K

Following
5,933

Followers
11.8K

Likes
80.5K

Lists
1

Moments
0

[Edit profile](#)

Differential timing

- We've been considering situations where treatment occurs in one area for the most part – the two group case
- But the modal situation is when there is *differential timing* – groups treated at different time periods
- This happens in America usually because each area (state, municipality) will adopt a policy whenever they want to, which creates tendencies for roll out to occur
- Turns out, this is actually problematic for TWFE.
- Remember: TWFE \neq DD. Not all estimators are made the same.

Strict exogeneity

- Two main identification assumptions for TWFE
 - 1. Strict exogeneity

$$E[\varepsilon_{it}|X_{i1}, X_{i2}, \dots, X_{iT}, c_i] = 0$$

- 2. Full rank (regressors vary over time for at least some i)
- This is violated with differential timing and heterogeneity, but let's see why

Regression equation

We define the ATT as

$$\beta_{gp} = E[(Y_{gpit}^1) - E(Y_{gipt}^0)|g, p]$$

which varies by group and period (i.e., differential timing)

TWFE specification

$$Y_{gpit} = \beta_{gp}X_{gp} + \lambda_g + \lambda_p + \varepsilon_{gp}$$

Violation of strict exogeneity

$$E[Y_{gpit}|g, p, X_{gp}] = E[\beta_{gp}|X_{gp} = 1]X_{gp} + \lambda_g + \lambda_p + \mu$$

The first term is the overall ATT. The last term is the composite error term where μ equals $(\beta_{gp} - E[\beta_{gp}|X_{gp} = 1]X_{gp})$. It is not necessarily mean-zero condition on g, p and the group and time varying X .

TWFE will only identify this “overall average ATT” if all groups have the same ATT or only one treatment group (i.e., no differential timing).

This implies strict exogeneity is violated with heterogeneity and differential timing because the composite error term is correlated with treatment and group fixed effects

Differential timing

- In summary, differential timing was thought to be a simple extension of two group case, so people used TWFE to estimate ATT
- But it turns out TWFE is misspecified under differential timing with heterogeneity
- Event studies are also flawed for similar reasons as static parameter
- Here we will discuss recent work in econometrics
- Let's learn a paper that I will use for illustration

Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine



Cheng Cheng

Mark Hoekstra

Abstract

From 2000 to 2010, more than 20 states passed so-called "Castle Doctrine" or "stand your ground" laws. These laws expand the legal justification for the use of lethal force in self-defense, thereby lowering the expected cost of using lethal force and increasing the expected cost of committing violent crime. This paper exploits the within-state variation in self-defense law to examine their effect on homicides and violent crime. Results indicate the laws do not deter burglary, robbery, or aggravated assault. In contrast, they lead to a statistically significant 8 percent net increase in the number of reported murders and nonnegligent manslaughters.

Summary

- Cheng and Hoekstra (2013) are interested in whether expansions to “castle doctrine statutes” at the state level increase or decrease gun violence.
- Prior to these expansions, English common law principle required “duty to retreat” before using lethal force against an assailant except when the assailant is an intruder in the home
 - The home is one’s “castle” – hence, “castle doctrine”
 - When intruders threatened the victim in the home, the duty to retreat was waived and lethal force in self-defense was allowed

Castle doctrine law explained

- In 2005, Florida passed a law that expanded self-defense protections beyond the house
 - 2000 to 2010, 21 states explicitly put “castle doctrine” into statute, and (more importantly) extended it to places outside the home
 - In other words, 21 states removed the duty to retreat in specified circumstances
- Other changes:
 - Presumption of reasonable fear is added
 - Civil liability for those acting under the law is removed

Economic theory predicts more lethal homicides

- Workers supply legal or illegal labor and are therefore responsive to costs and benefits
- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- If people are rational, then lowering the price of lethal self-defense should increase lethal homicides

Economic theory also predicts less crime from deterrence

- Although deterrence is a theoretical possibility, note that the goal of the laws was to protect enhance victim rights, not deter crime
- Testable prediction with data and same design

Treatment passage

- Summary:
 - 21 states passed laws removing “duty to retreat” in places outside the home
 - 17 states removed “duty to retreat” in any place one had a legal right to be
 - 13 states include a presumption of reasonable fear
 - 18 states remove civil liability when force was justified under law

Cheng and Hoekstra's identification strategy

- Panel fixed effects estimation

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3(CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes “region-by-year fixed effects”

Data

- FBI Uniform Crime Reports Part 1 Offenses (2000-2010)
 - State-level crime rates, or “offenses per 100,000 population”
 - Falsification outcomes: motor vehicle theft and larceny
- Dataset on justifiable homicides by private citizens

Outcomes (in order)

- Deterrence and homicide outcomes:
 1. Burglary: the unlawful entry of a structure to commit a felony or a theft
 2. Robbery: the taking or attempting to take anything of value from the care, custody or control of a person or persons by force or threat of force or violence and/or putting the victim in fear
 3. Aggravated assault: unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury
- Homicide categories
 1. Total homicides – murder plus non-negligent manslaughter (~14,000 per year)
 2. Justifiable homicides by private citizens (~250/year)

Inference: Clustering

- Statistical inference: cluster standard errors at the state level
 - Are disturbances random draws from individually identical distribution?
 - It's likely that within a state, unobserved determinants of crime are serially correlated
 - They follow Bertrand, Duflo and Mullainathan (2004) and adjust for serial correlation in unobserved disturbances within states at the level of the treatment

Inference: Fisher's sharp null

- How likely is it that we estimate effects of this magnitude when using randomly chosen pre-treatment time periods and randomly assigning placebo treatments?
- Randomizes dates within-state for the pre-treatment period (<2000)
- Randomization inference and exact p-values

Region-by-year fixed effects

- Absent passing castle doctrine laws, outcomes in these 21 states would have changed similar to other states in their same region
 - Recall the “region-by-year fixed effects” in the X term
 - By including “region-by-year fixed effects”, they are arguing that unobserved changes in crime are running “parallel” to the treatment states within region over time
 - Need not hold across regions since the across region variation is not being used in this analysis due to the saturation of the model with “region-by-year fixed effects”

State specific time trends

- Alabama, et al. dummy interacted with TREND which equals 1 in 2000, 2 in 2001, ..., 11 in 2010
- Forces the identification to come from variation in outcomes around the state-specific linear trend
 - Outcomes must be large enough and different enough from a state-specific linear trend otherwise it is collinear with the state-trend
 - Same argument applies to any control though
 - Goodman-Bacon (2019) suggests group-trends are less taxing and satisfying than unit-specific trends

Control variables

- Controls (X matrix in earlier equation)
 - Full-time police employment per 100,000 state residents from the LEKOA data (FBI data)
 - Persons incarcerated in state prison per 100,000 residents
 - Shares of white/black men in 15-24 and 25-44 age groups
 - State per capita spending on public assistance
 - State per capita spending on public welfare

Parallel Leads

- Look at each set of treatment states against never-treated figure by figure (rare)
- Use a one-period lead in the regression model (not as common)
- I'm going to look at event study coefficients (most common)

Step one: Falsification test

- Policy-makers are not just randomly flipping coins when passing laws, but presumably do so because of things they observe on the ground
- Address concerns up front this isn't driven by spurious crime results
- Cheng and Hoekstra (2013) present falsification of larceny and motor vehicle theft first, then results

Step one (cont.)

- Results will be presented separately under six different specifications
 - Each new specification adds more controls
- Pop quiz: What should you expect to find on key variables of interest when conducting a falsification and why?

Answer

- No statistically significant association between the CDL passage and the placebos; preferably precise zeroes
- No association on the one-year lead either
- Basically, you should not find effects where there are no theoretical policy effects; gun laws shouldn't affect non-violent offenses

Step one (cont.)

- How do you interpret coefficients?
 - His model is “log outcomes” regressed onto a dummy variable (level), so these are semi-elasticities and approximate percentage changes – but you should transform them by taking the exponential of each coefficient and then differencing it from one to find the actual percentage change
 - Ex: CDL = -0.0137 (column 12, Table 3, “Log (larceny rate)” outcome.) $\text{Exp}(-0.0137) = 0.986$, and so $1-0.986 = 1.4$. Thus, CDL reduced larceny rates by 1.4 percent, which is not statistically significant.

Results – Falsification Exercise

Table 3: Placebo Tests

	OLS - Unweighted					
	7	8	9	10	11	12
Panel A: Larceny	Log (Larceny Rate)					
Castle Doctrine Law	0.00745 (0.0227)	0.00145 (0.0205)	-0.00188 (0.0210)	-0.00445 (0.0226)	-0.00361 (0.0201)	-0.0137 (0.0228)
One Year Before Adoption of Castle Doctrine Law				-0.0103 (0.0114)		
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft	Log (Motor Vehicle Theft Rate)					
Castle Doctrine Law	0.0767* (0.0413)	0.0138 (0.0444)	0.00814 (0.0407)	0.00775 (0.0462)	0.00977 (0.0391)	-0.00373 (0.0361)
One Year Before Adoption of Castle Doctrine Law				-0.00155 (0.0287)		
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft					Yes	
State-Specific Linear Time Trends						Yes

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics.

Step two: testing the deterrence hypothesis

- Having found no effect on their placebos, Cheng and Hoekstra (2013) examine the effect of CDL on three deterrence outcomes: burglary, robbery and aggravated assault
 - They will, again, have six specifications per outcome in the “weighted” regression, and then another five for the “unweighted” regression
- Pop quiz: What does deterrence look like?

Answer

- Negative signs on the CDL variable is consistent with deterrence – these crimes were “deterred”, in other words
- Based on early work by Becker (1968) and 1970s work by his student Isaac Ehrlich; higher probabilities of getting hurt in public may cause offenders to avoid violence in public altogether
- Bounds on the magnitudes from the standard errors are used to provide some confidence about the estimates as well

Results – Deterrence

	OLS - Weighted by State Population						OLS - Unweighted					
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Burglary												
	Log (Burglary Rate)						Log (Burglary Rate)					
Castle Doctrine Law	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237	0.0572**	0.00961	0.00663	0.00277	0.00683	0.0207
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)	(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)	(0.0259)
One Year Before Adoption of					-0.0201							
Castle Doctrine Law					(0.0139)							
Panel B: Robbery												
	Log (Robbery Rate)						Log (Robbery Rate)					
Castle Doctrine Law	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*	0.0448	0.0320	0.00839	0.00552	0.00874	0.0267
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)	(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)	(0.0299)
One Year Before Adoption of					-0.0156							
Castle Doctrine Law					(0.0167)							
Panel C: Aggravated Assault												
	Log (Aggravated Assault Rate)						Log (Aggravated Assault Rate)					
Castle Doctrine Law	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414	0.0555	0.0698	0.0343	0.0305	0.0341	0.0317
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)	(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)	(0.0380)
One Year Before Adoption of					-0.00343							
Castle Doctrine Law					(0.0161)							
Observations	550	550	550	550	550	550	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes					Yes		
State-Specific Linear Time Trends						Yes					Yes	

Conclusion

- “In short, these estimates provide strong evidence against the possibility that castle doctrine laws cause economically meaningful deterrence effects” (p. 17)
 - Translation: They can't find evidence of large deterrence effects
- “Thus, while castle doctrine law may well have benefits to those legally justified in protecting themselves in self-defense, there is no evidence that the law provides positive spillovers by deterring crime more generally” (p. 17)
 - They note in footnote 24 that they cannot measure the benefits to victims whose crimes were deterred, or the benefits from lower legal costs; their focus is limited to whether it deterred the crimes, not whether the net benefits from the laws were positive
 - Obviously, if there is no deterrence, though, then the net benefits are lower from CDL than they would be if they did deter

Step 3: Homicides

- The key finding in this study focuses on CDL and its effect on homicides and non-negligent manslaughter
- Pop quiz: what should the sign on CDL be here?

