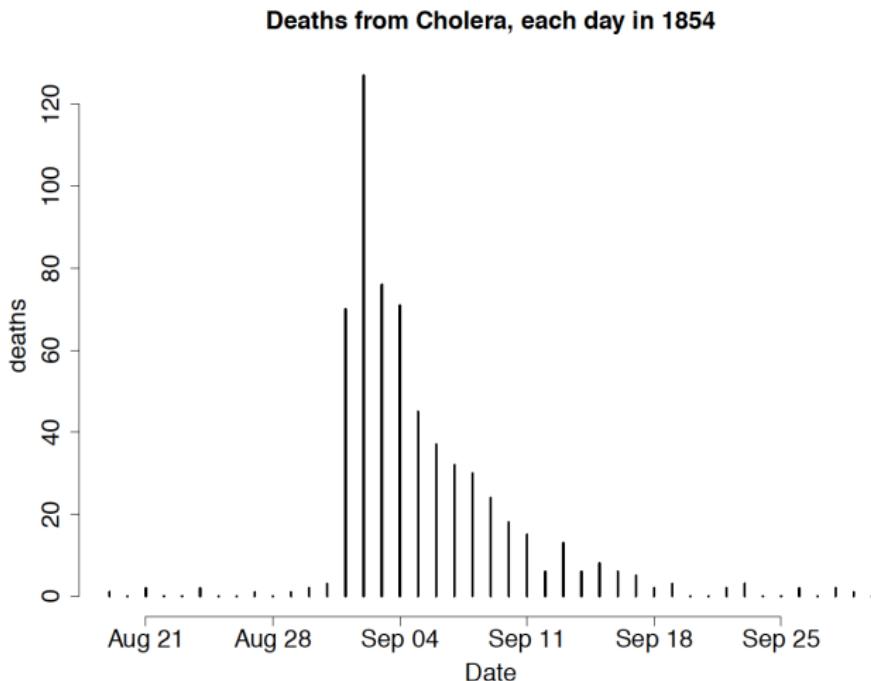


## John Snow

- John Snow was a practicing anesthesiologist in the mid 19th century London
- He was then famous for inventing a machine that would carefully deliver chloroform to patients in homogenous dosage which reduced mortality from anesthesia
- But he is now famous for providing convincing evidence that cholera was a waterborne disease during the 1854 outbreak
- Published two works on cholera – an essay in 1849, and a book in 1855
- Died of a stroke in 1858



**Figure:** Daily cholera deaths, London (Coleman 2019)

## Cholera background

- Cholera hits London three times in the early to mid 1800s causing large waves of tens of thousands of deaths
- Three London epidemics – 1831-1832, 1848-1849, 1853-1854
- Cholera attacked victims suddenly, with a 50% survival rate, and very painful symptoms included vomiting and acute diarrhea

## Miasmis

- 19th century London was a filthy place with waste collecting in cesspools under houses or emptied into open ditches and sewers
- Majority opinion about disease was *miasmis*
- Miasmis hypothesized that disease transmission was caused by vapors and smells; unclear its relevance for person-to-person

## Never before seen microorganism

- Microscopes were around but had horrible resolution
- Most human pathogens couldn't be seen
- Johnson (2007) reports Snow did track down a microscope but could only see blurry things moving around
- Isolating these microorganisms wouldn't occur for half a century

## Snow's hypothesis

- Snow (as well as a few others like Rev. Henry Whitehead) believe miasms is not relevant for explaining cholera
- Snow hypothesizes that the active agent was a living organism that entered the body, got into the alimentary canal with food or drink, multiplied in the body, and generated some poison that caused the body to expel water
- The organism passed out of the body with these evacuations, entered the water supply and infected new victims
- The process repeated itself, growing rapidly through the common water supply, causing an epidemic

## Thought Experiment

- How will he convince anyone that cholera is waterborne and not due to “bad air”?
- Consider the ideal experiment: randomize households by coin flip to receive water from runoff (control) vs. water without runoff (treatment)
- Unethical, impractical and unrealistic
- Even if the randomized experiment is not possible, the thought experiment suggests the observational equivalent

## Multiple sources of evidence, not just one

Snow makes his argument with many pieces of evidence that when taken together are very compelling that water, not air, is the cause of the cholera epidemics. These can be categorized as:

- ① Observation
- ② Broad Street Pump
- ③ Grand Experiment

## Observation

- Observed progression of the disease for years
- Tracked Patient Zero
- Treatments didn't work: Snow would cover with burlap sacks, which did nothing
- Strange irregular patterns – higher deaths in close proximity to a public pump on Broad Street, fewer deaths at a pub

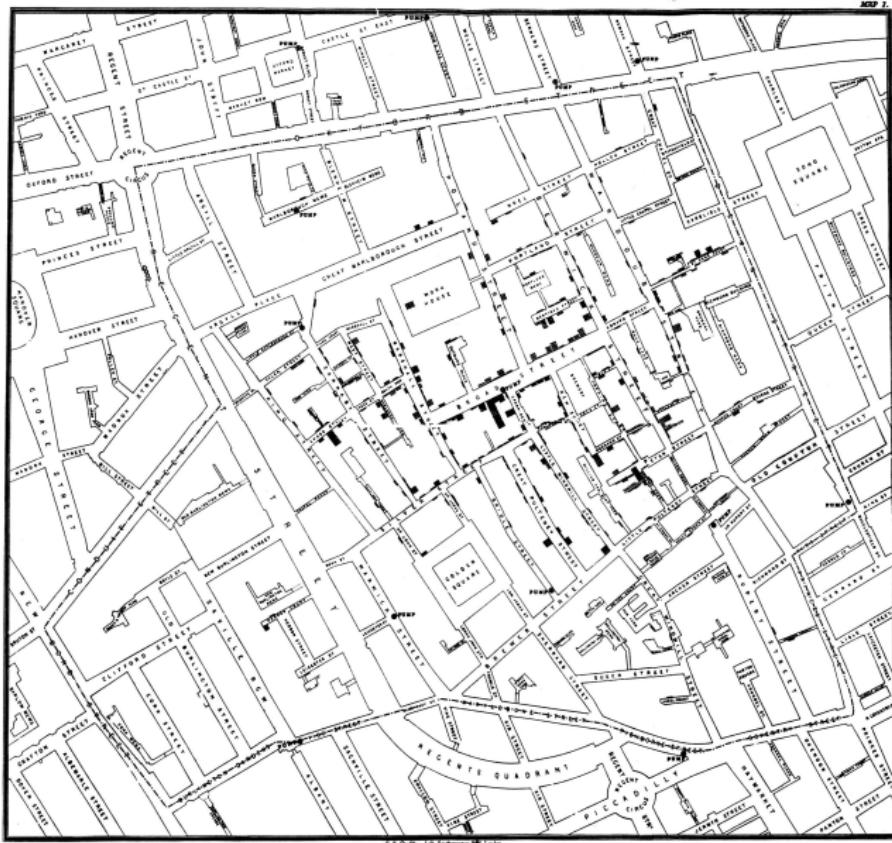
*"cholera extended to nearly all the houses in which the water was thus tainted, and to no others." (Snow 1849)*

## Broad street outbreak

*"The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street [now Lexington St.] joins Broad Street [now Broadwick], there were upwards of five hundred fatal attacks of cholera in ten days."* (Snow 1855)

## How he argues for the Broad street pump

- Famous map showing unusual mass of cholera deaths near the public Broad street pump
- He was looking for the source, but he was not inductively forming his theory with this map because he already knew the mechanism
- He was assembling evidence that would further refute the explanations of those who advocated an alternative explanation of the outbreak



**Figure:** Cholera deaths laid over a small area of London near Broad Street

## **Map was important but not enough on its own**

*"[Snow] could see at a glance that he'd be able to demonstrate that the outbreak was clustered around the pump, yet he knew from experience that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over that part of Soho, something emanating from the gully holes or cesspools – or perhaps even from the pump itself. Snow knew that the case would be made in the exceptions from the norm. Pockets of life where you could expect death, pockets of death where you would expect life." Johnson (2007) p. 140*

## Two companies fight for customers

- Southwark and Vauxhall Waterworks Company and the Lambeth Water Company competed over some of the regions south of the Thames
- In 16 sub-districts, with a population of 300,000, they competed directly, even supplying customers side-by-side

*"In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies." Snow (1855) p 75*

## Lambeth moves its pipe

- During the 1849 epidemic, both companies drew water from Thames which was polluted with sewage and cholera
- London passes legislation requiring utility companies to move their pipes above the city
- In 1852, the Lambeth Company, a water utility company, changed supply from Hungerford Bridge
- It moved its intake pipe upstream to cleaner water and in response to legislation (SV delayed)
- This created a natural experiment because Southwark and Vauxhall left its intake pipe in place

## Meticulous Data Collection

- Two types of data: DD uses aggregate deaths bc of mixing of customers whereas his Broad Street evidence focused on individuals
- Collected detailed information from households with cholera deaths on utility subscription (Lambeth or SV)
- Many residents didn't know their water company – distant landlords paid for it
- He knew Lambeth water was four times saltier, so he'd take a sample and test it using a saline test back at his office

## **Shoelather and knowledge of institutional details**

- Careful balance checks – “the pipes of each Company go down all the streets into nearly all the courts and alleys”
- Concern for sample selection bias –“No fewer than 3000 people of both sexes [of all types affected]”
- Treatment assignment was arbitrary – “a few houses supplied by one Company and a few by the other”

**Table XII**

**Modified Table XII (Snow 1854)**

<b>Company name</b>	<b>1849</b>	<b>1854</b>
Southwark and Vauxhall	135	147
Lambeth	85	19

Estimated ATT using DD is 78 fewer deaths per 10,000

## **Failure to convince**

*"In spite of what has since been recognized as a classic exercise in data, analysis, and argument, Snow failed to convince the medical profession, the policy-making establishment, or the public." (Coleman 2019)*

## Final victory

- Another cholera outbreak in 1866, east of London, is when Snow's ideas were gradually and reluctantly accepted by public officials and the scientific community
- 1866 outbreak was confined only to the east of London, which was the last area not yet connected to the newly constructed sewage system which discharged sewage below the Thames
- The rest of London didn't have an outbreak
- This was the final piece of evidence that swayed skeptics and led to a more reasoned assessment of Snow's data and analysis

## **Merits of Snow's work**

- Long commitment to the topic led him to reject unsound hypotheses and form new ones based on observation and experience (shoe leather)
- Expert handling of data analysis, data visualization, and a framing of evidence with a ladder of reasoning

## **Layered rhetoric of research**

*"The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself." (Johnson, Ghost Map)*

Hidden curriculum  
Foundational causality stuff  
Regression discontinuity designs  
Instrumental variables  
Two-way fixed effects estimator  
**Differences-in-differences**  
Comparative case studies  
Matching and weighting  
Concluding remarks

John Snow  
The simple 2x2  
Event study  
Differential timing

## 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	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
Southwark and Vauxhall	Before	$Y = SV$		$D$
	After	$Y = SV + T$	$T$	

## Sample averages

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

## **Population expectations**

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

## Potential outcomes and the switching equation

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

## Parallel trends bias

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

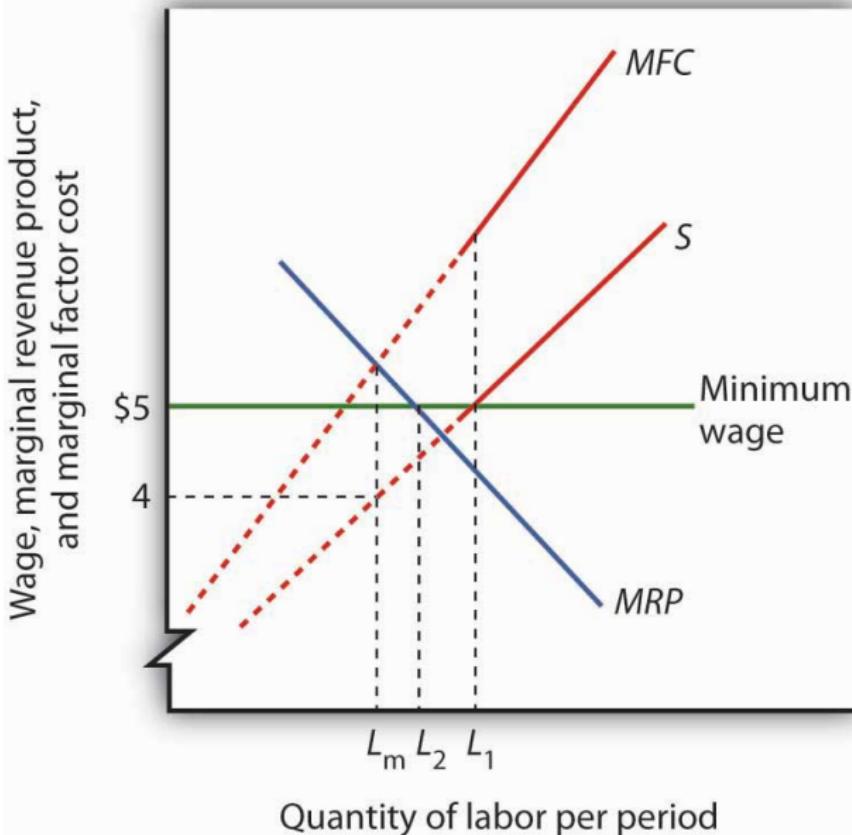
## Another famous DD study

- 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

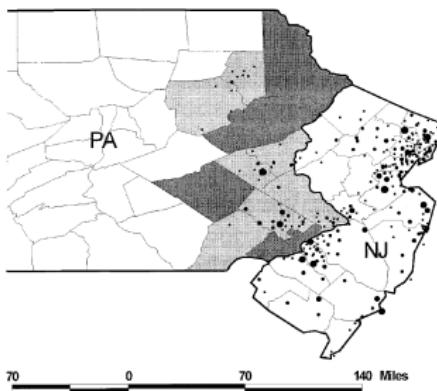
- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In a competitive input market, increases in the minimum wage would move us up a downward sloping labor demand curve → employment would fall
- Monopsony (imperfect labor markets) suggest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

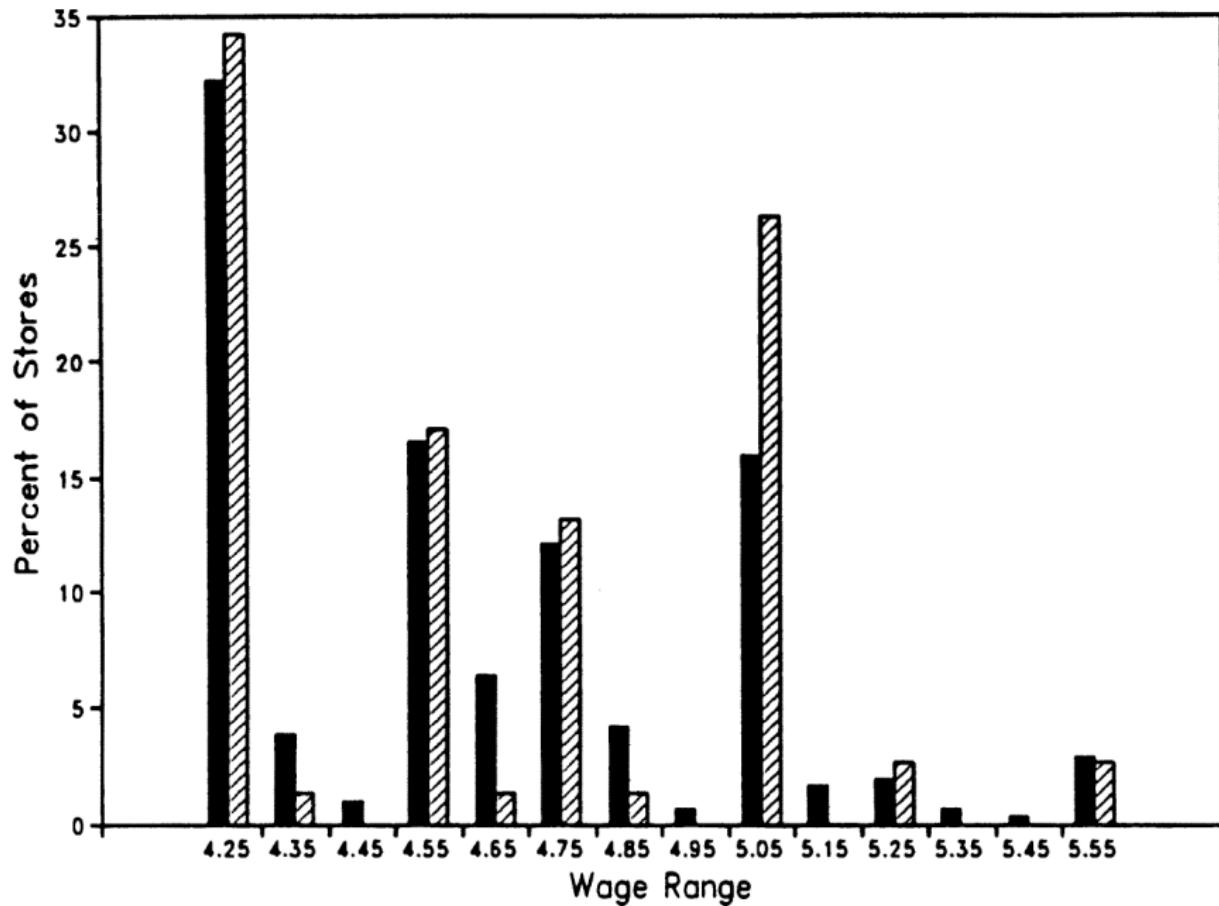
## Parallel trends assumption

- Key identifying assumption is the “parallel trends” assumption

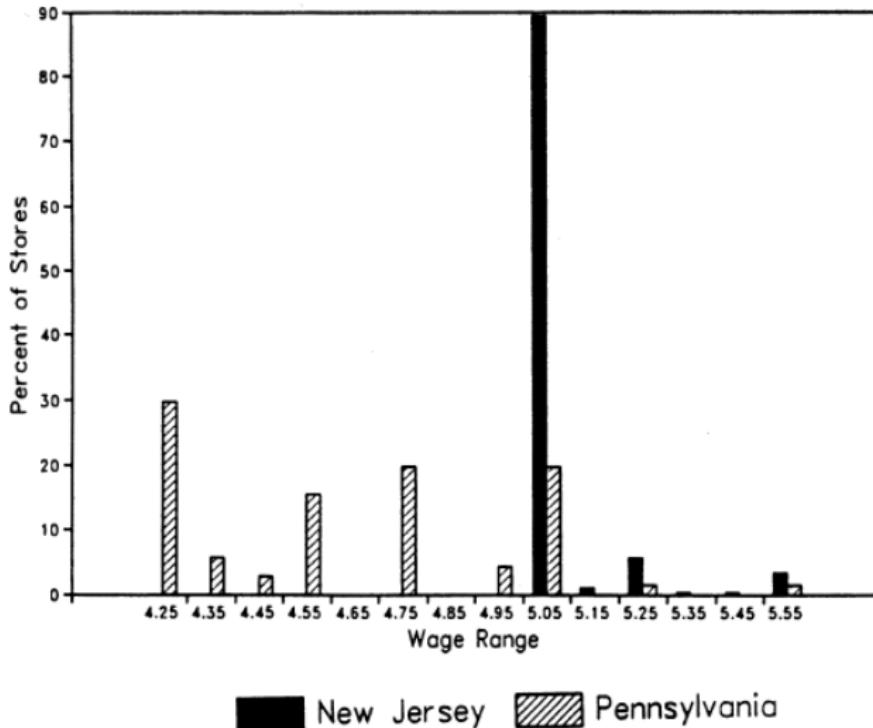
$$\underbrace{[E[Y_{NJ}^0 | Post] - E[Y_{NJ}^0 | Pre]] - [E[Y_{PA}^0 | Post] - E[Y_{PA}^0 | Pre]]}_{\text{Non-parallel trends bias}}$$

- 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

- Remember, I said there are good reasons to use TWFE
  - It's easy to calculate the standard errors
  - We can control for other variables which may reduce the residual variance (lead to smaller 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

## Regression DD

The typical regression model we estimate is

$$Y_{it} = \beta_1 + \beta_2 \text{Treat}_i + \beta_3 \text{Post}_t + \beta_4 (\text{Treat} \times \text{Post})_{it} + \varepsilon_{it}$$

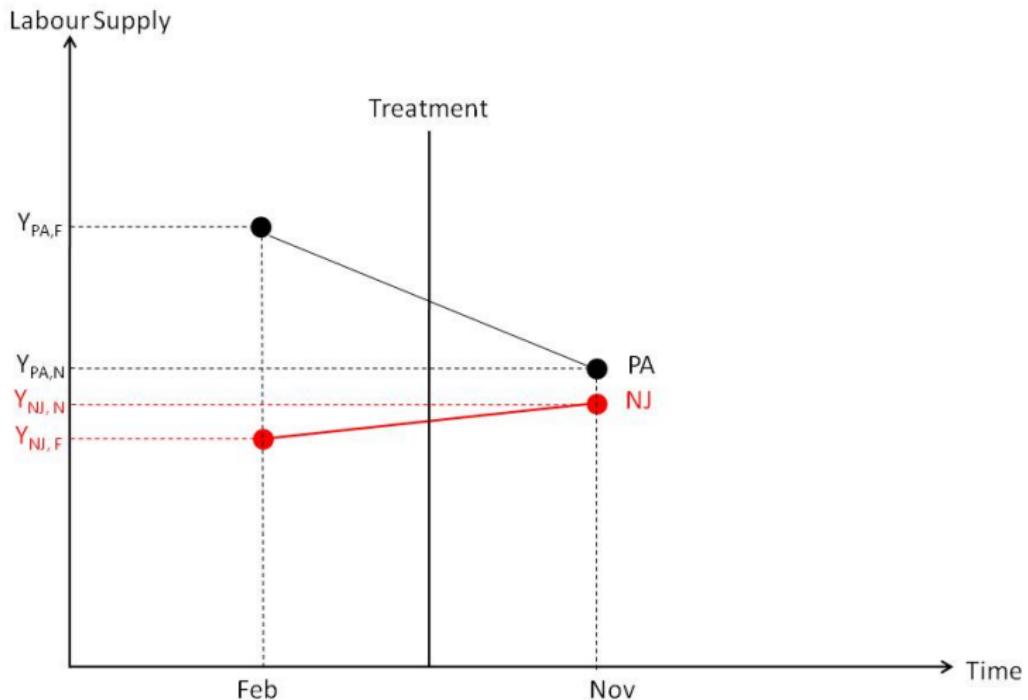
where Treat is a dummy if the observation is in the treatment group and Post is a post treatment dummy