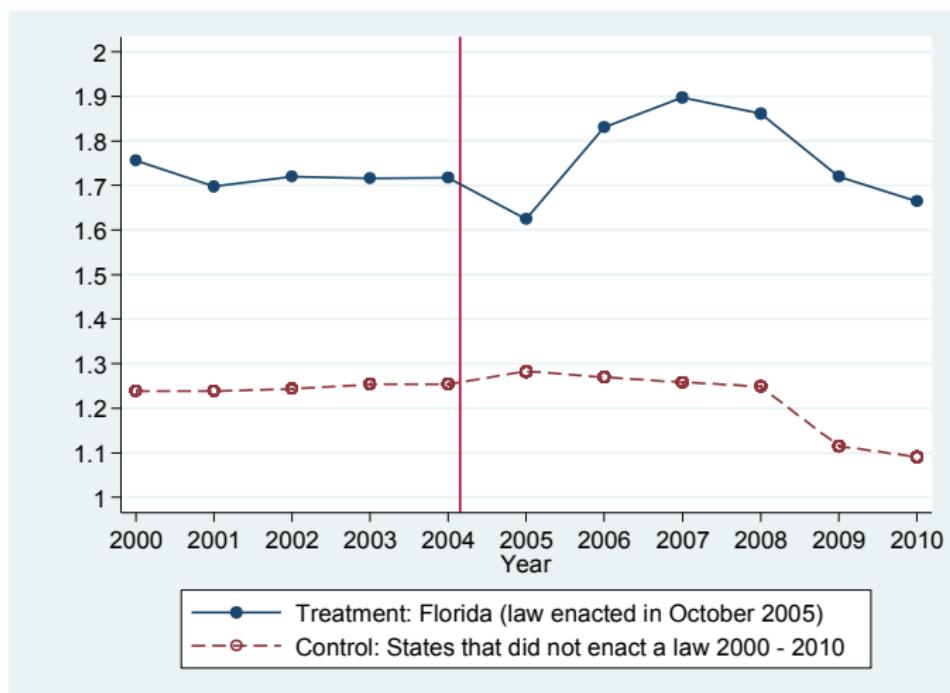
Answer

- Effects should be **positive**
- Cheng and Hoekstra want to show the raw data, but have differential timing
- Differential timing means you can't show pre-treatment raw data for the never-treated groups
- So they show it one by one – which isn't the most aesthetically pleasing way to do it, but which has the benefit of being transparent

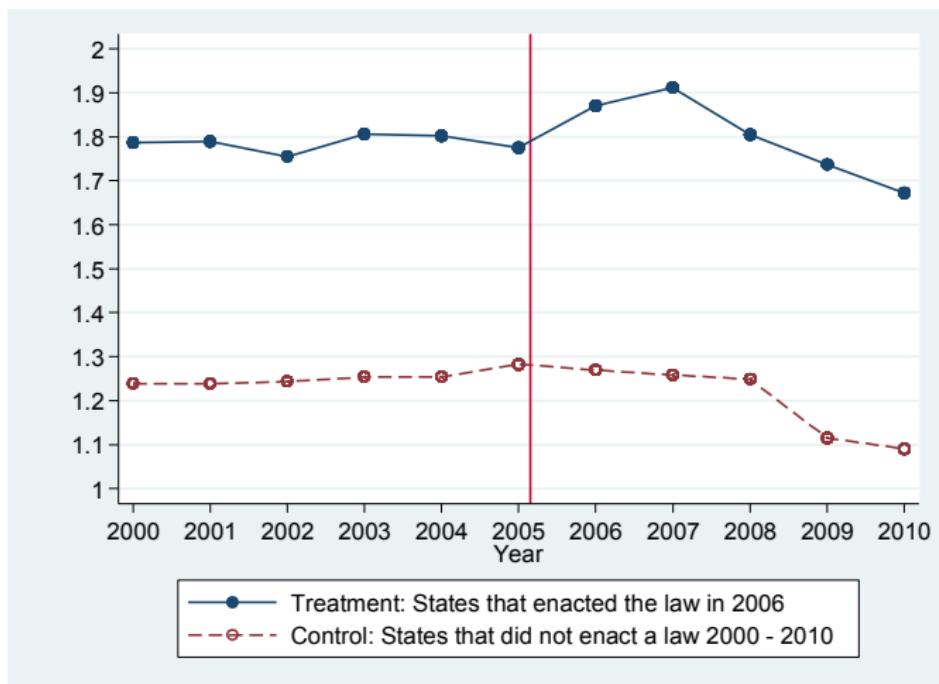
Parallel pre-treatment trends

- Keep your eyes on whether pre-treatment trends are parallel for treatment and control groups
 - Remember, though – he needs parallel trends within-region – these figures don't show that
 - But starting with pictures and raw data has value

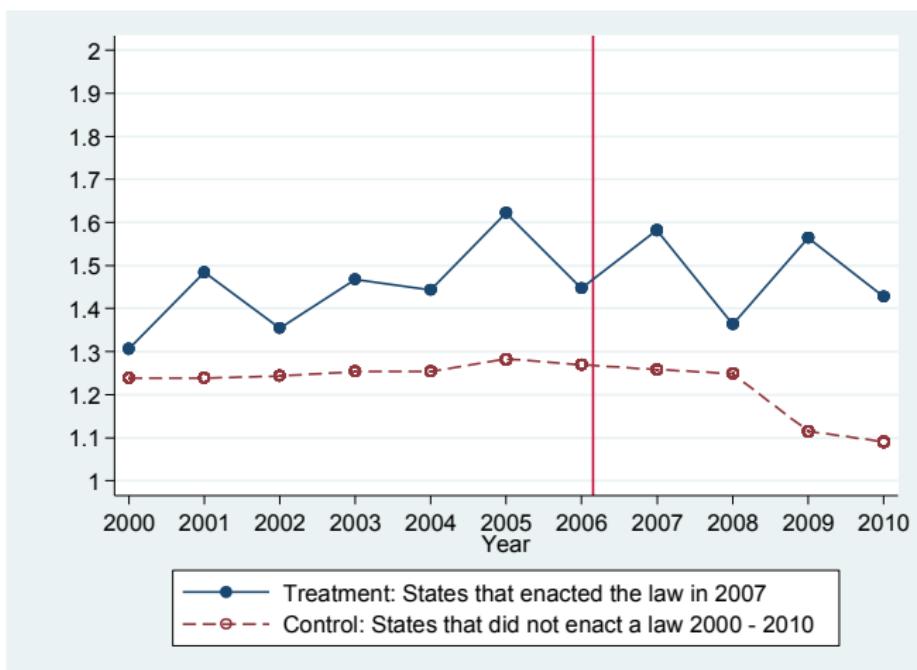
Log Homicide Rates – 2005 Adopter = Florida



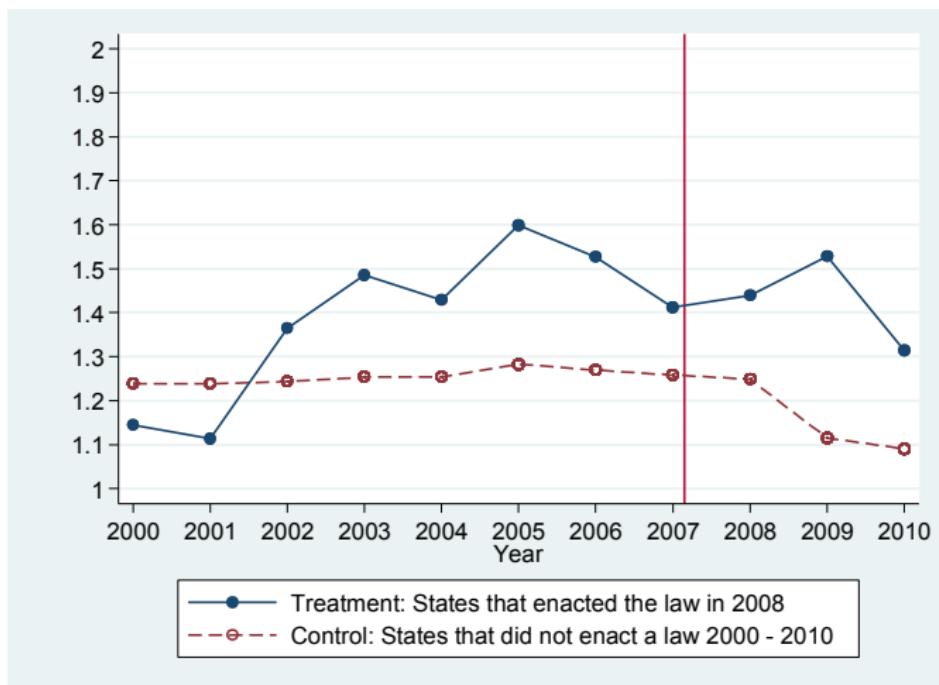
Log Homicide Rates – 2006 Adopter (13 states)



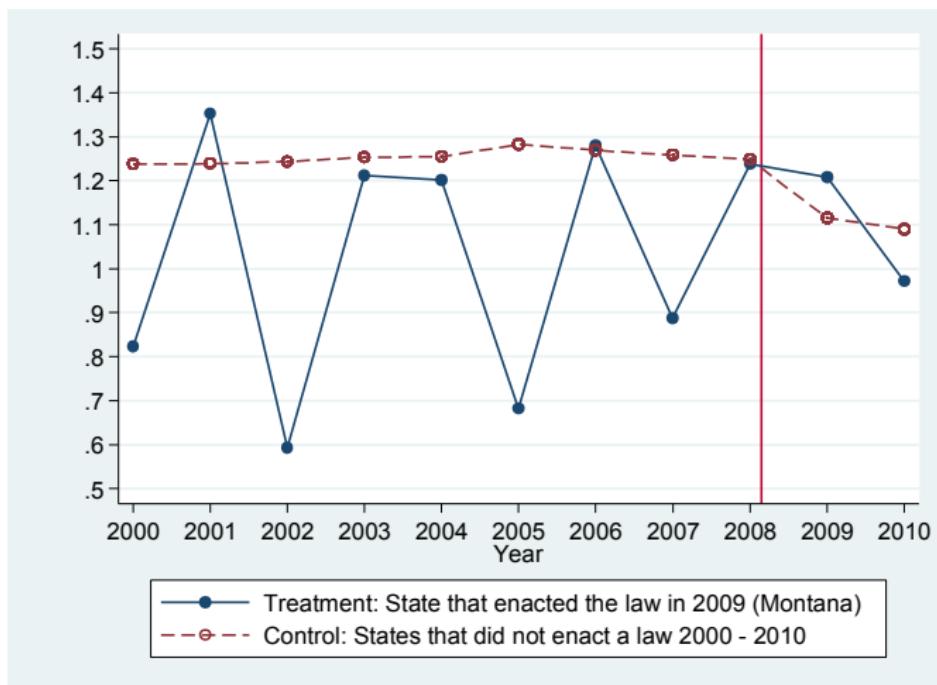
Log Homicide Rates – 2007 Adopter (4 states)



Log Homicide Rates – 2008 Adopter (2 states)



Log Homicide Rates – 2009 Adopter = Montana



Modeling

- He uses a class of estimators more appropriate for “counts” called “count models”, like the negative binomial estimated with maximum likelihood
- Results are robust to least squares and count models

Homicide – Negative Binomial; Murder – OLS

	1	2	3	4	5	6
<u>Panel C: Homicide (Negative Binomial - Unweighted)</u>						
Castle Doctrine Law	0.0565*	0.0734**	0.0879***	0.0783**	0.0937***	0.108***
	(0.0331)	(0.0305)	(0.0313)	(0.0355)	(0.0302)	(0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352		
				(0.0260)		
Observations	550	550	550	550	550	550
<u>Panel D: Log Murder Rate (OLS - Weighted)</u>						
Castle Doctrine Law	0.0906**	0.0955**	0.0916**	0.0884**	0.0981**	0.0813
	(0.0424)	(0.0389)	(0.0382)	(0.0404)	(0.0391)	(0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110		
				(0.0230)		
Observations	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Fisher sharp null

Move the 11-year panel back one year at a time (covering 1960-2009) and estimate 40 placebo “effects” of passing CDL 1 to 40 years earlier

Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	-0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