## Regression DD - Card and Krueger

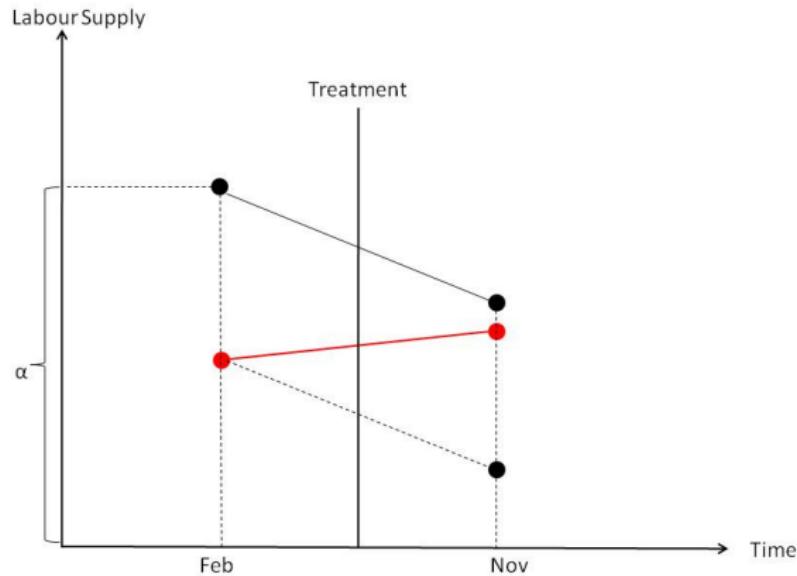
- In the Card and Krueger case, the equivalent regression would be:

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

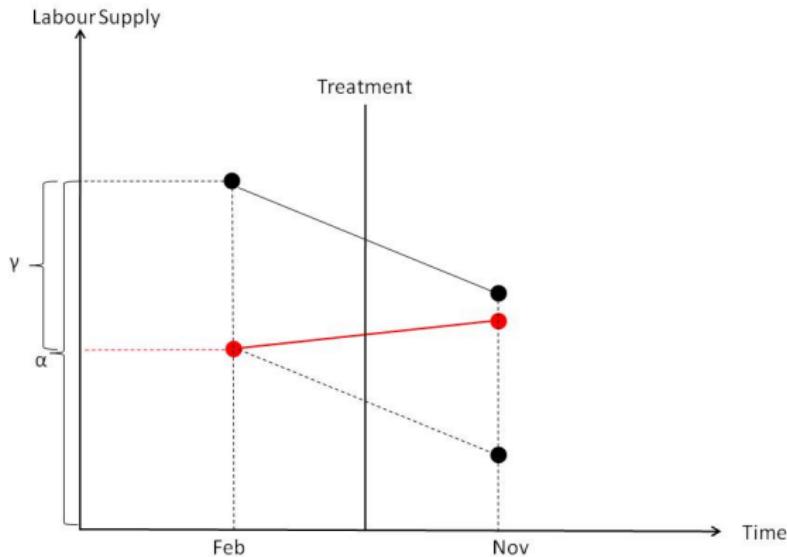
- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
  - PA Pre:  $\alpha$
  - PA Post:  $\alpha + \lambda$
  - NJ Pre:  $\alpha + \gamma$
  - NJ Post:  $\alpha + \gamma + \lambda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



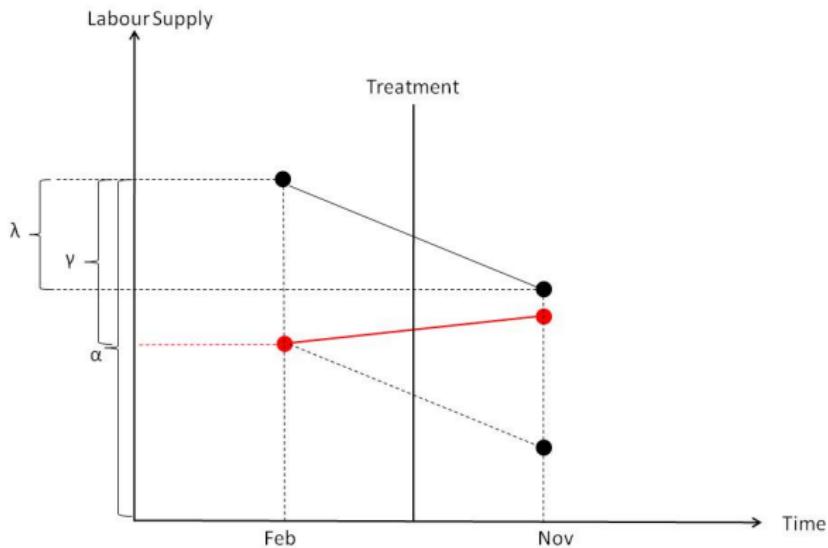
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



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



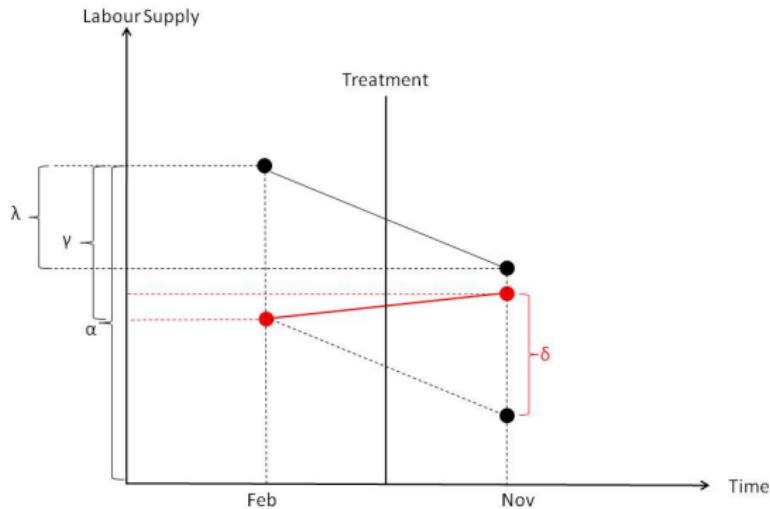
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



## **Key assumption of any DD strategy: Parallel trends**

- The key assumption for any DD strategy is that the outcome in treatment and control group would follow the same time trend in the absence of the treatment
- This doesn't mean that they have to have the same mean of the outcome
- But regardless of parallel trends, OLS always estimates the vertical bar on next slide

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



## 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 for the treatment group

## Labor Supply

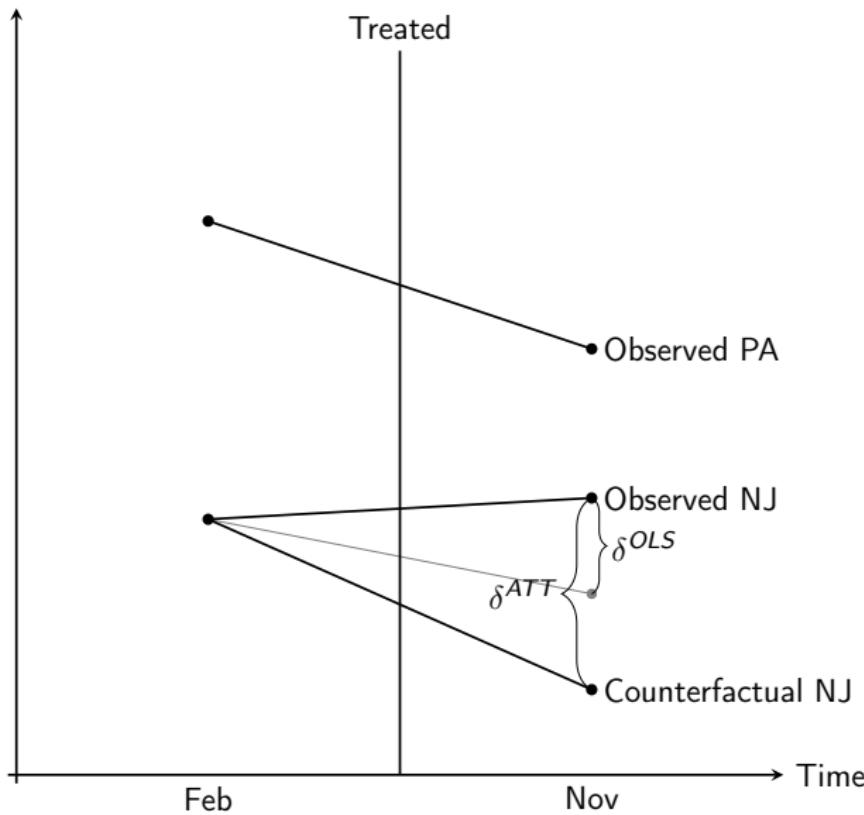


Figure: DD regression diagram without parallel trends

## 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

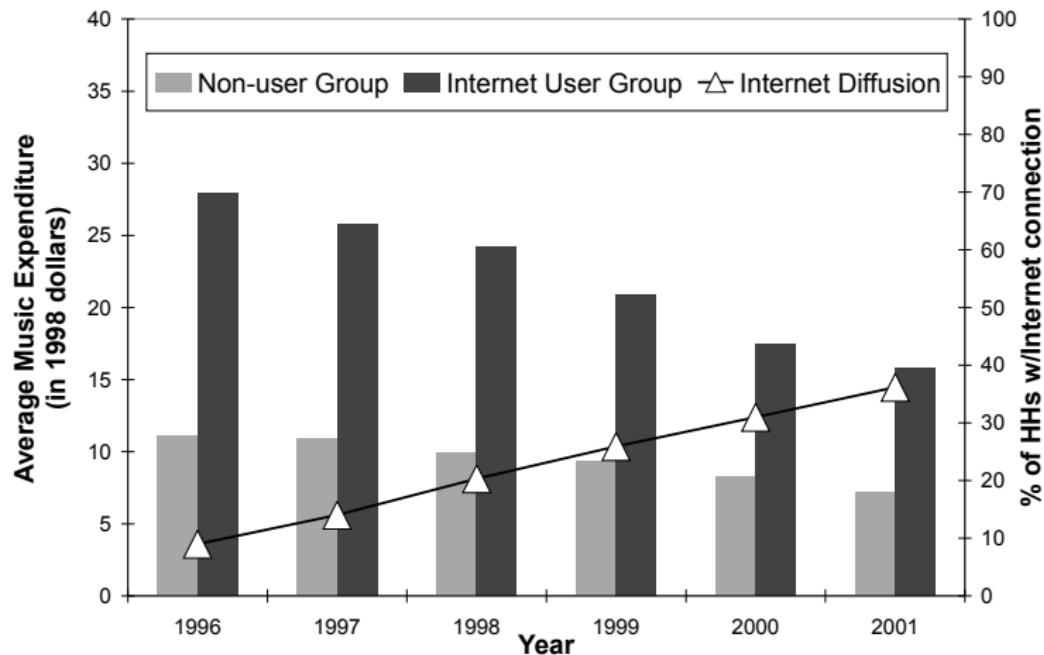


Table 1: Descriptive Statistics for Internet User and Non-user Groups<sup>a</sup>