My replication using event study plots

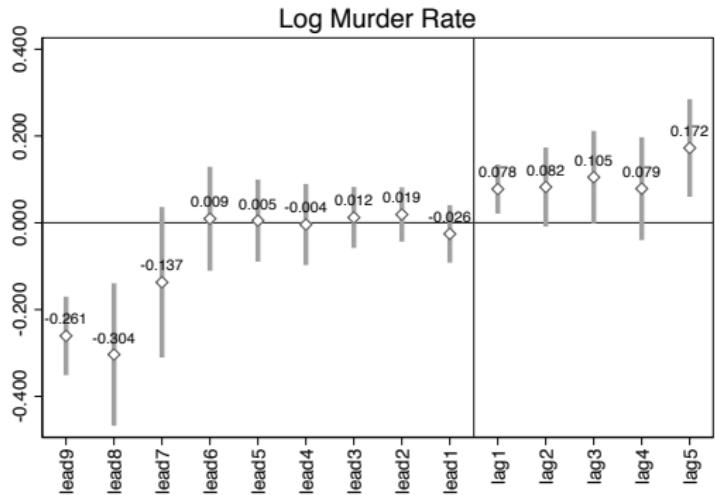


Figure: Homicide event study plots using coefplot

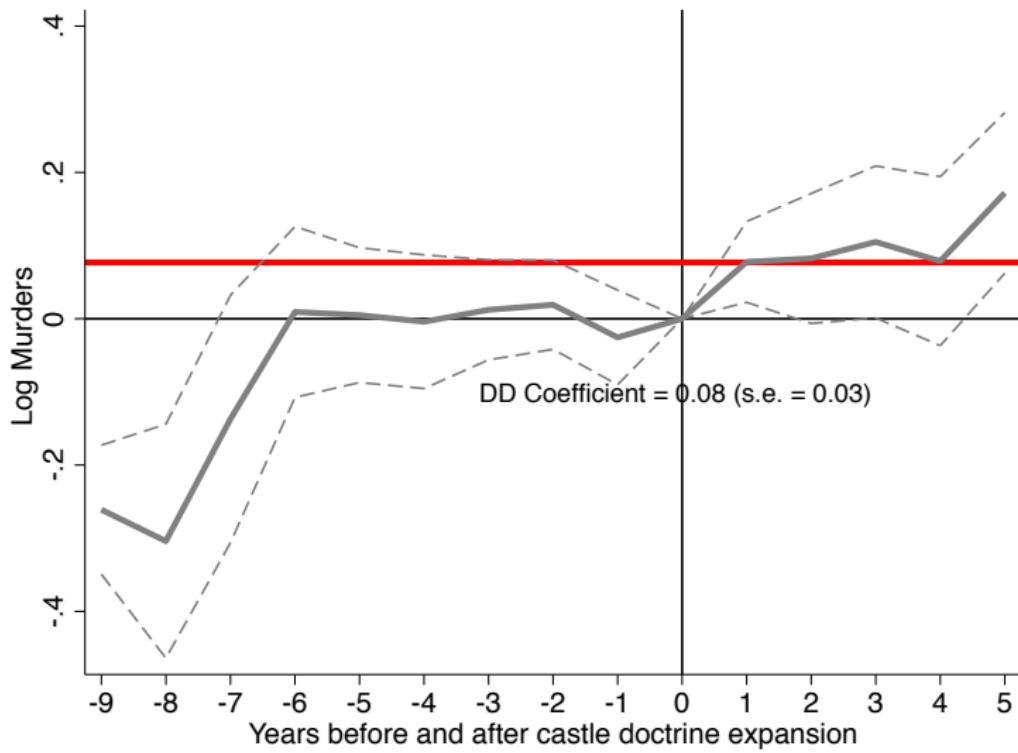


Figure: Homicide event study plots using two way

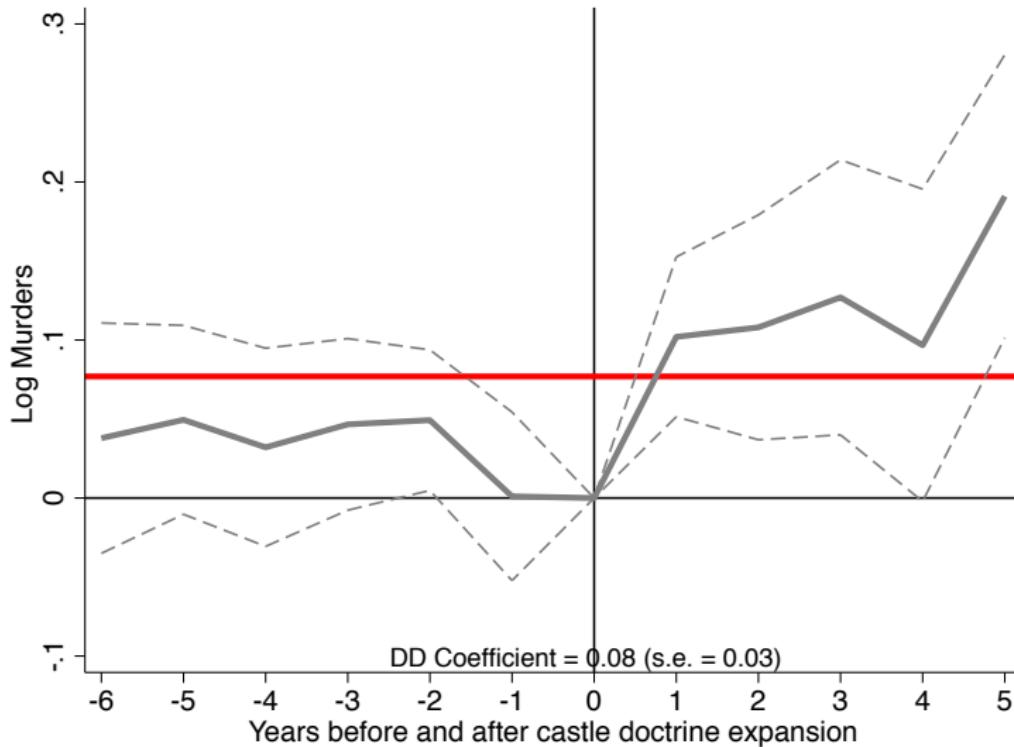


Figure: Homicide event study plots using two way and force early leads into one coefficient

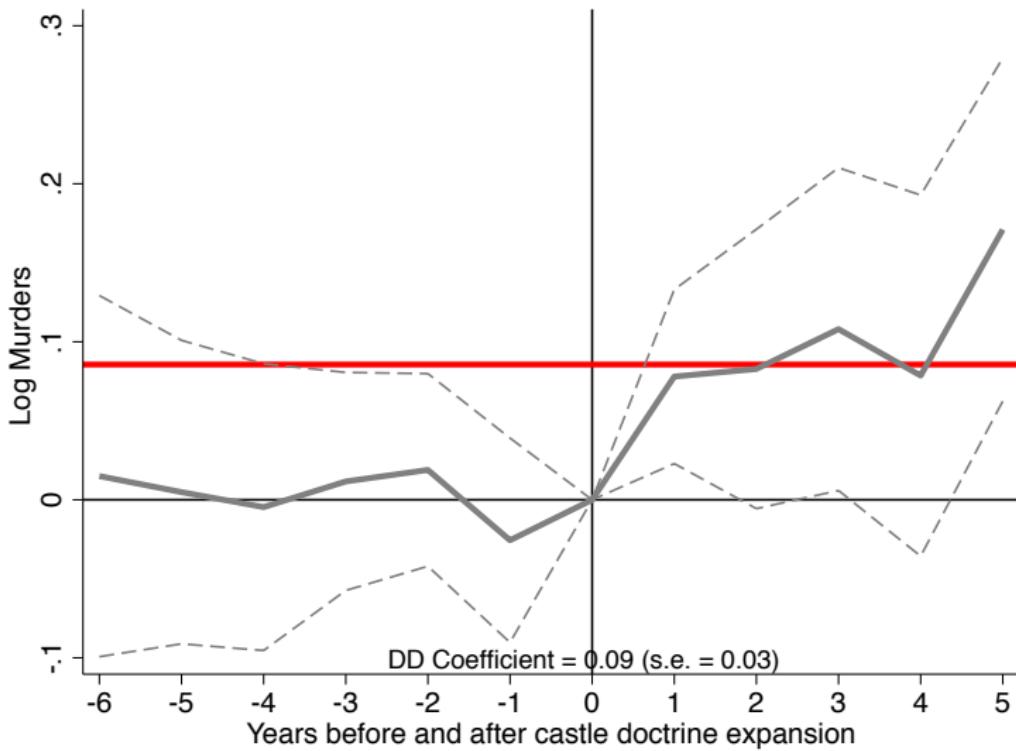


Figure: Homicide event study plots using two way dropping imbalanced states

Interpretation

- No evidence that Castle Doctrine/Stand Your Ground Laws deter violent crimes such as burglary, robbery and aggravated assault
- These laws do lead to an 8% net increase in homicide rates, translating to around 600 additional homicides *per year* across the 21 adopting states
- Unlikely that all of the additional homicides were legally justified
- Incentives matter in some contexts (lethal force) but not others (deterrence)

Where to from here?

- Now that we've reviewed the two-way fixed effects with treatment that differed across time, how does this more general form of "differential timing" compare with the 2x2 DD that we reviewed?
- Complicated derivation, but simple interpretation - two-way fixed effects with differential timing estimates a weighted average of all 2x2
- Andrew Goodman-Bacon (2018; 2019) and Callaway and Sant'ann (2019)
- I will be making the argument that under certain *modal* situations, the two-way fixed effects model has major problems, even fatal ones, due to biases even when parallel trends plausibly holds

Difference-in-Differences with Variation in Treatment Timing

Andrew Goodman-Bacon

NBER Working Paper No. 25018

Issued in September 2018

NBER Program(s): Children, Development of the American Economy, Labor Studies, Public Economics

The canonical difference-in-differences (DD) model contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper derives an expression for this general DD estimator, and shows that it is a weighted average of all possible two-group/two-period DD estimators in the data. This result provides detailed guidance about how to use regression DD in practice. I define the DD estimand and show how it averages treatment effect heterogeneity and that it is biased when effects change over time. I propose a new balance test derived from a unified definition of common trends. I show how to decompose the difference between two specifications, and I apply it to models that drop untreated units, weight, disaggregate time fixed effects, control for unit-specific time trends, or exploit a third difference.



Reminder of 2x2 DD

To understand differential timing, we need to remind ourselves 2x2 form

$$\widehat{\delta}_{kU}^{2x2} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

Post to pre difference for treatment group compared to the post to pre difference for *never treated*

Different treatment dates by panel unit

$$y_{it} = \underbrace{\beta D_i + \tau Post_t + \delta(D_i \times Post_t) + X_{it} + \alpha_i + \alpha_t + \varepsilon_{it}}_{2x2 DD}$$
$$y_{it} = \underbrace{\delta D_{it} + X_{it} + \alpha_i + \alpha_t + \epsilon_{it}}_{Two way FE}$$

We know a lot about 2x2, but about the twoway fixed effects estimator when it comes to DD designs

Decomposition Preview

- Linear panel models estimate a treatment parameter that is a weighted average over all 2×2 in your sample
- The estimator is a weighted average of all potential $\delta^{2 \times 2}$ in which treated units act as both controls and treatment depending on the situation
- Weights are function of sample sizes of each “group” and the variance of the treatment dummies for the groups

Decomposition (cont.)

- Under the assumptions of variance weighted common trends (VWCT) and time invariant treatment effects, the estimator called the variance weighted ATT is a weighted average of all possible ATTs
- Under more restrictive assumptions it perfectly matches the ATT
- Time varying treatment effects generate a bias that needs to be accounted for

3 Group Example

- Suppose two treatment groups (k,l) and one untreated group (u)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it later than l
- Denote \bar{D}_k as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{ab}^{2x2,j}$ as the canonical 2×2 DD estimator for groups a and b where j is the treatment group
- So what are the possible 2×2 combinations?

How many 2x2?

- A lot!
- When there's three groups - a never treated (U), an early treated (k) and a late treated (l), there are four 2x2s
- But typically, we have more than 3 groups making the number of potential 2x2 even larger
- With K timing groups and one untreated group, there are K^2 distinct 2x2 DDs

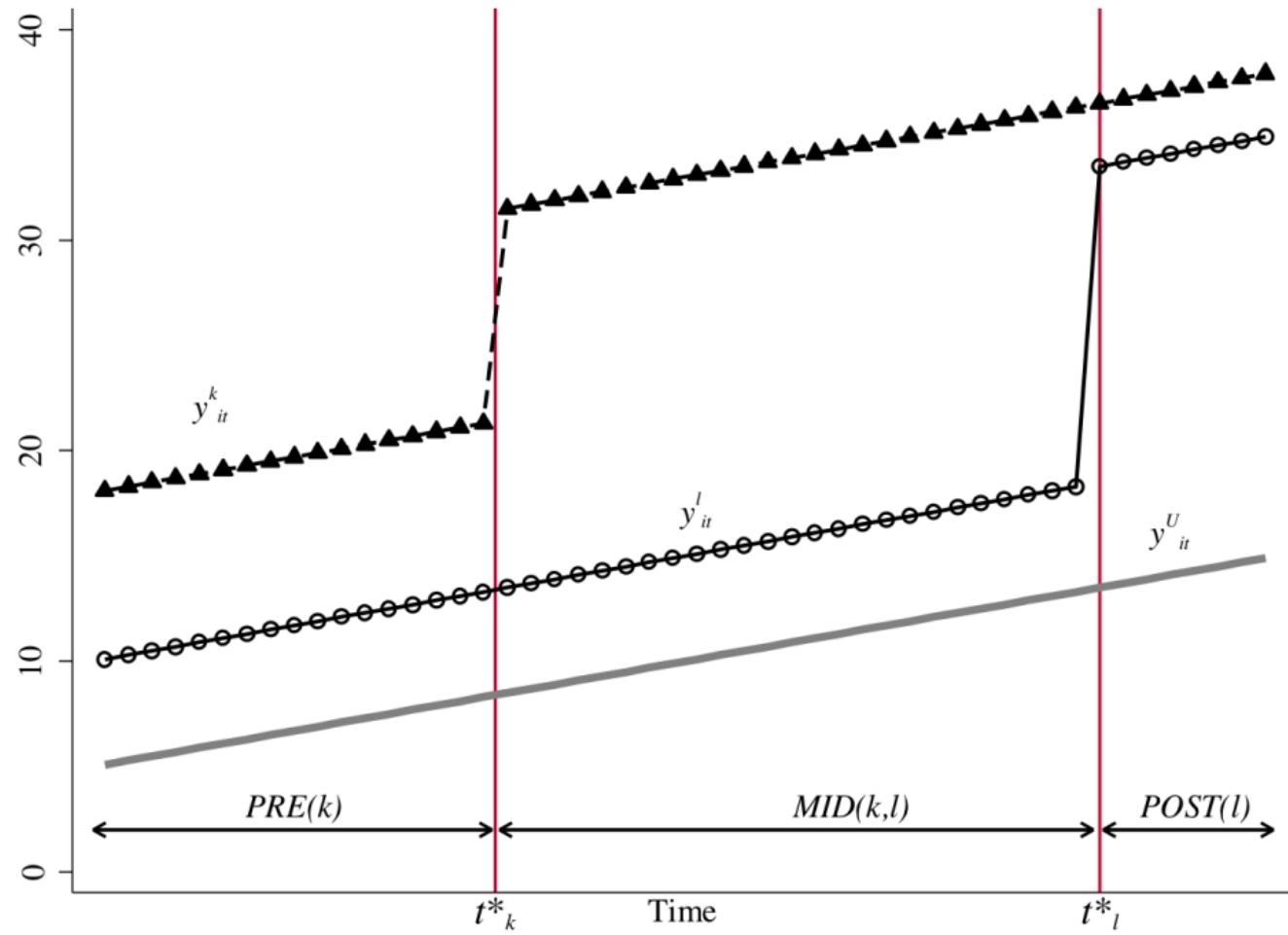
K^2 distinct DDs

Assume 3 timing groups (a, b and c) and one untreated group (U). Then there should be 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

Simple example with 3 groups

- We'll stick with two groups, k and l , who will get the treatment at t_k^* and t_l^* , and the third group U will never get treated
- The earlier period before anyone is treated is “pre”, the period between k and l treatment is “mid”, and the period after l is treated is “post”



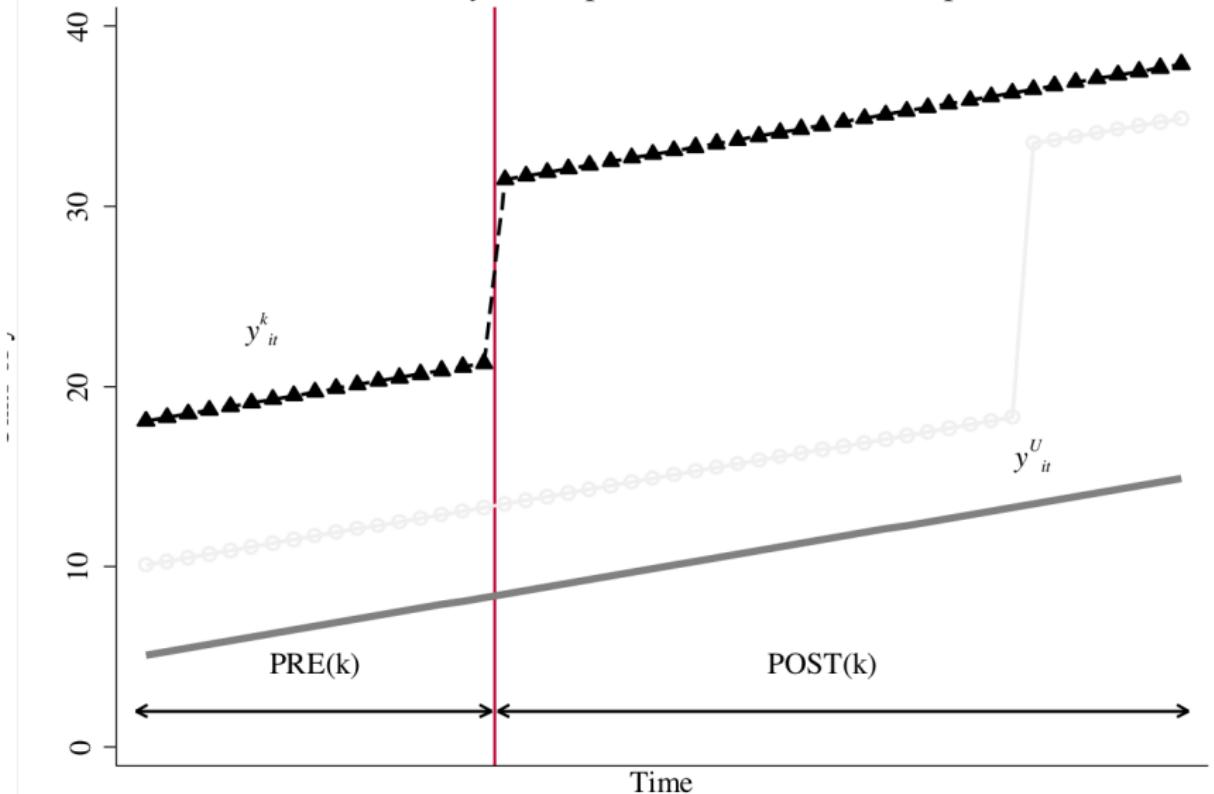
Three important 2x2 DDs

$$\begin{aligned}\widehat{\delta}_{kU}^{2x2} &= \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right) \\ \widehat{\delta}_{kl}^{2x2} &= \left(\bar{y}_k^{mid(k,l)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_l^{mid(k,l)} - \bar{y}_l^{pre(k)} \right) \\ \widehat{\delta}_{lk}^{2x2} &= \left(\bar{y}_l^{post(l)} - \bar{y}_l^{mid(k,l)} \right) - \left(\bar{y}_k^{post(l)} - \bar{y}_k^{mid(k,l)} \right)\end{aligned}$$

where the first 2x2 is any timing group compared to untreated, the second is a group compared to yet-to-be-treated timing group, and the last is the eventually-treated compared to the already-treated controls.

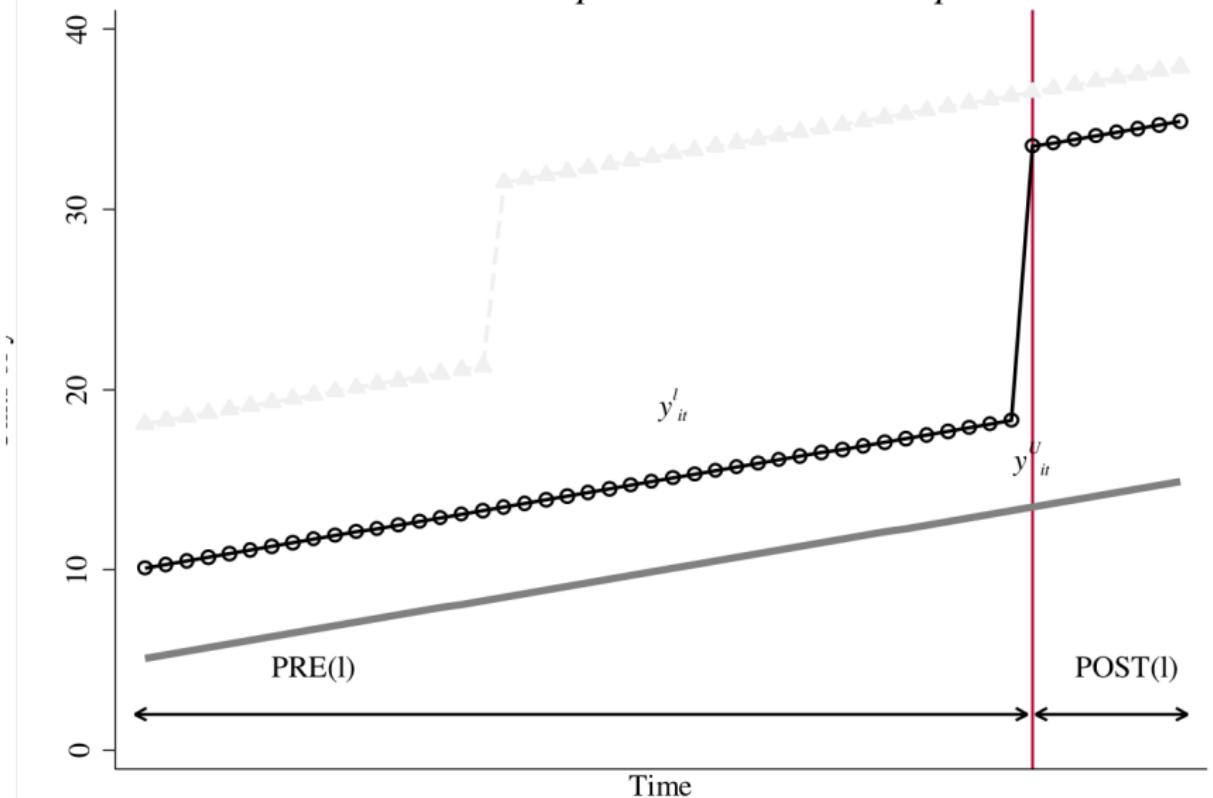
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$

A. Early Group vs. Untreated Group



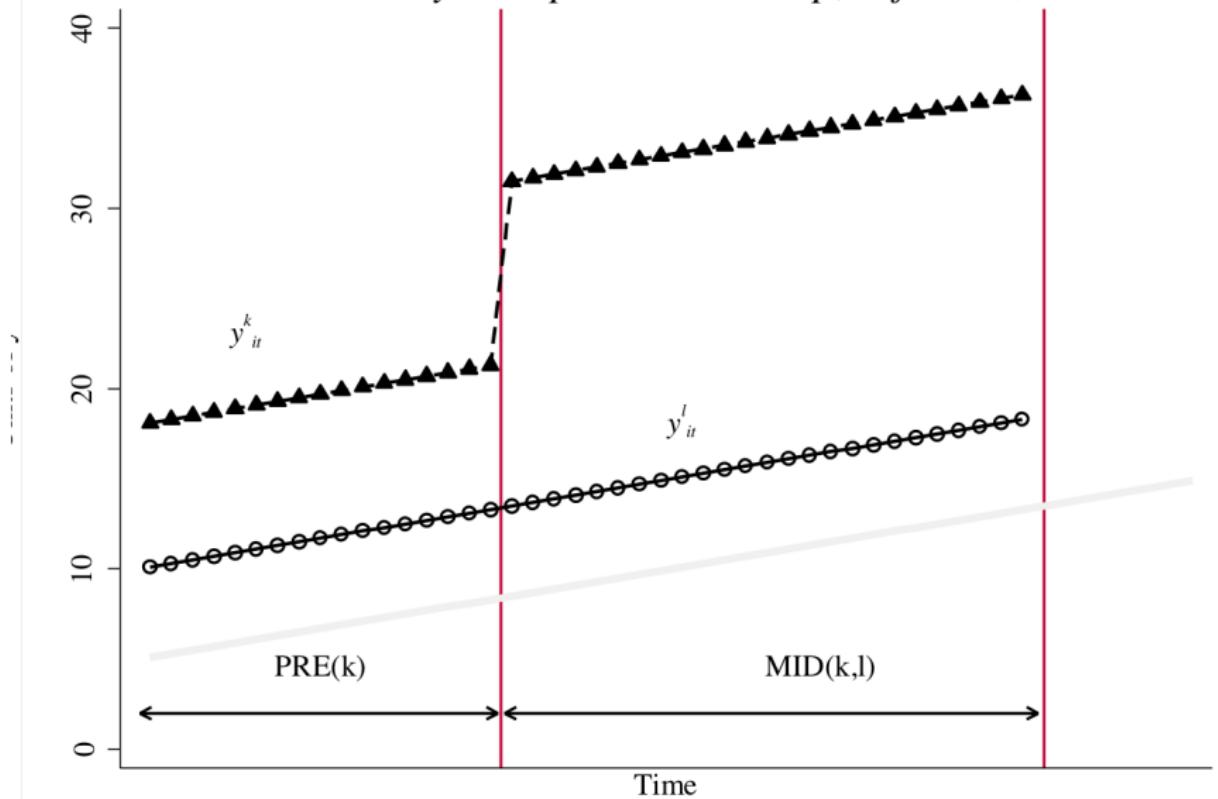
$$\widehat{\delta}_{lU}^{2x2} = \left(\bar{y}_l^{post(l)} - \bar{y}_l^{pre(l)} \right) - \left(\bar{y}_U^{post(l)} - \bar{y}_U^{pre(l)} \right)$$

B. Late Group vs. Untreated Group

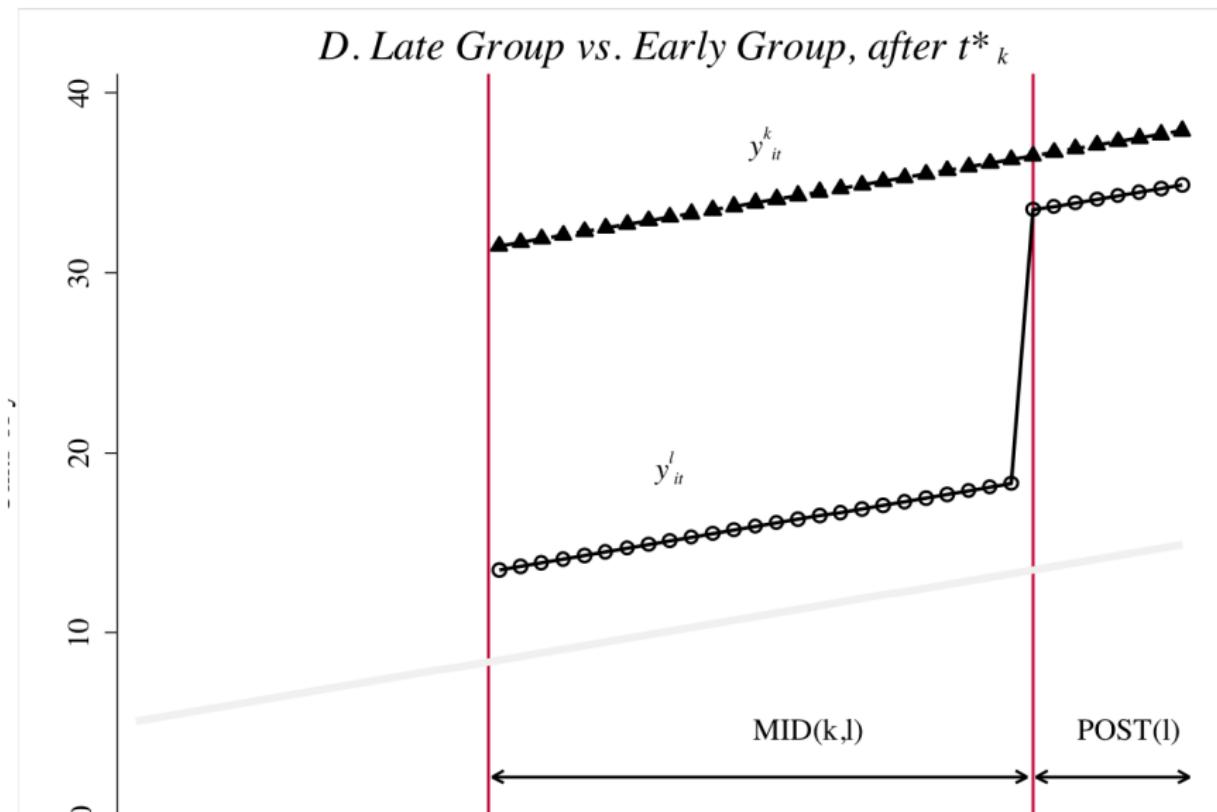


$$\delta_{kl}^{2x2,k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{Pre(k,l)} \right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k,l)} \right)$$

C. Early Group vs. Late Group, before t^*_l



$$\delta_{lk}^{2x2,l} = \left(\bar{y}_l^{POST(k,l)} - \bar{y}_l^{MID(k,l)} \right) - \left(\bar{y}_k^{POST(k,l)} - \bar{y}_k^{MID(k,l)} \right)$$



Second, what makes up the DD estimator?