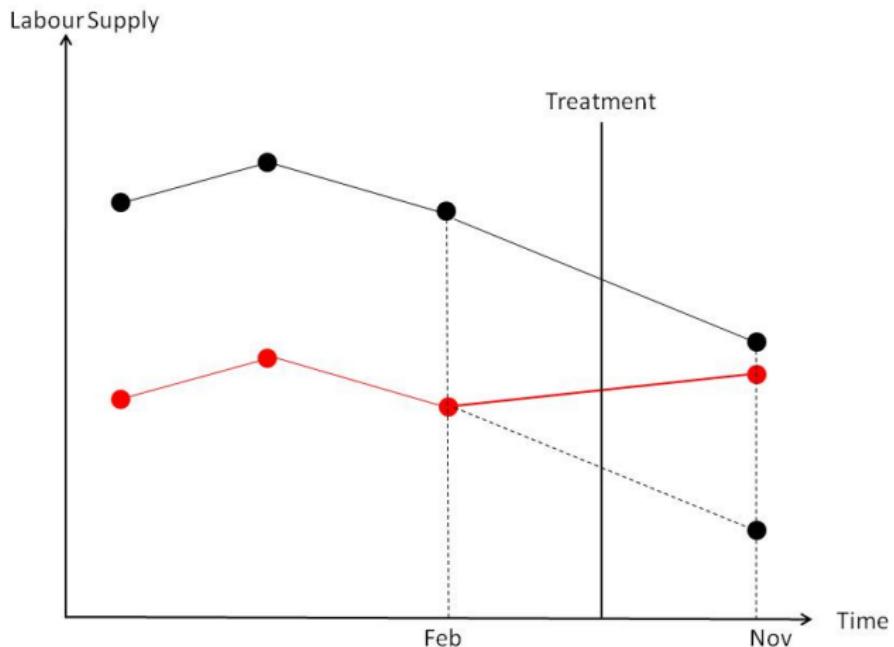
Year	1997		1998		1999	
	Internet	User	Non-user	Internet	User	Non-user
<b>Average Expenditure</b>						
Recorded Music	\$25.73	\$10.90		\$24.18	\$9.97	\$20.92
Entertainment	\$195.03	\$96.71		\$193.38	\$84.92	\$182.42
<b>Zero Expenditure</b>						
Recorded Music	.56	.79		.60	.80	.64
Entertainment	.08	.32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

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

## Parallel leads, not trends

- The identifying assumption for all DD designs is some representation of a counterfactual parallel trend
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check using pre-treatment data to show that the trends had been the same prior to treatment
- 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



## Differential timing makes pre-treatment undefined for untreated groups

- 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?

## Randomize treatment dates to control units

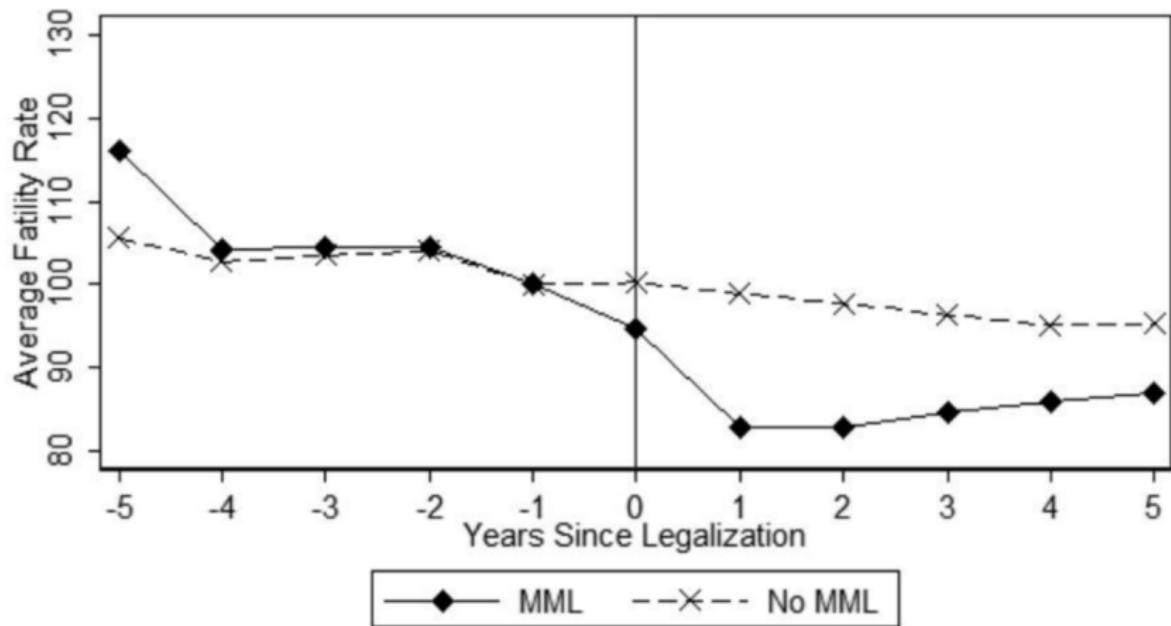


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

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

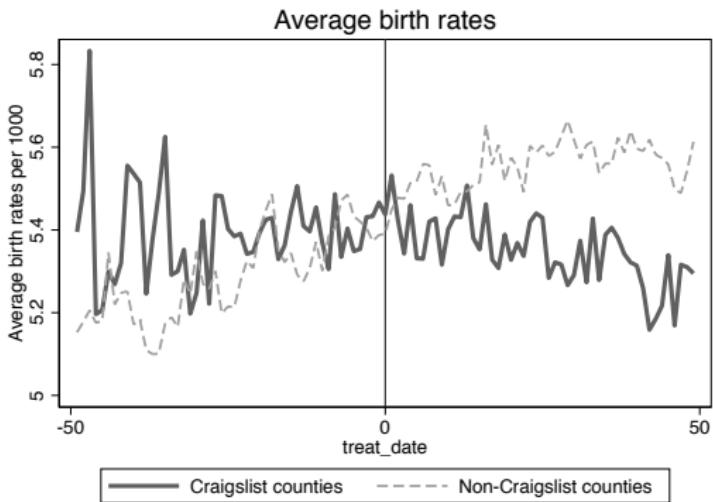


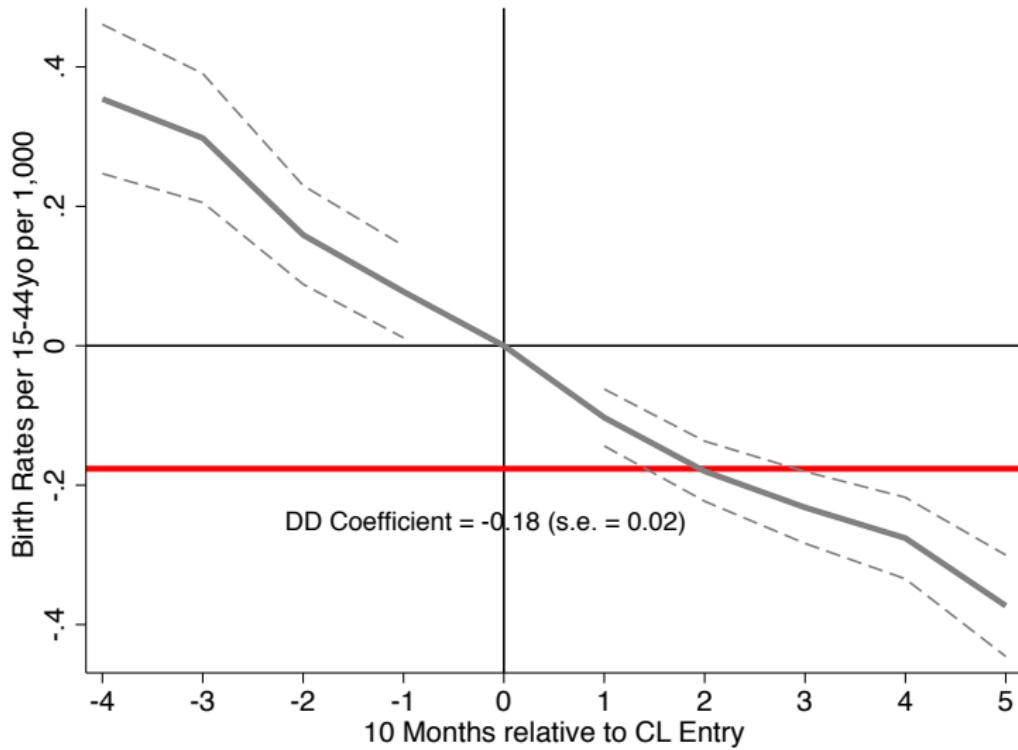
Figure: From one of my studies. Looks decent right?

## 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:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-1}^{-q} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

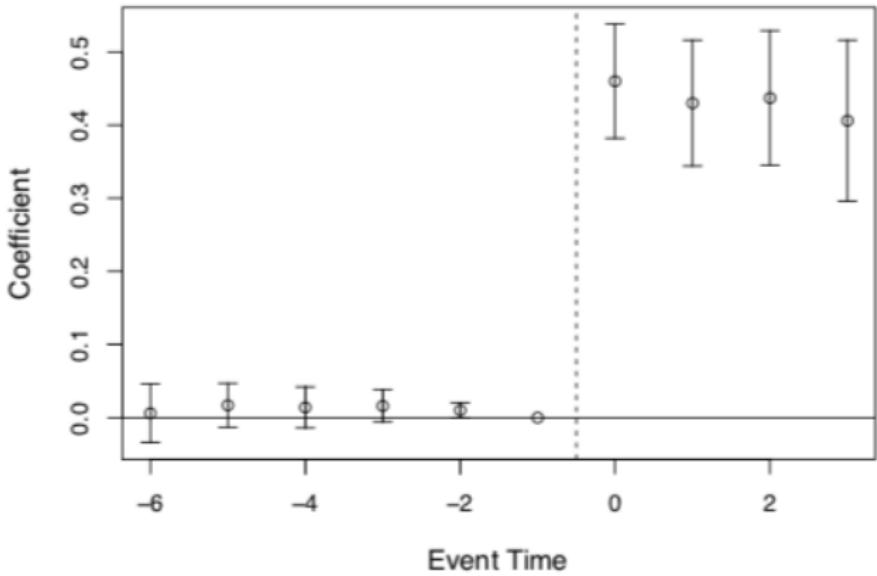
- Treatment occurs in year 0
- Includes  $q$  leads or anticipatory effects
- Includes  $m$  leads or post treatment effects



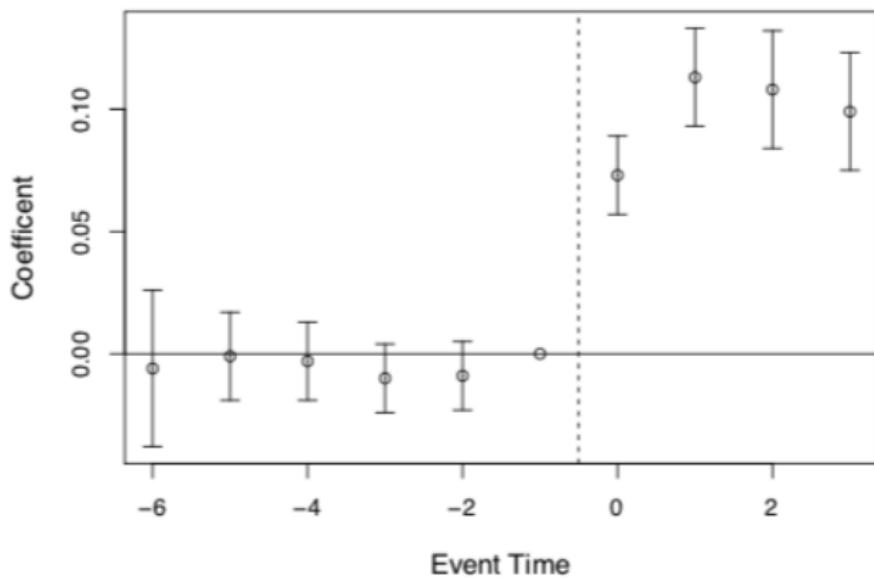
Same data as a couple slides ago, leads don't look good