The least squares estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \widehat{\delta}_{lk}^{2x2,l} \right]$$

where that first 2x2 is the k compared to U and the l compared to U
(combined to make the equation shorter)

Third, the Weights

$$\begin{aligned}s_{ku} &= \frac{n_k n_u \bar{D}_k (1 - \bar{D}_k)}{\widehat{Var}(\tilde{D}_{it})} \\ s_{kl} &= \frac{n_k n_l (\bar{D}_k - \bar{D}_l) (1 - (\bar{D}_k - \bar{D}_l))}{\widehat{Var}(\tilde{D}_{it})} \\ \mu_{kl} &= \frac{1 - \bar{D}_k}{1 - (\bar{D}_k - \bar{D}_l)}\end{aligned}$$

where n refer to sample sizes, $\bar{D}_k(1 - \bar{D}_k)$ ($\bar{D}_k - \bar{D}_l$) $(1 - (\bar{D}_k - \bar{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.

Weights discussion

- Two things pop out of these weights
 - “Group” variation matters more than unit-level variation. A group is if two states got treated in 1995. They are the 1995 group. More units in a group, the bigger that 2×2 is practically
 - Within-group *treatment* variance matters a lot.
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
 1. $\bar{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
 2. $\bar{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
 3. $\bar{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
- What's this mean? The weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the “treated on treated” weights? What’s this $\bar{D}_k - \bar{D}_l$ business about?
- Well, same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\bar{D}_k - \bar{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

TWFE and centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- But when looking at treated to treated comparisons, when differences in timing have a spacing of around 1/2, those also weight up the respective 2s2s via variance weighting
- But there's no theoretical reason why should prefer this as it's just a weighting procedure being determined by how we drew the panel
- This is the first thing about TWFE that should give us pause, as not all estimators do this

Potential outcomes

- Previous just showed that DD was based on a weighted “adding up” of particular 2x2s. That tells us what DD is numerically. But that’s not the end
- Because the decomposition theorem expresses the DD coefficient in terms of sample averages, the movement to potential outcomes is easy.
- Now we express DD in terms of ATT which is essential for understanding identification and bias

Average treatment effect on the treatment group (ATT)

- Define the year-specific ATT as

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 | k, t = \tau]$$

- Now define it over a time window W (e.g., a post-treatment window)

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 | k, \tau \in W]$$

- Define differences in average potential outcomes over time as:

$$\Delta Y_k^h(W_1, W_0) = E[Y_{it}^h | k, W_1] - E[Y_{it}^h | k, W_0]$$

for $h = 0$ (i.e., Y^0) or $h = 1$ (i.e., Y^1)

Changing potential outcomes

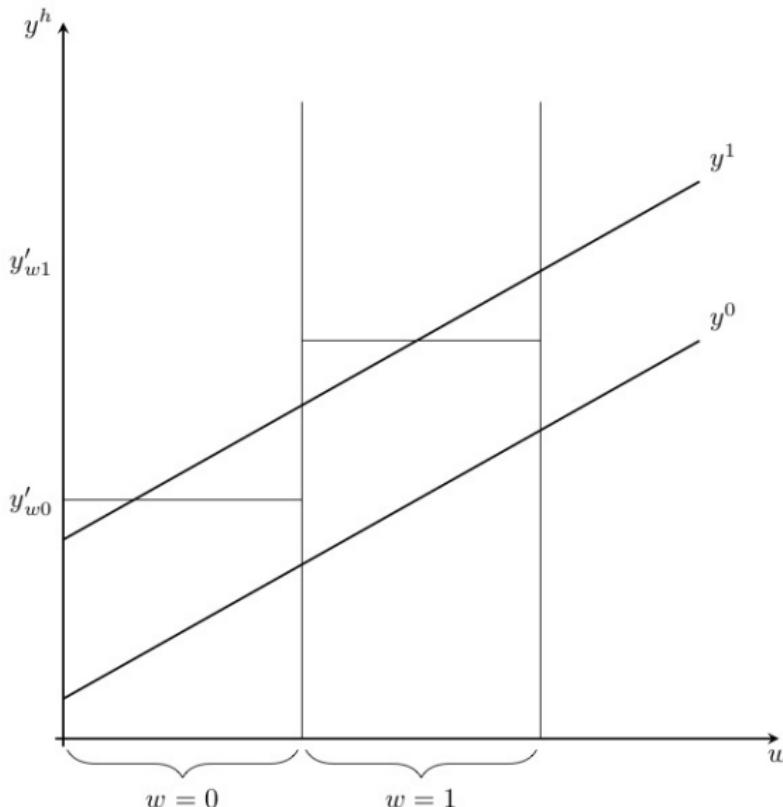


Figure: With trends, differences in mean potential outcomes is non-zero

From 2x2 to ATT

$$\begin{aligned}\hat{\delta}_{kU}^{2x2} &= \left(E[Y_j|Post] - E[Y_j|Pre] \right) - \left(E[Y_u|Post] - E[Y_u|Pre] \right) \\ &= \underbrace{\left(E[Y_j^1|Post] - E[Y_j^0|Pre] \right) - \left(E[Y_u^0|Post] - E[Y_u^0|Pre] \right)}_{\text{Switching equation}} \\ &\quad + \underbrace{E[Y_j^0|Post] - E[Y_j^0|Post]}_{\text{Adding zero}} \\ &= \underbrace{E[Y_j^1|Post] - E[Y_j^0|Post]}_{\text{ATT}} \\ &\quad + \underbrace{\left[E[Y_j^0|Post] - E[Y_j^0|Pre] \right] - \left[E[Y_u^0|Post] - E[Y_u^0|Pre] \right]}_{\text{Non-parallel trends bias in 2x2 case}}\end{aligned}$$

Potential outcomes

$$\widehat{\delta}_{kU}^{2x2} = ATT_{Post,j} + \underbrace{\Delta Y_{Post,Pre,j}^0 - \Delta Y_{Post,Pre,U}^0}_{\text{Selection bias!}}$$

Hah! It's that another selection bias term, like when we decomposed the simple difference in outcomes! But here we see it's basis - non-parallel trends in potential outcomes themselves. Notice one of these is counterfactuals, but which one?

Two benign 2x2

$$\begin{aligned}\widehat{\delta}_{kU}^{2x2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\ \widehat{\delta}_{kl}^{2x2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)\end{aligned}$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated yet).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions like we did we get:

$$\widehat{\delta}_{lk}^{2x2} = ATT_{l,Post(l)} + \underbrace{\Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

Heterogeneity bias?

That old decomposition of the simple difference in outcomes rears its ugly head!

$$\begin{aligned}\widehat{\delta}_{kl}^{2x2} &= ATT_{l,Post(l)} \\ &\quad + \Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

- The first part is the ATT we are looking for
- The selection bias which only zeroes out if Y^0 for k and l has the same parallel trends from mid to post period
- The heterogeneity bias (3) occurs if the ATT for k differs over time. If not, then it just zeroes out.

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \widehat{\delta}_{kl}^{2x2,l} \right]$$

where we will make these substitutions

$$\begin{aligned}\widehat{\delta}_{kU}^{2x2} &= ATT_k(Post) + \Delta Y_l^0(Post, Pre) - \Delta Y_U^0(Post, Pre) \\ \widehat{\delta}_{kl}^{2x2,k} &= ATT_k(Mid) + \Delta Y_l^0(Mid, Pre) - \Delta Y_l^0(Mid, Pre) \\ \widehat{\delta}_{lk}^{2x2,l} &= ATT_l Post(l) + \Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID) \\ &\quad - (ATT_k(Post) - ATT_k(Mid))\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

$$\begin{aligned} p\lim_{n \rightarrow \infty} \widehat{\delta}_{n \rightarrow \infty}^{DD} &= \delta^{DD} \\ &= VWATT + VWCT - \Delta ATT \end{aligned}$$

- Notice the number of assumptions needed even to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Let's look at each of these three parts more closely

Variance weighted ATT

$$\begin{aligned} VWATT &= \sum_{k \neq U} \sigma_{kU} ATT_k(Post(k)) \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} ATT_k(MID) + (1 - \mu_{kl}) ATT_l(POST(l)) \right] \end{aligned}$$

where σ is like s only population terms not samples.

- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant.
- But otherwise, it's basically weighting each of the individual sets of ATT we have been discussing, where weights depend on group size and variance

Variance weighted common trends

- VWCT can be understood as a variance weighted common trends component,
- This is the collection of selection biases we previously wrote out,
- But notice – identification requires *variance weighted* common trends to hold.
- You get this with identical trends, but you don't need identical trends anymore as the weights can make it hold without.
- Huge pain to write out, unfortunately.

Variance weighted common trends

$$\begin{aligned} VWCT &= \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0(Post(k), Pre) - \Delta Y_U^0(Post(k), Pre) \right] \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} \{ \Delta Y_k^0(Mid, Pre(k)) - \Delta Y_l^0(Mid, Pre(k)) \} \right. \\ &\quad \left. + (1 - \mu_{kl}) \{ \Delta Y_l^0(Post(l), Mid) - \Delta Y_k^0(Post(l), Mid) \} \right] \end{aligned}$$

This is new. But while this is a lot to be equalling zero, it's ironically a weaker identifying assumption than we thought bc you don't need identical common trends since the weights can technically correct for unequal trends.

Heterogeneity bias

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \left[ATT_k(Post(l)) - ATT_k(Mid) \right]$$

Now, if the ATT is constant over time, then this difference is zero, but what if the ATT is not constant? Then TWFE is biased, and depending on the dynamics and the VWATT, may even flip signs

Case 1: ATT varies across units but not time

$$p \lim_{n \rightarrow \infty} \widehat{\delta}_{n \rightarrow \infty}^{DD} = VWATT + VWCT$$

because $\Delta ATT = 0$ here. Assume VWCT=0. Then the VWATT equals

$$\begin{aligned} VWATT &= \sum_{k \neq U} ATT_k \left[\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}(1 - \mu_{jk}) + \sum_{j=k+1}^K \sigma_{jk}\mu_{jk} \right] \\ &= \sum_{k \neq U} ATT_k w_k^T \end{aligned}$$

the VWATT weights together group-specific ATTs by a function of sample shares and treatment variance.