## Medicaid and Affordable Care Act example

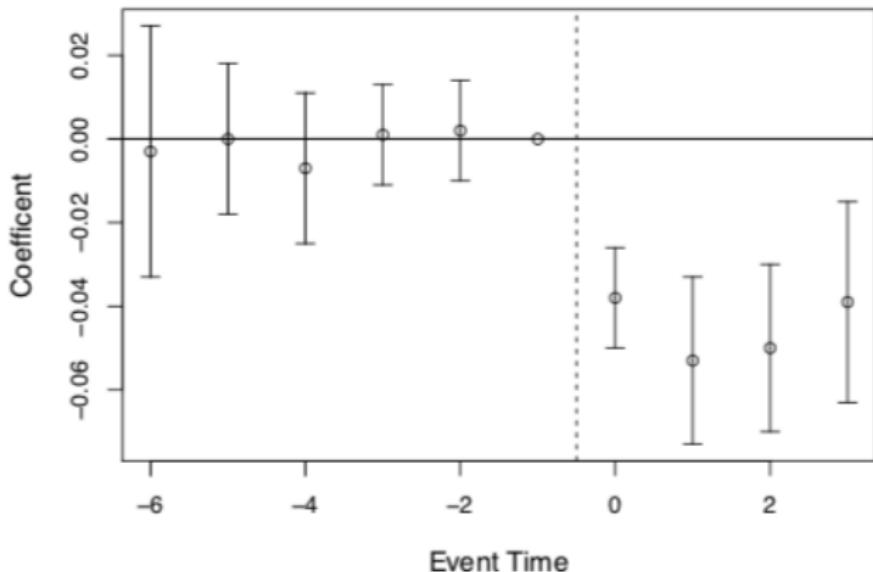
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

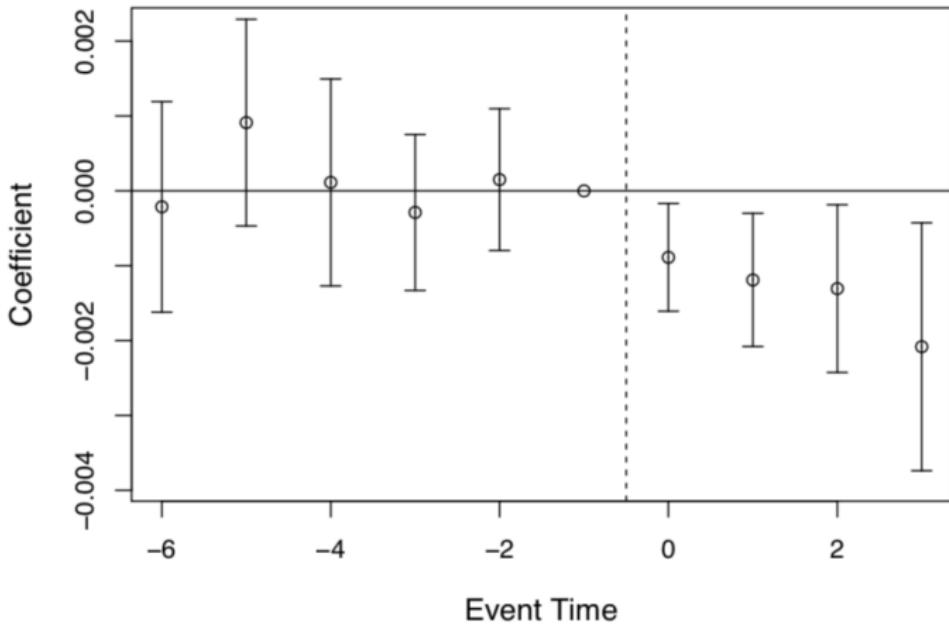


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

## 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 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

## Note about groups

- Correct treatment of standard errors sometimes makes the number of groups very small: in the Card and Krueger study the number of groups is only 2.

## DD Robustness

- 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 leads and lags we already discussed
  - Event study (already discussed)
  - Falsification test using data for alternative control group
  - Falsification test using alternative “placebo” outcome that should not be affected by the treatment

## Within group controls - triple diff

**Table:** Differences-in-differences-in-differences

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

# 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
<b>A. Treatment Individuals: Married Women, 20–40 Years Old:</b>			
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>B. Control Group: Over 40 and Single Males 20–40:</b>			
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)	
<b>DDD:</b>		<b>-0.054</b> <b>(0.026)</b>	

## DDD in Regression

$$Y_{ijt} = \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{jj} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}$$

- The DDD estimate is the difference between the DD of interest and a placebo DD (which is supposed to be zero)
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias
- If the placebo DD is zero, then DD and DDD give the same results but DD is preferable because standard errors are smaller for DD than DDD
- But now you have multiple parallel trends assumption - both the control group trends are good counterfactuals, and within-state placebo trends for within-state treatment unit counterfactual trends

## Implementing DDD

- Have to get the structure of the data correct because now you have (1) before and after, (2) treatment and control states, and (3) within state placebo
- I give an example in my Mixtape (p. 278) looking at abortion legalization's effect on longterm risky sexual behavior, including do file
- Let's review first the paper, then work through the exercise itself using data.

# The Long-run Effect of Abortion on Sexually Transmitted Infections

---

Christopher Cornwell, *University of Georgia*, and Scott Cunningham,  
*Baylor University*

Send correspondence to: Scott Cunningham, Department of Economics, Baylor University, One Bear Place #98003, Waco, TX 76798-8003, USA; Tel: 254-710-4753; Fax: 254-710-6142; E-mail: scott\_cunningham@baylor.edu

There is a growing literature on the effects of abortion legalization on a range of fertility outcomes. The now-famous paper by Donohue and Levitt [2001. “The Impact of Legalized Abortion on Crime,” 116 *Quarterly Journal of Economics* 379–420], linking abortion to the decline in crime in the 1990s, has shifted the focus to non-fertility outcomes. We focus on STIs, specifically gonorrhea, exploiting the states that legalized abortion prior to *Roe v. Wade* as a quasi-experiment. Using data from the CDC,

**Figure:** Longrun effects of abortion legalization on Risky Sex

## Motivation

- Legalization caused teen childbearing to fall by 12% (Levine 2004)
- Gruber, et al. (1999) showed that the marginal child would have been 60% more likely to live in a single-parent household, 50% more likely to live in poverty, and 45% more likely to be a recipient of public services
- Mechanism was believed to be non-random selection associated with high risk conditions

## Emerging influence

- Donohue and Levitt (2001) linked abortion legalization to declining crime in the 1990s, one of several reasons given for his John Bates Clark award
- Freakonomics popularizes the sensational theory
- Other papers followed like Charles and Stephens (2006) who find that children exposed *in utero* to legalization were less likely to use illegal substances

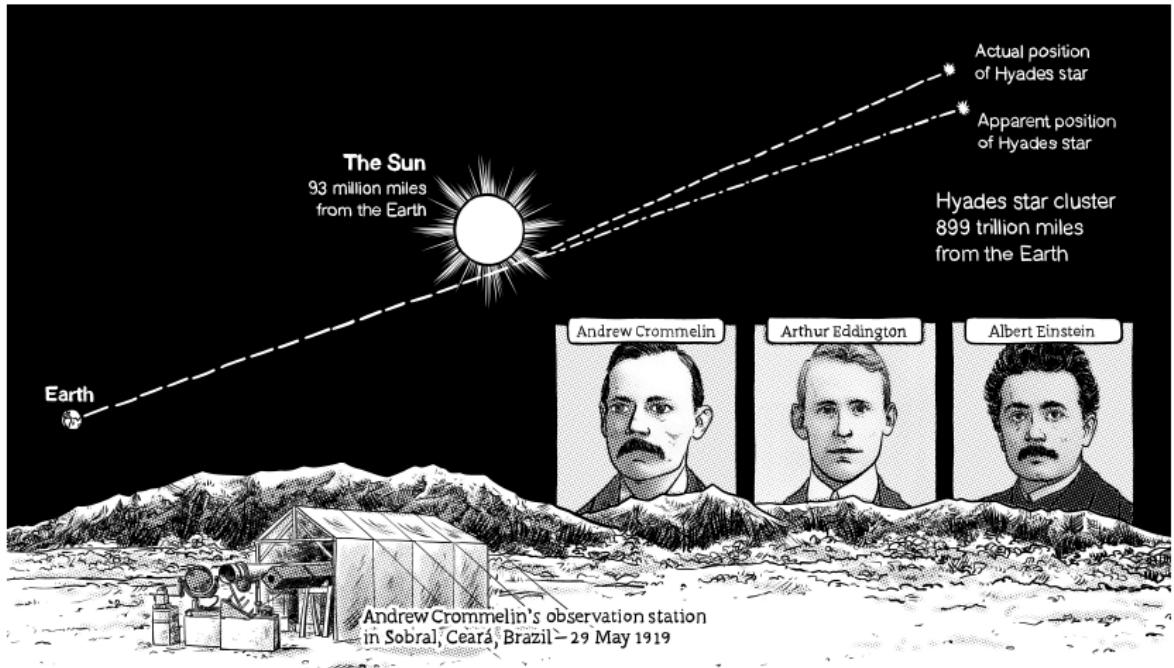
## Controversy

- Triple diff by Joyce finds no evidence for it when using an (arbitrary) cutoff of the median abortion rate within early repeal treatment states
- Foote and Goetz (2008) argue the abortion ratio was constructed incorrectly, and report a coding error leaving out state-year fixed effects; construction problem destroys results, state-year fixed effects somewhat attenuates
- Literature stops and theory is ignored

## In defense of Steve Levitt

- I want to remind people though: we only know about the coding error bc Levitt posted his do files and gave them to anyone who asked (very easy to “lose do files”)
- Levitt had and has oodles of scientific integrity for his willingness to cooperate; not always the case

*"If abortion lowers homicide rates by 20 – 30%, then it is likely to have affected an entire spectrum of outcomes associated with well-being: infant health, child development, schooling, earnings and marital status. Similarly, the policy implications are broader than abortion. Other interventions that affect fertility control and that lead to fewer unwanted births – contraception or sexual abstinence – have huge potential payoffs. In short, a causal relationship between legalized abortion and crime has such significant ramifications for social policy and at the same time is so controversial, that further assessment of the identifying assumptions and their robustness to alternative strategies is warranted." Ted Joyce in his triple diff paper*



**Figure:** Light bending around the sun, predicted by Einstein, and confirmed in a natural experiment involving an eclipse. Artwork by Seth Hahne ©.