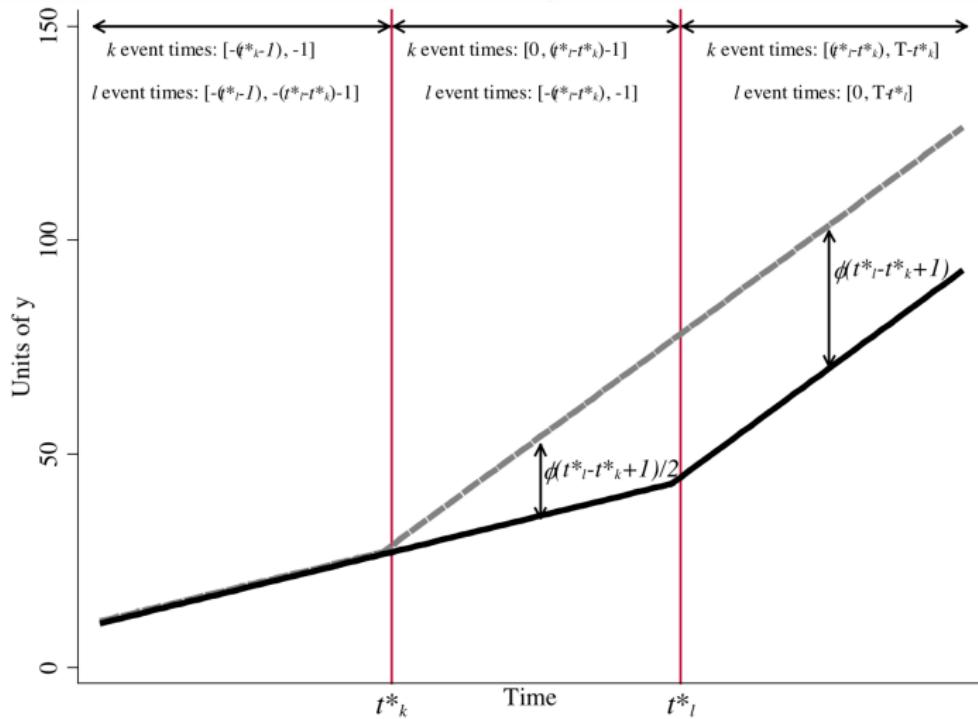
Case 1 cont.

- The processes that determine treatment timing are central to the interpretation of VWATT.
- Assume treatment rolls out first to units with the largest ATTs.
 - Then regression DD underestimates the sample-weighted ATT if t_1^* is early enough, or if there are a lot of post periods, so that \bar{D}_1 very small and $\bar{D}_k \approx 0.5$
 - Regression DD overestimates if t_1^* is late enough (or if there are a lot of pre periods) so that $\bar{D}_1 \approx 0.5$ and \bar{D}_k is small
- Goodman-Bacon (2018) suggests scattering the weights against each group's sample share. They may be close if there is little variation in treatment timing, if the untreated group is very large, or if some timing groups are very large

Case 2: Constant ATT across units, but heterogenous over time

- Time varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows
- Let's consider a case where the counterfactual outcomes are identical, but the treatment effect is a linear break in the trend. For instance, $Y_{it}^1 = Y_{it}^0 + \theta(t - t_1^* + 1)$ similar to Meer and West (2013)

Treatment effect is break in trend



Case 2 cont.

- The first 2x2 uses the later group as its control in the middle period. But in the late period, the later treated unit is using the earlier treated as its control
- But notice, this effect is biased because the control group is experiencing a trend in outcomes (heterogeneous treatment effects)
- This bias feeds through to the later 2x2 according to the size of the weight $(1 - \mu_{kl})$

Variance weighted common trends

- If treatment effects are constant over time, then we only need $VWCT = 0$ to identify VWATT. “Only”!
- The assumption itself is not testable because common trends is based on counterfactual Y^0 for the treatment groups in the post-treatment period, and we only have pre-treatment data
- But let’s assume differential counterfactual trends Y_k^0 are linear throughout the panel. Then we can get a convenient approximation to the $VWCT$ on the next slide

Variance weighted common trends

$$\begin{aligned} VWCT &= \sum_{k \neq U} \Delta Y_k^0 \left[\sigma_{kU} + \sum_{j=1}^{k-1} \sigma_{jk}(1 - 2\mu_{jk}) + \sum_{j=k+1}^K \sigma_{kj}(2\mu_{kj} - 1) \right] \\ &\quad - \Delta Y_U^0 \sum_{k \neq U} \sigma_{kU} \end{aligned}$$

Obviously, for this bias to be inconsequential, we need the sum of the two weighted counterfactual trends to be zero. You get this with identical trends, but those are not necessary due to the weights ability to shift non-identical trends so as to satisfy the zero condition.

Variance weighted common trends

The weight on each group's counterfactual trend equals the difference between the total weight it gets when it acts as a treatment group (w_k^T) minus the total weight it gets when it acts as a control (w_k^c).

$$\sum_k \Delta Y_k^0 [w_k^T - w_k^C] = 0$$

where w_k^T is the sum of all weights where group k is the treatment group

$$w_k^T = \sigma_{kU} + \sum_{k=1}^{k-1} \sigma_{jk}(1 - \mu_{jk}) + \sum_{j=k+1}^K \sigma_{kj}\mu_{kj}$$

and w_k^c is the sum of all weights where group k is the control group

$$w_k^c = \sum_{k=1}^{k-1} \sigma_{jk}\mu_{jk} + \sum_{j=k+1}^K \sigma_{jk}(1 - \mu_{jk})$$

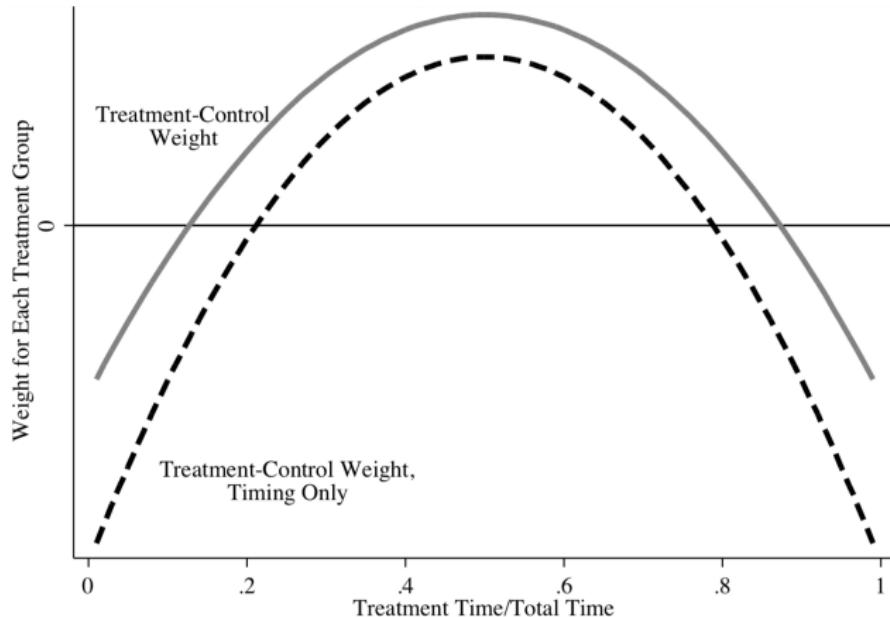
Variance weighted common trends

- The bias induced by each group will depend on whether it is a net treatment/control group
- A positive pre-trend for group j will bias the results upwards if j is a net treatment group ($w_j^T > w_j^C$) or down if its a net control group, and if they are equal, then the bias will be zero regardless of group pre-trend
- Units treated towards the ends of the panel get relatively more weight when they act as controls.
- Needless to say, the size of the bias from a given trend is larger for groups with more weight

Variance weighted common trends

- What this means is that while all units are acting as controls, treatment timing causes some units to be controls more often - hence why they become negative (e.g., $w_k^T - w_k^C < 0$ implies w_k^C has become relatively large)
- The earliest and/or latest units get more weight as controls than treatments
- Units treated in the middle of the panel have high treatment variance as we've noted repeatedly, and so get more weight when they act as the treatment group

Variance weighted common trend weights



Testing VWCT

The identifying assumption $\sum_k \Delta Y_k^0 [w_k^T - w_k^C] = 0$ shows us how to exactly weight averages of x_{it} and perform a single t -test that directly captures the identifying assumption.

1. Generate a dummy for the effective treatment group

$$1[B_k] = w_k^T - w_k^C > 0$$

2. Estimate

$$\bar{x}_k = \beta B_k + \varepsilon_k$$

weighted by $|w_k^T - w_k^C|$

The coefficient $\hat{\beta}$ equals covariate differences weighted by the actual identifying variation and its t -statistic tests the null of reweighted balance implied the VWCT equality

Software to check the 2x2s and weights

- Austin Nichols and Thomas Goldring have made available a package in Stata called `ddtiming.ado`
- This will estimate each individual 2x2 and the weights associated with a simple two-way fixed effects model
- Let's look at it. First download Cheng and Hoekstra data from earlier (`castle-doctrine-2000-2010.dta`)
- Now install `ddtiming.ado` and use the do file that I've supplied called `hoekstra-cheng.do`

Stata

```
. use castle-doctrine-2000-2010.dta, replace  
. areg l_murder post i.year, a(sid) robust
```

Dep var	Log homicide
Castle doctrine law	0.105 (0.032)

Recall the estimated ATT is 0.105

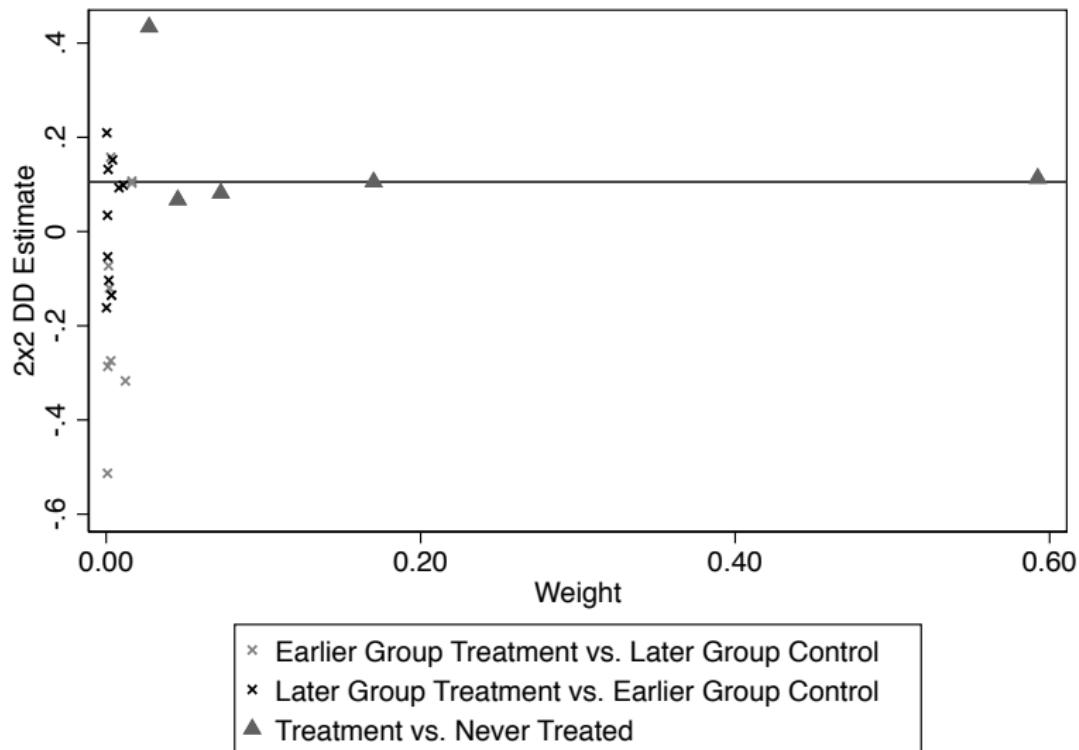
```
. ddtiming l_murder post, i(sid) t(year)
```

DD Comparison	Weight	Avg DD Est
Earlier T vs. Later C	0.060	-0.039
Later T vs. Earlier C	0.032	0.063
T vs. Never treated	0.908	0.116

```
. di (0.060*-0.039) + (0.032*0.063) + (0.908*0.116)  
. 0.105
```

Most of the 0.105 is coming from comparing treatment units to never treated units; the others cancel out

2x2s and their corresponding weights



Biased DD with OLS

- Review baker.do
- So we see – with differential timing, and heterogeneous treatment effects over time, the TWFE bias can be gigantic because:

$$plim = VWATT + VWCT - \Delta ATT_{lk}$$

- New papers are coming out focused on the issues that we are seeing with TWFE
- Callaway and Sant'anna (2019) is one of these (currently R&R at Journal of Econometrics)

Preliminary

Callaway and Sant'anna consider identification, estimation and inference procedures for ATE in DD models with

1. multiple time periods
2. variation in treatment timing (i.e., differential timing)
3. parallel trends only holds after conditioning on observables

Group-time ATE

Key concept: the ATE for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Their method will calculate an ATE per group/time which yields *many* individual ATE estimates
- Group-time ATE estimates are not determined by the estimation method one adopts (first difference or FE)
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Provides a way to aggregate over these to get a single ATE

Another contribution

- Typical econometrics paper: they propose estimators and provide asymptotically valid inference procedures for the causal parameter of interest
 - Uses a particularly kind of bootstrapping that is computationally convenient to obtain confidence intervals
- This is an extension of an older Abadie (2006) paper on semi-parametric DD with some subtle and substantive differences
- The estimator will look awfully similar to an inverse probability weighting estimator down to the use of propensity scores

Parallel trends assumption

- Parallel trends is *never* directly testable
- If you assume though that it holds in the pre-treatment period that therefore it holds in the counterfactual periods, then fine
- (IMO, this begs the question [as in assumes the conclusion]. Obviously if treatment is endogenous then parallel trends doesn't hold even if it did hold prior (see Kahn-Lang and Lang 2018))

Notation

- T periods going from $t = 1, \dots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t .
- C is also binary and indicates a control group unit equalling one if “never treated”
 - Recall the problem with OLS on using treatment units as controls
 - Callaway and Sant’anna seem to know this and working to specifically address it by essentially not using those units at all as controls
- Generalized propensity score: $p(\hat{X}) = Pr(G_g = 1 | X, G_c + C = 1)$

Propensity scores

- They'll estimate a propensity score based on group covariates using probit or logit (but not OLS)
- That score will then be normalized (e.g., Hajek weight) which improves finite sample bias
- You may need to trim it on the [0.1,0.9] interval as is commonly suggested in other applications
- Essentially, units in control group will be weighted up if their propensity scores are high, and weighted down if low, making more apple-to-apples comparisons

Detour into IPW

Horvitz weights

$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \widehat{p}(X_i)}{1 - \widehat{p}(X_i)}$$

Harjek weights

$$\widehat{\delta}_{ATT} = \left[\sum_{i=1}^N \frac{Y_i D_i}{\widehat{p}} \right] / \left[\sum_{i=1}^N \frac{D_i}{\widehat{p}} \right] - \left[\sum_{i=1}^N \frac{Y_i (1 - D_i)}{(1 - \widehat{p})} \right] / \left[\sum_{i=1}^N \frac{(1 - D_i)}{(1 - \widehat{p})} \right]$$

Parameter of interest

$$ATT(g, t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

Potential uses of this estimator

1. Are treatment effects heterogenous by time of adoption?
2. Does treatment effect change over time?
3. Are shortrun effects more pronounced than longrun effects?
4. Do treatment effect dynamics differ if people are first treated in a recession relative to expansion years?

Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Estimator

Theorem 1

$$ATT(g, t) = E \left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E \left[\frac{\hat{p}(X)C}{1-\hat{p}(X)} \right]} \right) (Y_t - Y_{g-1}) \right]$$

Which units will and will not be controls?

- Callaway and Sant'anna are keeping us from calculating DD's using TWFE, which is problematic in part bc you're implicitly calculating 2x2s by comparing later treated units to early treated units, which is a sin
- But what if you never have a true control group, or "never treated"?

Remarks about “staggered adoption” with universal coverage

Proof.

Remark 1: In some applications, eventually all units are treated, implying that C is never equal to one. In such cases one can consider the “not yet treated” ($D_t = 0$) as a control group instead of the “never treated?” ($C = 1$). □

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”

“We can aggregate the group-time treatment effects into fewer interpretable causal effect parameters, which makes interpretation easier, and also increases statistical power and reduces estimation uncertainty.” - Andrew Baker

Interesting Parameter 1

$$\frac{2}{T(T-1)} \sum_{g=2}^T \sum_{t=2}^T \mathbf{1}\{g \leq t\} ATT(g, t)$$

where T is number of pre-treatment years (Assumption 2 regarding conditional parallel trends). Let's look at an example.

Aggregating the first way

$$ATT(1986, 1986) = 10$$

$$ATT(1986, 1987) = 15$$

$$ATT(1986, 1988) = 20$$

Let data run from 1983 - 1988. Thus $T = 3$. ATT simple average is 15.

Interesting Parameter 2

$$\frac{1}{k} \sum_{g=2}^T \sum_{t=2}^T \mathbf{1}\{g \leq t\} ATT(g, t) P(G = g)$$

This is a weighted average of each $ATT(g, t)$ putting more weight on $ATT(g, t)$ with larger group sizes

Bootstrap inference

They propose a bootstrap procedure to conduct asymptotically valid inference which can adjust for autocorrelation and clustering

Coding example

See `baker.do` to illustrate basics of CS, and `castle`cs.R` to see CS in action

Comparing TWFE and CS for Cheng and Hoekstra

- Results are similar for the two estimators
- Event study plots seem similar as well
- Likely because of the large pool of never treated units which get a weight over 0.9 in TWFE
- Very little treatment effect heterogeneity – except for last period, group ATT is similar across all groups
- Be more concerned with most or all units are eventually treated by end of sample, as then you can't have the never treated comparisons

Stacking

- TWFE seems like it should identify ATT with staggered adoption since it does in simple 2x2
- Several papers show that this is not true with heterogeneity
- Several authors have shown that TWFE identifies some weighted average of group and time-specific ATT but the weights can be negative and non-interpretable

Alternatives – aggregation

- Alternatives we've examined so far are SA and CS
- Both estimate group-time ATT (or cohort-time ATT) – many parameters
- These can then be aggregated into whatever parameter you're interested in
- Easily implementable in available R or Stata software

Alternatives – stacking

- A separate stream went a different route
- Cengiz, et al. (2019) is a minimum wage study that used a “stacking” method
- Intuition is to transform the staggered adoption setting to a two-group two-period design (Gardner 2020)
- Done by creating many different datasets centering each treatment and control group on the same relative event time

Method

- For each treatment group, create a new dataset spanning a periods before and b periods after treatment adoption
- This dataset will consist of observations on the treatment group and the group of units that never receive the treatment
- So long as there are treated and untreated observations for each group group and period, you can do this
- Then stack these group-specific datasets and regress the outcomes onto treatment dummy and group-specific dummies and period dummies

Differential timing complicates plotting sample averages

- New Jersey treated in late 1992, New York in late 1993, Pennsylvania never treated
- Pre-treatment:
 - New Jersey: <1992
 - New York: <1993
 - Pennsylvania: undefined
- So how do we check parallel leads?

Early efforts at event studies

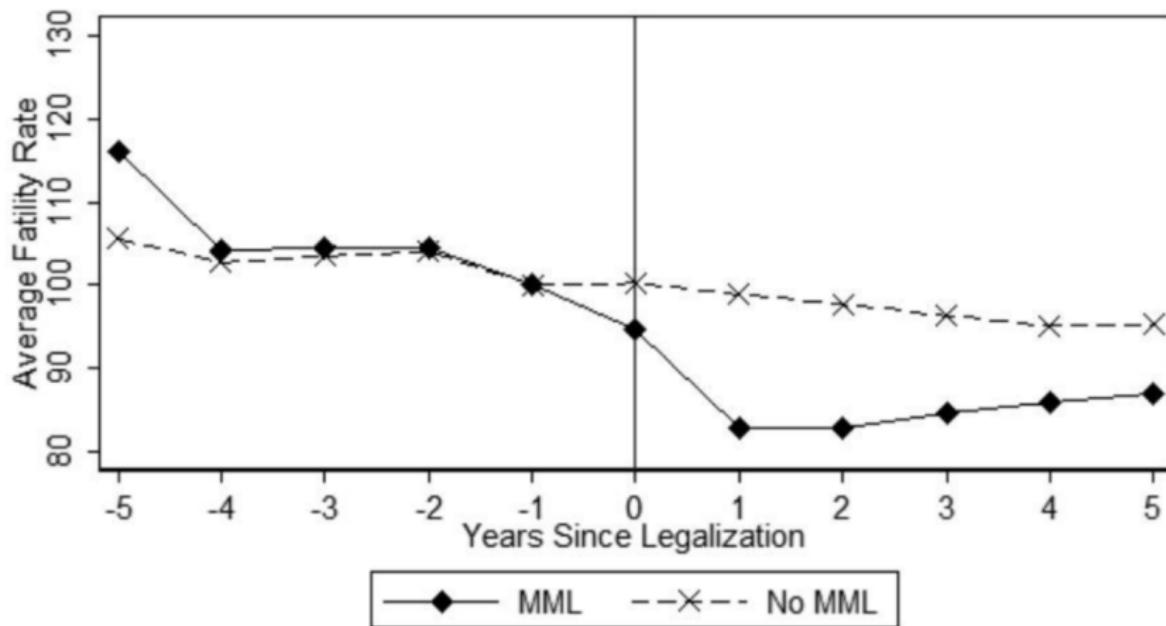


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Randomized control counties to receive arbitrary dates as treatment can be misleading

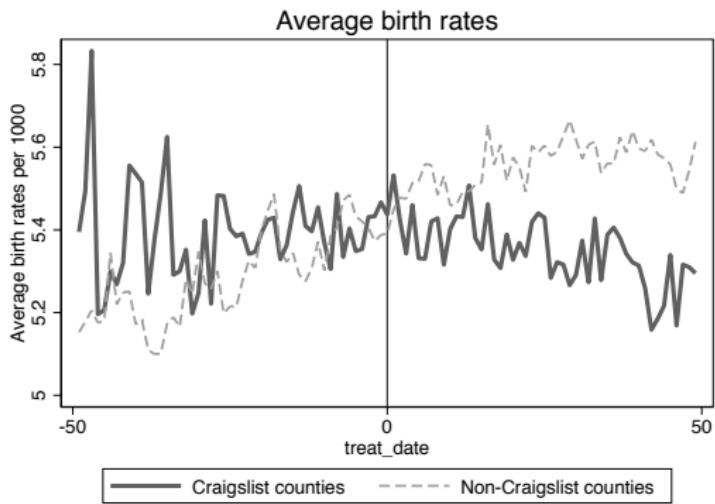
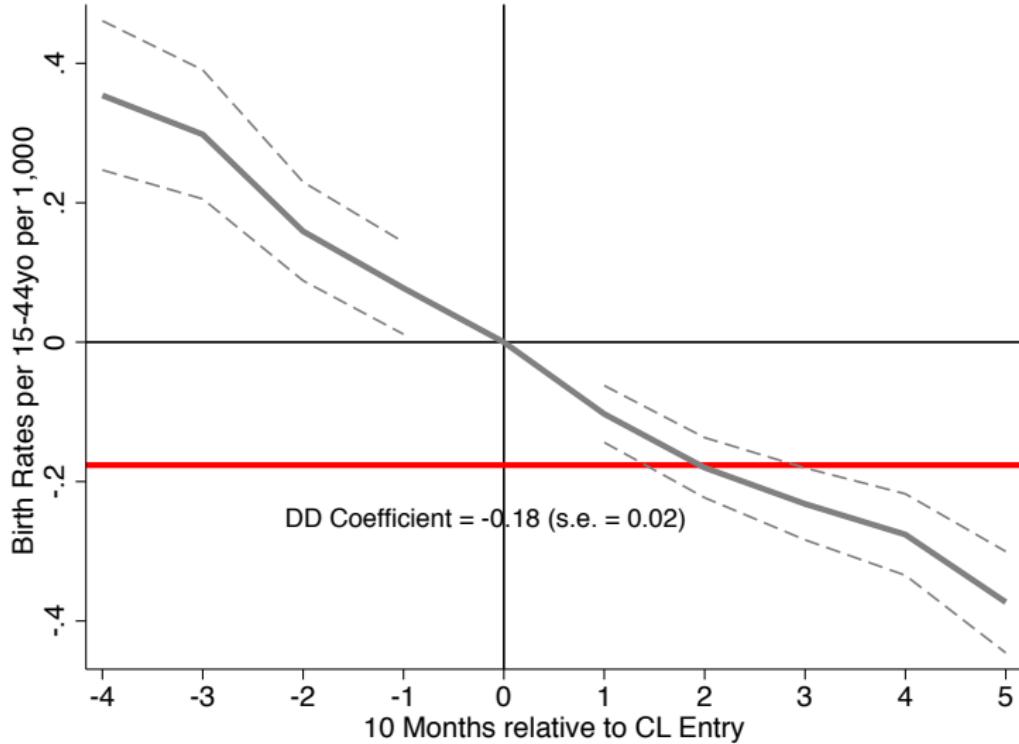


Figure: From one of my studies. Looks decent right?



Same data as a couple slides ago, leads don't look good

Revisiting the event study

- Recall our discussion of event studies estimated with TWFE under differential timing
- Now that we know about the biases of TWFE when estimating aggregate DD parameters, let's revisit event studies under differential timing
- Callaway and Sant'Anna (2020) propose alternative estimators for event studies that estimate group-time ATT in relative event time
- But now we will discuss Sun and Abraham (2020) [SA] which is like a blend of Goodman-Bacon's decomposition and Callaway and Sant'anna alternative estimator to TWFE

Summarizing

- Goodman-Bacon (2019) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

Summarizing (cont.)

1. SA is a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
2. They show that the population regression coefficient is "contaminated" by information from other leads and lags
3. SA presents an alternative estimator that is not so dissimilar to CS

Summarizing (cont.)

- Problems seem to occur with DD when we introduce treatment effect heterogeneity
- Under treatment effect heterogeneity, spurious non-zero positive lead coefficients even when there is no pretrend
- This problem is exacerbated by the TWFE related weights as under some scenarios, the weights sum to zero and “cancel out” the treatment effects from other periods
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Summarizing (cont.)

- Only decomposition of TWFE estimating dynamic leads and lags (Goodman-Bacon focused on a “static” specification)
- Contamination of coefficients on leads and lags by treatment effects depends on the magnitude of the weights on the true group-time ATT, or “cohort-specific ATT”
- Weights are a function of cohort composition
- Examining weights lets you gauge how treatment effect heterogeneity would interact with potential non-zero and non-convex weighting in population regression coefficients on the leads and lags

Difficult notation sadly

- When treatment occurs at the same time, we say they are part of the same cohort, e
- If we bin the data, then a lead or lag l will appear in the bin g so sometimes they use g instead of l or $l \in g$
- Building block is the “cohort-specific ATT” or $CATT_{e,l}$ – same thing as CS group-time ATT
- Estimate $CATT_{e,l}$ with population regression coefficient μ_l

Difficult notation (cont.)