## In defense of falsifiable predictions

- Theories which make falsifiable predictions (comparative statics) are *more* convincing of causal effects than simpler reduced form studies
- Great paper by Coleman on (2019) Snow's rhetoric in his 1849 essay and his 1855 book on cholera – mounts different data to make his argument, some of which is of this nature
- Those predictions are threefold:
  - Where we should find effects
  - Where we should not find effects
  - The kind of effects we should find
- If all three are met, an identified causal effect becomes epistemologically more credible

## Falsifiable predictions contained in a diff-in-diff

		CDC Surveillance Data in Calendar Year																	
		1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000		
Age in calendar year		15	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	
15		69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84		
16		68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83		
17		67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82		
18		66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81		
19		65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80		
20		63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78		
21		62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77		
22		61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76		
23		60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75		
24		59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74		
25		58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73		
26		57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72		
27		56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71		
28		0	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5		
29		0	0	0	0	1	2	3	4	5	5	5	5	5	5	5	5		
Number of cohorts (age 15-19 exposed, reforms in 71,74)	Repeal (1)	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	
	No Repeal (2)	1	2	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	
	Difference (3)	0	1	2	3	3	3	1	0	0	0	0	0	0	0	0	0	0	

Figure: Group-time differential exposure predicts a temporary parabolic ATT

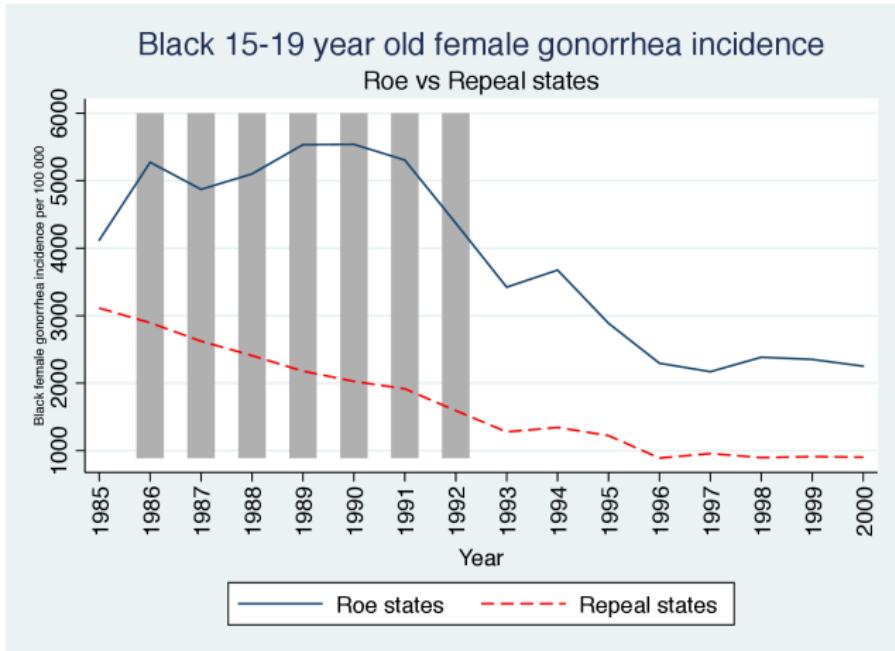


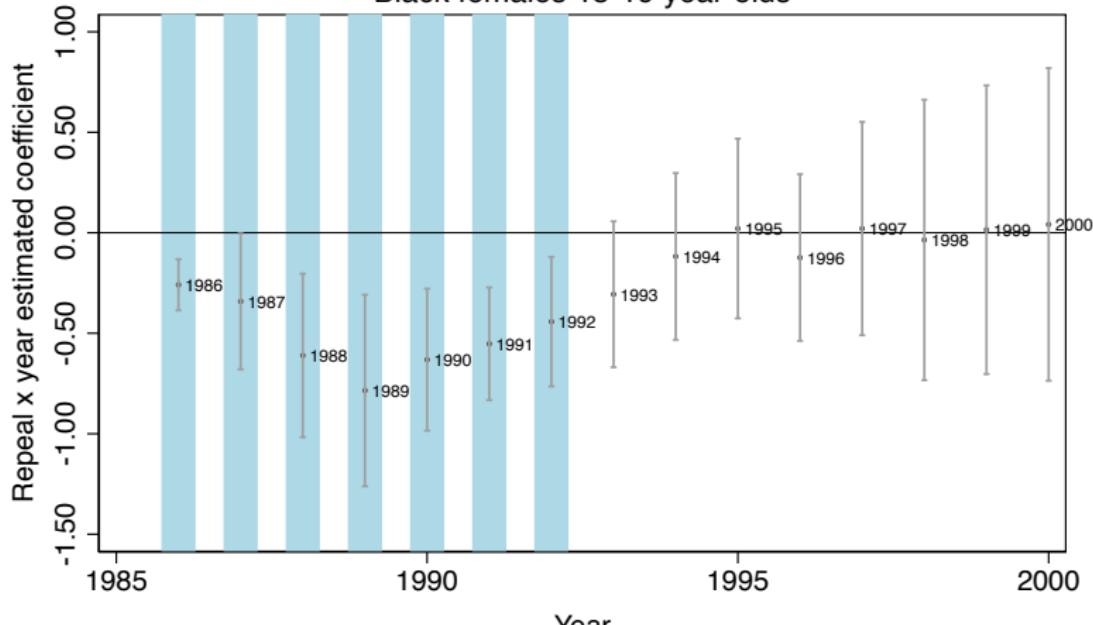
Figure: Raw data for repeal and Roe states.

## Estimating equation

$$\begin{aligned} Y_{st} = & \beta_1 Repeals + \beta_2 DT_t + \beta_3 Repeal_s \times DT_t + X_{st}\psi + \alpha_s DS_s \\ & + \gamma_1 t + \gamma_2 s \times t + \varepsilon_{st} \end{aligned}$$

## Estimated effect of abortion legalization on gonorrhea

Black females 15-19 year-olds



Whisker plots are estimated coefficients of DD estimates

**Figure:** Differences in black female gonorrhea incidence between repeal and Roe cohorts.

## Assuaging doubt

- Maybe spurious - something happened in those years, but what?
- Crack epidemic maybe? But we control for the crack index by Fryer, et al.
- Maybe something else - let's try a within-state control group (the older cohort)

## DDD Equation

$$\begin{aligned}Y_{ast} = & \beta_1 Repeal_s + \beta_2 DT_t + \beta_3 DA + \beta_{4t} Repeal_s \cdot DT_t + \\& + \beta_5 Repeal_s \cdot DA + \beta_{6t} DA \cdot DT_t + \beta_{7t} Repeal_s \cdot DA \cdot DT_t \\& + X_{st}\xi + \alpha_{1s} DS_s + \alpha_{2s} DS_s \cdot DA + \gamma_1 t + \gamma_{2s} DS_s \cdot t + \gamma_3 DA \cdot t \\& + \gamma_{4s} DS_s \cdot DA \cdot t + \epsilon_{ast}\end{aligned}$$

One will be dropped, but I want to focus your attention on the number of interactions needed to identify DDD parameters

# Stacking Structure

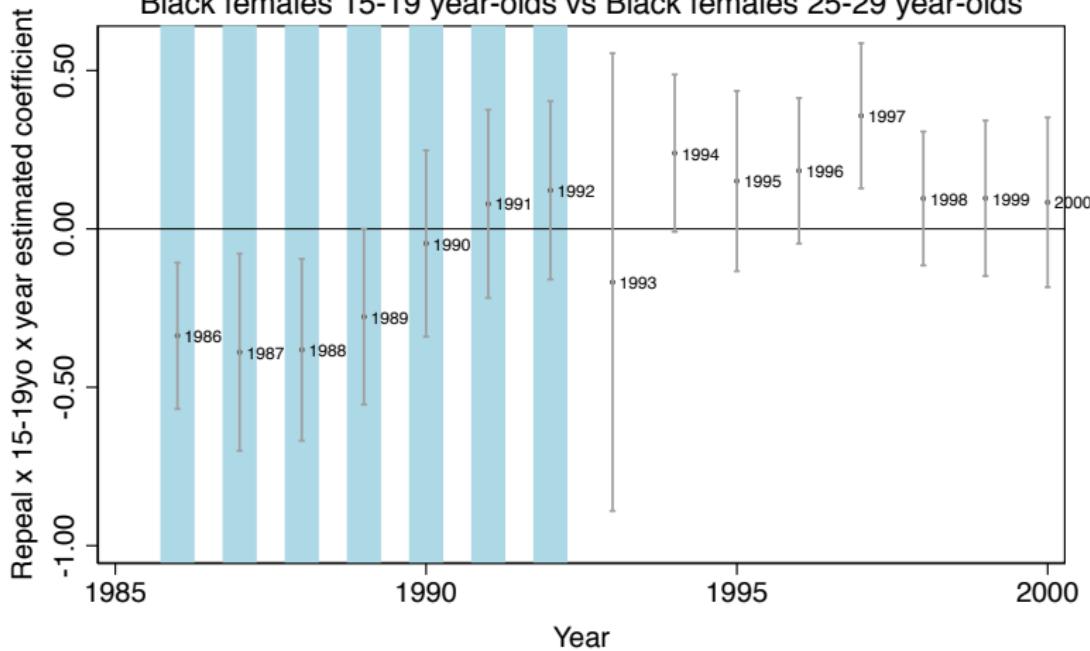
```
. list id wm wf bf bm age repeal year in 1/30
```

	id	wm15	wf15	bf15	bm15	age	repeal	year
1.	1	1	0	0	0	15	0	1985
2.	1	1	0	0	0	15	0	1986
3.	1	1	0	0	0	15	0	1987
4.	1	1	0	0	0	15	0	1988
5.	1	1	0	0	0	15	0	1989
6.	1	1	0	0	0	15	0	1990
7.	1	1	0	0	0	15	0	1991
8.	1	1	0	0	0	15	0	1992
9.	1	1	0	0	0	15	0	1993
10.	1	1	0	0	0	15	0	1994
11.	1	1	0	0	0	15	0	1995
12.	1	1	0	0	0	15	0	1996
13.	1	1	0	0	0	15	0	1997
14.	1	1	0	0	0	15	0	1998
15.	1	1	0	0	0	15	0	1999
16.	1	1	0	0	0	15	0	2000
17.	2	0	1	0	0	15	0	1985

# DDD Results

## Estimated effect of abortion legalization on gonorrhea

Black females 15-19 year-olds vs Black females 25-29 year-olds



Whisker plots are estimated coefficients of DDD coefficients

## My original conclusions

- Model made narrow predictions of a *parabola* within a given window but only for the treatment cohort
- Amazingly we actually found that very shape in the DD – did we vindicate Gruber, et al. and Donohue and Levitt then?
- Also used older group as within-state controls in a DDD, and still found the parabola, though not as great a look as DD which is a bit of a red flag
- Paper also illustrates the usefulness of having a specific theoretical prediction. Limits the number of competing hypotheses (Popperian type of reasoning).
- But was I done? Look back at the table

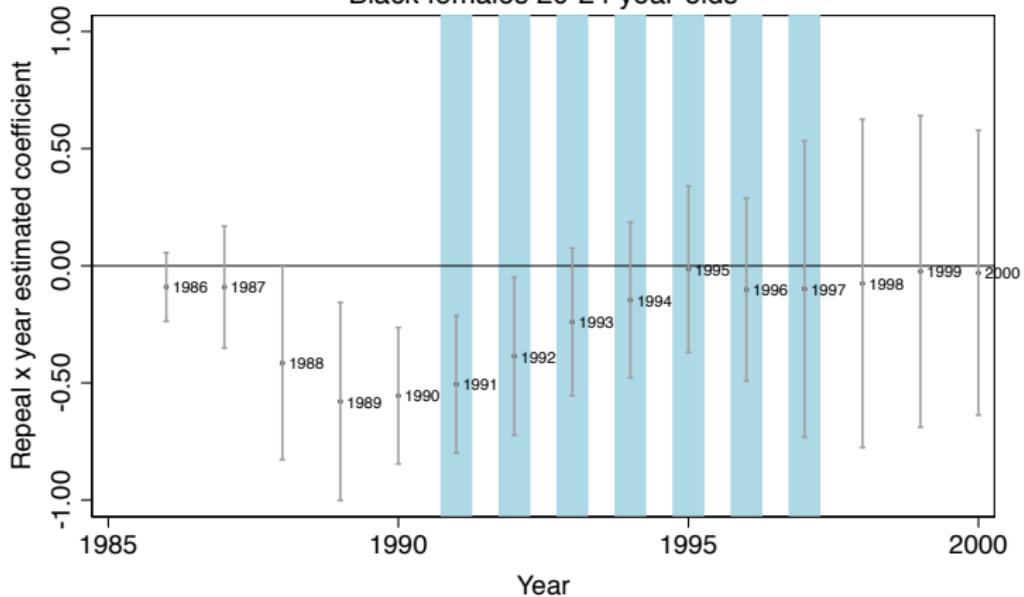
# Going beyond Cornwell and Cunningham (2013)