- At each time t there are two possible treatment status $D_{i,t} \in \{0, 1\}$ over $T + 1$ time periods
- Path of treatment status scales exponentially with T and can take on 2^{T+1} possible values
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- If a group is never treated, the ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome: $Y_{i,t} - Y_{i,t}^{\infty}$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as $l = \text{year} - \text{treatment date}$

Definition 1

Definition 1: The cohort-specific ATT l periods from initial treatment date e is:

$$CATT_{e,l} = E[Y_{i,e+l} - Y_{i,e+l}^\infty | E_i = e]$$

Identifying assumption 1

Assumption 1: Parallel trends in baseline outcomes:

$E[Y_{i,t}^\infty - Y^\infty + i, s | E_i = e]$ is the same for all $e \in \text{supp}(E_i)$ and for all s, t and is equal to $E[Y_{i,t}^\infty - Y_{i,s}^\infty]$

Interesting SA comment: Never-treated units are likely to differ from ever-treated units in many ways; think of a Roy model. What does it imply that they chose not to get treated? It may imply net negative treatment effects and that could mean they may not share the same evolution of baseline outcomes as the treatment groups. If you think they are unlikely to satisfy this assumption, then drop them. Almost like a synthetic control approach.

Assumption 2

Assumption 2: No anticipator behavior in pre-treatment periods:

There is a set of pre-treatment periods such that

$$E[Y_{i,e+l}^e - Y_{i,e+l}^\infty | E_i = e] = 0 \text{ for all possible leads.}$$

Basically means that potential outcomes prior to treatment at baseline by on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

Assumption 3

Assumption 3: Treatment effect homogeneity: For each relative time period l , the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_l$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn't rule out parallel trends

TWFE Regression

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

They say E_i is the initial time of a binary variable absorbing treatment for unit i . Fixed effects should be obvious. μ_g is the population regression coefficient on the leads and lags that we want to estimate. We estimate this using OLS and get $\widehat{\mu}_g$.

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

Specifying the leads and lags

How will we specify the $1\{t - E_i \in g\}$ term? SA considers a couple:

1. Static specification:

$$Y_{i,t} = \alpha_i + \delta_t + \mu_g \sum_{l \geq 0} D_{i,t}^l + \varepsilon_{i,t}$$

2. Dynamic specification:

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_l D_{i,t}^l + \sum_{l=0}^L \mu_l D_{i,t}^l + \varepsilon_{i,t}$$

Multicollinearity

Dynamic specification requires deciding which leads to drop. They recommend dropping two: $l = -1$ and some other one (they seem to favor $l = -4$). The reason is twofold. You drop one of them to avoid multicollinearity in the relative time indicators. You drop a second one because of the multicollinearity coming from the linear relationship between TWFE and the relative period indicators.

Trimming and binning

- First some terms: trimming and binning, I do both in the Mixtape when analyzing Cheng and Hoekstra (2013)
- Binning means placing all “distant” relative time indicators into a single one. Done because of the sparseness of units in such distant bins. So if there’s 3 distant leads and lags that aren’t balanced, combine them all into the last lead and lag
- Trimming means excluding any relative period for which you don’t have balance in relative time. This creates a balanced panel “in relative time”, but imbalanced panel length overall.
- They’ll analyze both and how they affect $\widehat{\mu}_g$ estimation using TWFE

Interpreting $\widehat{\mu}_g$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$\begin{aligned}\mu_g = & \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Good stuff}} \\ & + \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{Bleh - Other included relative time}} \\ & + \underbrace{\sum_{l \in g^{excl}} \sum_e w_{e,l}^g (E[Y_{i,e+l} - Y_{i,0}^\infty | E_i = e] - E[Y_{i,e+l}^\infty - Y_{i,0}^\infty])}_{\text{More bleh - Excluded}}\end{aligned}$$

Superscript g associates the weight with coefficient μ_g . The weight associated with cohort e in relative period l is equal to the population regression coefficient on the $1\{t - E_i \in g\}$ from regression $D_{i,t}^l \times 1\{E_i = e\}$ on all bin indicators included in the regression and TWFE. Just the mechanics of double demeaning from TWFE

Weight ($w_{e,l}^g$) summation cheat sheet

1. For relative periods of μ_g own $l \in g$, $\sum_{l \in g} \sum_e w_{e,l}^g = 1$
2. For relative periods belonging to some other bin $l \in g'$ and $g' \neq g$,
 $\sum_{l \in g'} \sum_e w_{e,l}^g = 0$
3. For relative periods not included in G , $\sum_{l \in g^{excl}} \sum_e w_{e,l}^g = -1$

Estimating the weights

Regress $D_{i,t}^l \times 1\{E_i = e\}$ on:

1. all bin indicators included in the main TWFE regression,
2. $\{1\{t - E_i \in g\}\}_{g \in G}$ (i.e., leads and lags) and
3. the unit and time fixed effects

Interpretation of coefficients under parallel trends only

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bin g is a linear combination of $CATT_{e,l \in g}$ as well as $CATT_{d,l'}$ from other relative periods $l' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g = & \underbrace{\sum_{l \in g} \sum_e w_{e,l}^g CATT_{e,l}}_{\text{Desirable}} \\ & + \underbrace{\sum_{g' \neq g, g' \in G} \sum_{l' \in g'} \sum_e w_{e,l'}^g CATT_{e,l'}}_{\text{Undesirable - other specified bins}} \\ & + \underbrace{\sum_{l' \in g^{excl}} \sum_e w_{e,l'}^g CATT_{e,l'}}_{\text{Undesirable - excluded relative time indicators}}\end{aligned}$$

Comment on Proposition 2

The coefficient μ_g can be written as an average of $CATT_{e,l}$ from own periods but also $CATT_{e,l'}$ from other periods.

The weights are still functions of cohort comparisons, like in Proposition 1, which means μ_g can be written as non-convex averages of not only $CATT_{e,l}$ from own periods $l \in g$, but also $CATT_{e,l'}$ from other periods.

Means μ_g could in fact be the wrong sign to all $CATT_{e,l \in g}$.

Weights can help us gauge the severity of this problem.

When the weights have larger magnitude, treatment effect heterogeneity matters more as a particular $CATT_{e,l}$ can drive the overall estimates. But when weights are uniform, treatment effect heterogeneity matters less.

Interpretation under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$\begin{aligned}\mu_g = & \sum_{l' \in g, l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_e w_{e,l'}^g CATT_{e,l'} \\ & + \sum_{l' \in g^{excl}, l' \geq 0} \sum_e w_{w,l'}^g CATT_{e,l'}\end{aligned}$$

Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods even *though parallel trends hold in the pre period.*

Proposition 4

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,l} = ATT_l$ is constant across e for a given l , and the population regression coefficient μ_g is equal to a linear combination of $ATT_{l \in g}$, as well as $ATT_{l' \notin g}$ from other relative periods

$$\begin{aligned}\mu_g &= \sum_{l \in g} w_l^g ATT_l \\ &+ \sum_{g' \neq g} \sum_{l' \in g'} w_{l'}^g ATT_{l'} \\ &+ \sum_{l' \in g^{excl}} w_{l'}^g ATT_{l'}\end{aligned}$$

Proposition 4 comment

The weight $w_l^g = \sum_e w_{e,l}^g$ sums over the weights $w_{e,l}^g$ from Proposition 1 and is equal to the population regression coefficient from the following auxiliary regression:

$$D_{i,t}^l = \alpha_i + \lambda_t + \sum_{g \in G} w_l^g \cdot 1\{t - E_i \in g\} + u_{i,t}$$

which regresses $D_{i,t}^l$ on all bin indicators and TWFE

On binning

- Many propose either binning or trimming to create “balanced” panels (in relative event time)
- But SA notes that binning in simulations creates uninterpretable weights (due to the binned $CATT_{e,l'}$ inclusion in μ_g), whereas trimming creates weights that are more reasonable
- This may be because trimming subtracts the corresponding $CATT_{e,l'}$ from μ regression coefficient

Intuition for contamination

- Stupid notation make Hulk smash!
- Let's do a simple toy example instead

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

Toy example

$$\begin{aligned}\mu_{-2} = & \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}} \\ & + \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}\end{aligned}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

Interaction-weighted estimator

- They propose an interacted weighted estimator (IW) as a consistent estimator for μ_g
- Estimator uses either never-treated as controls or “last cohort treated” if no never-treated (contra CS which uses “not yet treated”)
- No covariates bc this is a regression with fixed effects and time-varying covariates create own biases, although they note you can plug in CS for the DD calculation and recover *CATT* that way
- The interaction is a TWFE regression specification that interacts relative period indicators with cohort/group indicators, excluding indicators for never-treated cohorts

Interaction-weighted estimator

- **Step one:** Do this DD regression and hold on to $\widehat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated. The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

Interaction-weighted estimator

- **Step two:** Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

IW estimator

- **Step three:** Take a weighted average of estimates for $CATT_{e,l}$ from Step 1 with weight estimates from step 2

$$\hat{v}_g = \frac{1}{|g|} \sum_{l \in g} \sum_e \hat{\delta}_{e,l} \widehat{Pr}\{E_i = e | E_i \in [-l, T - l]\}$$

Consistency and Inference

- Under parallel trends and no anticipation, $\hat{\delta}_{e,l}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IV estimator is consistent for a weighted average of $CATT_{e,l}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

DD Estimator of CATT

Definition 2: DD estimator with pre-period s and control cohorts C estimates $CATT_{e,l}$ as:

$$\widehat{\delta}_{e,l} = \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} \times 1\{E_i \in C\})]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,l}$.

Software

Use `staggered` from

<https://github.com/jonathandroth/staggered> by Jon Roth
(Brown University). There is also a Stata wraparound using the `rcall` package in Stata. See instructions on the URL above.

Conclusion of SA

- Bacon shows the TWFE coefficient on the static parameter is “contaminated” by other periods leads and lags
- Three strong assumptions needed for TWFE to be unbiased: parallel trends, no anticipation, and treatment homogeneity
- Three step interaction-weighted estimator is an alternative
- Doesn’t restrict to treatment profile homogeneity
- Callaway and Sant’Anna (2020) and Sun and Abraham (2020) use different controls, but under certain situations (no covariates, never treated) they are the same (“nested”)
- Software in R and Stata exist

Sharp DD

- In a “sharp” DD, a group gets treated in period 1, a control group does not
- Parallel trends allows you to identify ATT
- We discussed several methods
- But sometimes the lines between treatment and control groups get “fuzzy”

Fuzziness

- In a “fuzzy” DD design, there’s growth in treatment occurring naturally in the control group
 - They discuss an early 2000s Duflo paper where Indonesia pushed for more primary schooling
 - Used earlier cohorts as controls bc they were already past the age
 - But they saw growth in schools too
- In many applications, the “treatment rate” increase more in some groups than in others but there is no group that goes from fully untreated to fully treated
- But there is no group that also remains fully untreated

Earlier fuzzy estimators

- Popular estimator (10% of AERs from 2010-2012) divides DiD by the DiD of the treatment

$$Wald_{DiD} = \frac{\left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)}{\left(E[D_k|Post] - E[D_k|Pre] \right) - \left(E[D_U|Post] - E[D_U|Pre] \right)}$$

- It's Wald IV in that we scale the reduced form by the first stage but they call it Wald DiD
- de Chaisemartin and D'Haultfoeuille (2017) estimates the LATE for group's who go from untreated to treated

Personal takeaway

- Two main values of this paper that I found:
 - Situations where the control group is getting treated with unrelated policy shocks
 - Continuous treatments
- Code to do it is simple but in Stata

Most basic notation

For any random variable, R, we interpret as R_{dgt} as treatment status, treatment group, time

$$R_{101} \sim R|D = 1, G = 0, T = 1$$

Treatment status (D) is whether a unit is treated regardless of group; Group (G) is treatment or control groups; Time (T) is before or after

Cases under consideration

Case 1: Share of treated units in control don't change between periods

$$E[D_{01}] = E[D_{00}]$$

Wald_{DiD} identifies the LATE parameter for “switchers” (i.e., people whose treatment status changed between 0 and 1) if parallel trends holds and if the ATE of treated units at both dates is stable over time; proposes new estimators that don’t depend on this

Stable ATE isn’t required in a typical “sharp” DiD

Cases under consideration

Case 2: Share of treated units changes over time in control

$$E[D_{01}] > E[D_{00}]$$

Wald_{D_iD} identifies the LATE of switchers under PT and stable ATE assumption and LATE of treatment and control group switchers are the same

Under certain assumptions, their alternative estimator will only be partially identified, and it depends on the size of the change of treated units in the control.

Concluding remarks on DD

- Chances are you are going to write more papers using DD than any other design
- Goodman-Bacon (2018, 2019) and Sun and Abraham (2020) is *worth your time* because their decompositions show sources of bias in TWFE under reasonable scenarios
- Callaway and Sant'anna (2020) is an extremely useful contribution to the DD toolbox for showing a way to estimate the group-time ATT using any variety of approaches, including regression
- Sun and Abraham (2020) also provides a way forward for event studies (as does CS)