	CDC Surveillance Data in Calendar Year																		
Age in calendar year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000			
15	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85			
16	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84			
17	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83			
18	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82			
19	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81			
20	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80			
21	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79			
22	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78			
23	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77			
24	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76			
25	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75			
26	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74			
27	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73			
28	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72			
29	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71			
Number of cohorts (age 20-24) exposed, reforms in 71, 74	Repeal (1)																		
	0	0	0	0	0	0	0	1	2	3	4	5	5	5	5	5	5	5	
	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	5	5	5	
Difference (3)	No Repeal (2)																		
	0	0	0	0	0	0	0	1	2	3	3	3	2	1	0	0	0	0	

Figure: Second theoretical prediction - this time for 20-24 year olds

## Estimated effect of abortion legalization on gonorrhea

Black females 20-24 year-olds



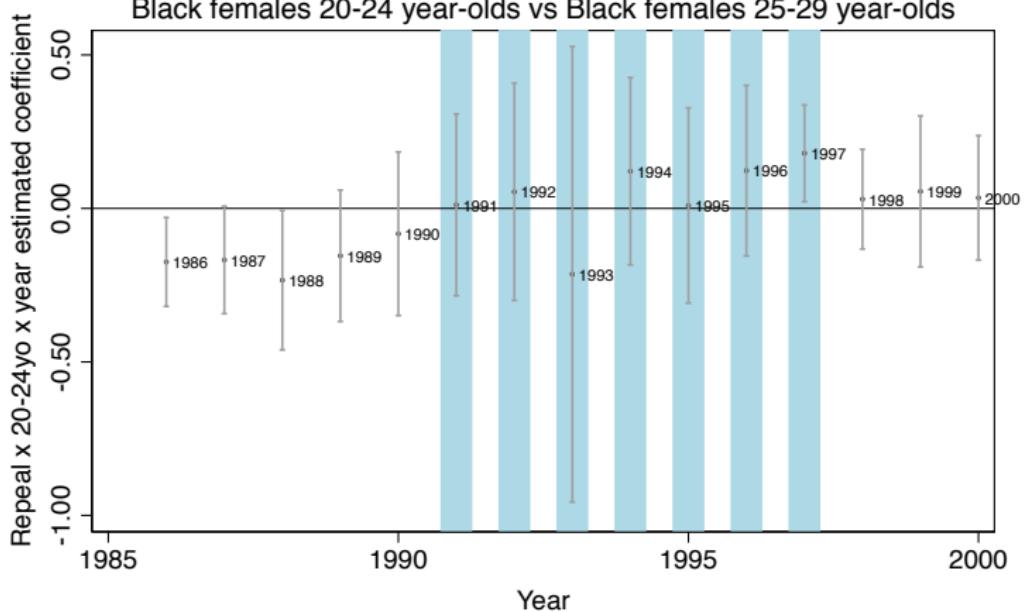
Whisker plots are estimated coefficients of DD estimates

## **Second prediction fails second DD model**

- Ugh. *lo tov* (Hebrew to English: not good)
- Well, maybe DDD will look better?

## Estimated effect of abortion legalization on gonorrhea

Black females 20-24 year-olds vs Black females 25-29 year-olds



Whisker plots are estimated coefficients of DDD coefficients

## Second predictions fails DDD too

- Notice that when we exploited just one testable prediction, we found evidence
- But when we exploit all of the testable predictions, the results fall apart, suggesting original DD was spurious
- Imagine for a moment, though – what if we had seen the group-time ATT moving with the cohort as they aged?
- Other alternative is the repeal-Roe effects dissipate by early to late 20s, but what does Ockham's Razor say is the more credible explanation?
- Perhaps the Gruber, et al. (1999) and Donohue and Levitt (2001) hypothesis was always spurious

## **Stata replication**

Let's replicate this using the abortion.do file. Pay close attention to the stacking of the data by group-state, not just state, and the exact way in which the interactions must therefore be constructed

## 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

Hidden curriculum  
Foundational causality stuff  
Regression discontinuity designs  
Instrumental variables  
Two-way fixed effects estimator  
**Differences-in-differences**  
Comparative case studies  
Matching and weighting  
Concluding remarks

John Snow  
The simple 2x2  
Event study  
Differential timing

I ❤️ federalism  
(for the natural experiments)

Tweets 30.4K Following 5,933 Followers 11.8K Likes 80.5K Lists 1 Moments 0

Edit profile

## State federalism and differential timing

- We've been considering situations where treatment occurs in one area for the most part
- But the modal situation is when there is *differential timing*
- 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
- Example might be the minimum wage though we will look at others

## 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 Bertand, 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													
Castle Doctrine Law	0.0780*** (0.0255)	0.0290 (0.0236)	0.0223 (0.0223)	0.0164 (0.0247)	0.0327* (0.0165)	0.0237 (0.0207)	Log (Burglary Rate)	0.0572** (0.0272)	0.00961 (0.0291)	0.00663 (0.0268)	0.00277 (0.0304)	0.00683 (0.0222)	0.0207 (0.0259)
One Year Before Adoption of				-0.0201								-0.0154 (0.0214)	
Castle Doctrine Law				(0.0139)									
Panel B: Robbery					Log (Robbery Rate)								
Castle Doctrine Law	0.0408 (0.0254)	0.0344 (0.0224)	0.0262 (0.0229)	0.0216 (0.0246)	0.0376** (0.0181)	0.0515* (0.0274)	Log (Robbery Rate)	0.0448 (0.0331)	0.0320 (0.0421)	0.00839 (0.0387)	0.00552 (0.0437)	0.00874 (0.0339)	0.0267 (0.0299)
One Year Before Adoption of				-0.0156								-0.0115 (0.0283)	
Castle Doctrine Law				(0.0167)									
Panel C: Aggravated Assault					Log (Aggravated Assault Rate)								
Castle Doctrine Law	0.0434 (0.0387)	0.0397 (0.0407)	0.0372 (0.0319)	0.0362 (0.0349)	0.0424 (0.0291)	0.0414 (0.0285)	Log (Aggravated Assault Rate)	0.0555 (0.0604)	0.0698 (0.0630)	0.0343 (0.0433)	0.0305 (0.0478)	0.0341 (0.0405)	0.0317 (0.0380)
One Year Before Adoption of				-0.00343								-0.0150 (0.0251)	
Castle Doctrine Law				(0.0161)									
Observations	550	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	Yes
Region-by-Year Fixed Effects		Yes	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?

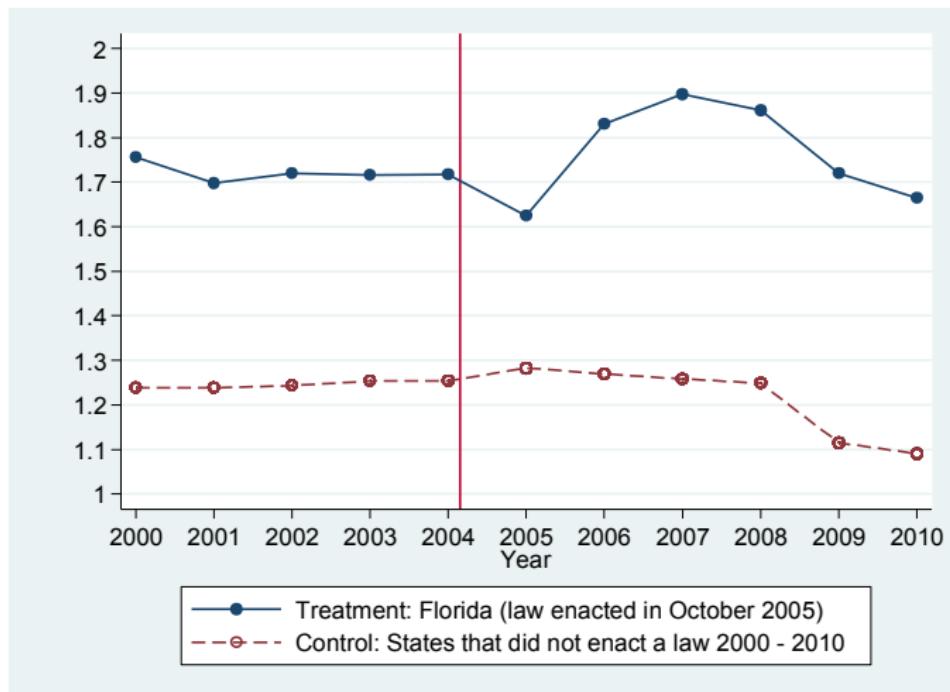
## Answer

- 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

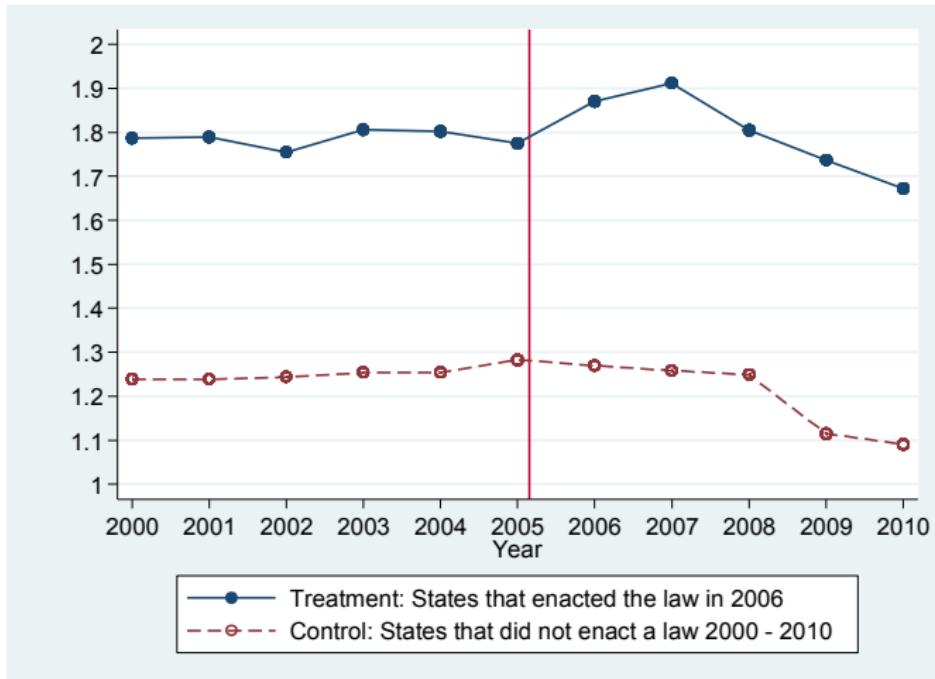
## Parallel pre-treatment trends

- 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

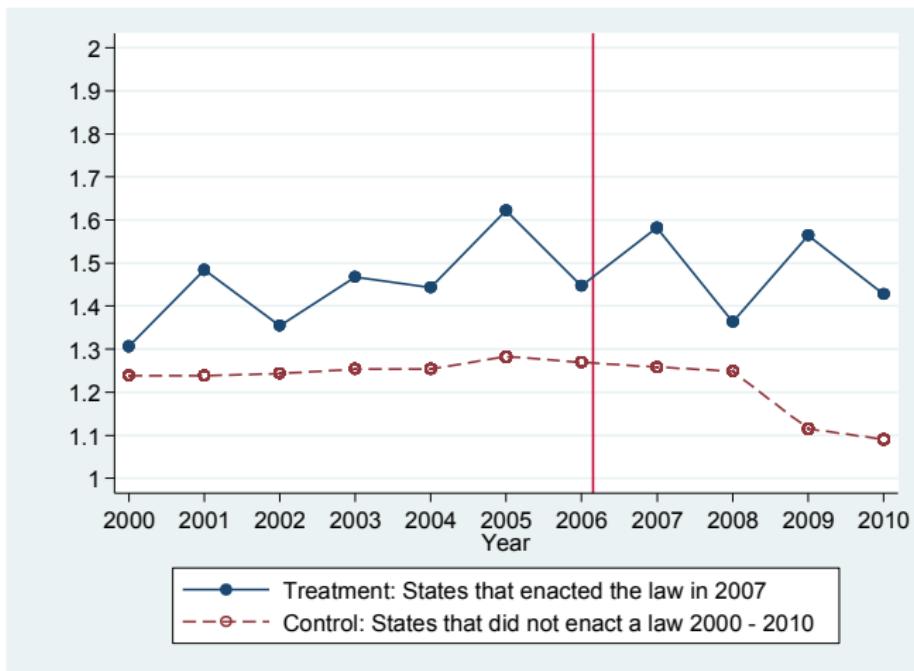
## Log Homicide Rates – 2005 Adopter = Florida



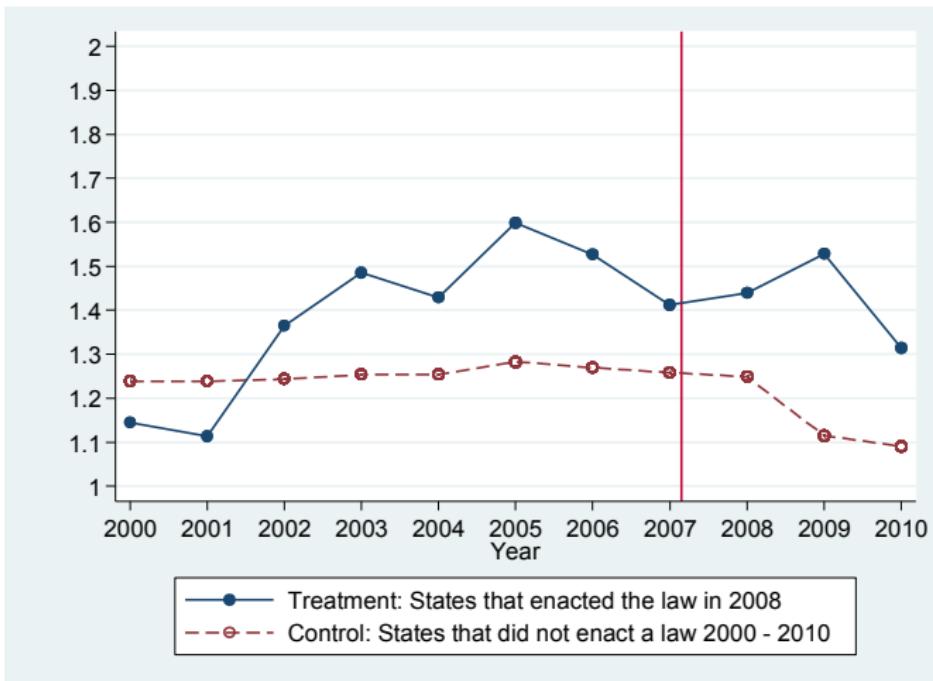
## Log Homicide Rates – 2006 Adopter (13 states)



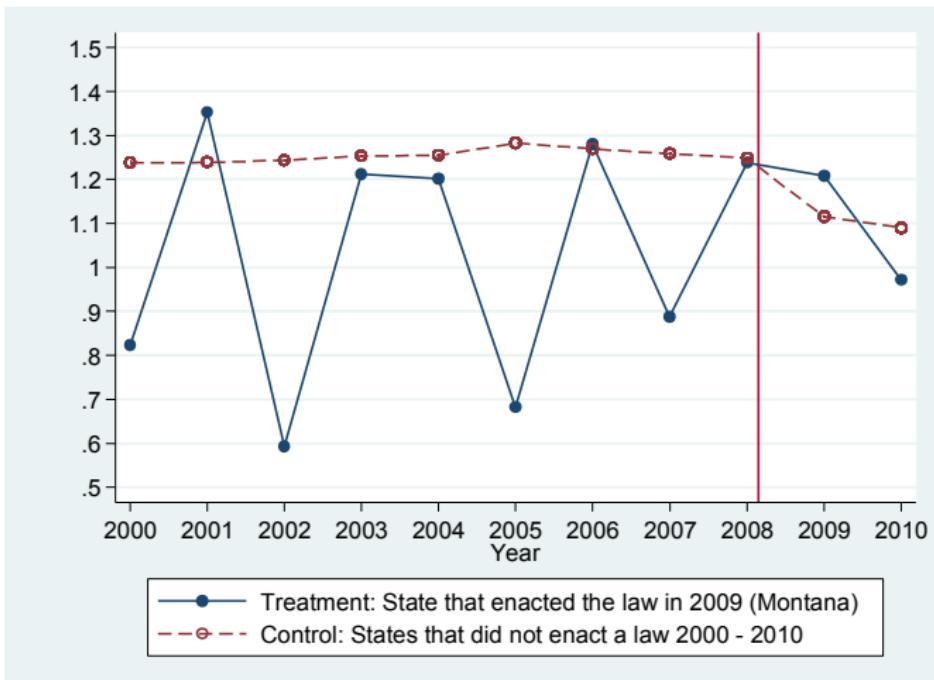
## Log Homicide Rates – 2007 Adopter (4 states)



## 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

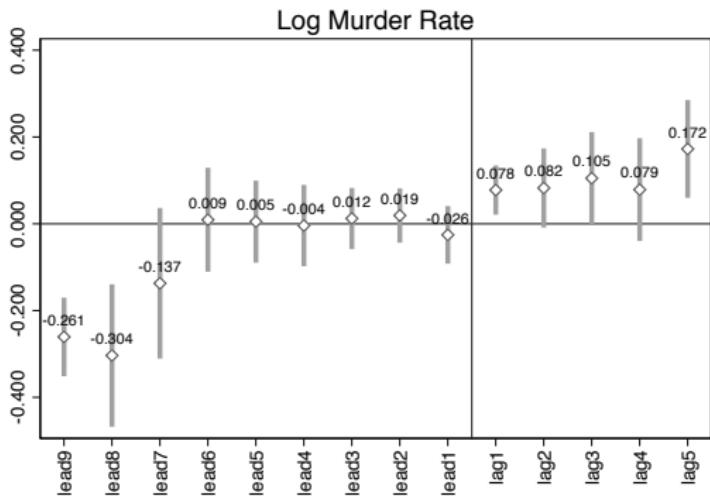


Figure: Homicide event study plots using coefplot

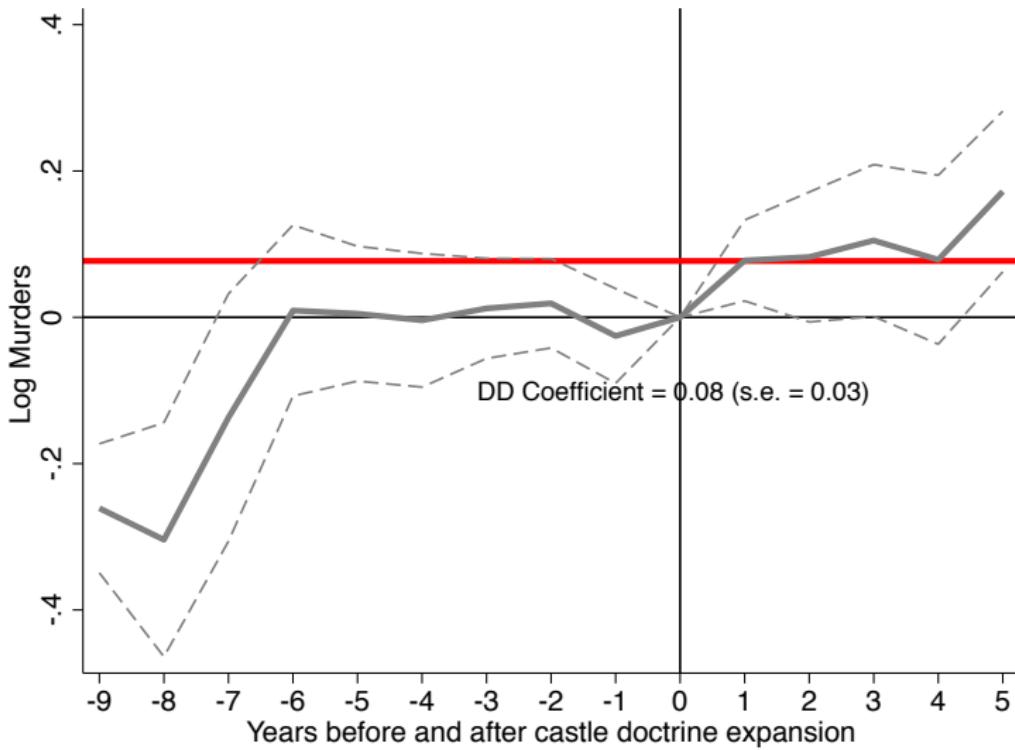


Figure: Homicide event study plots using two way

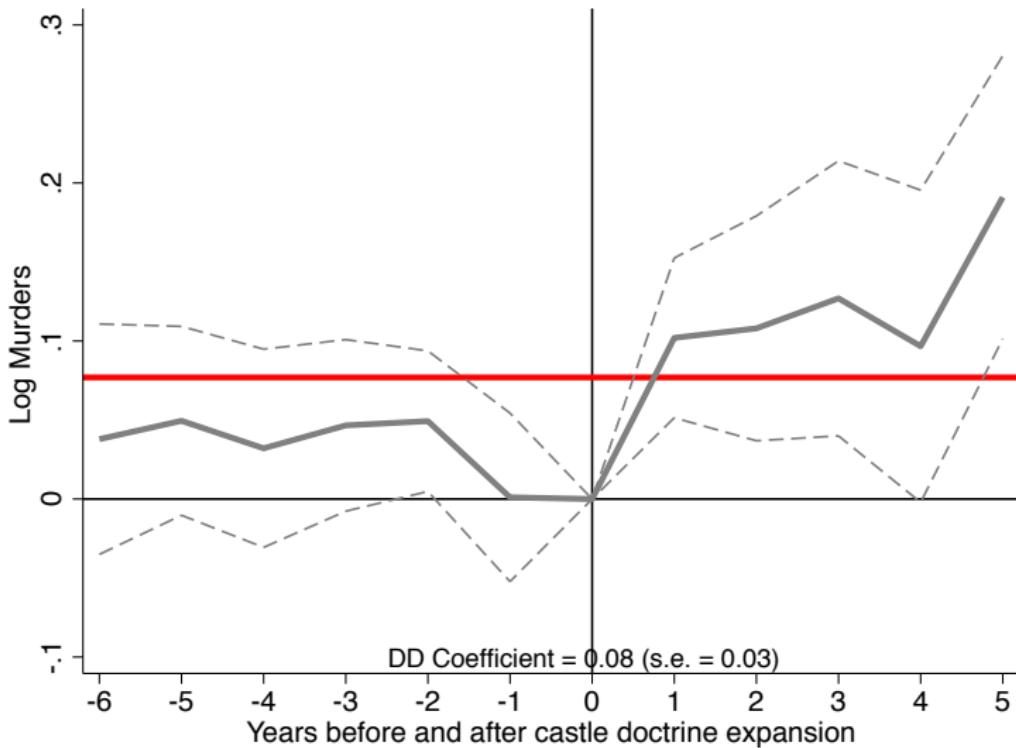


Figure: Homicide event study plots using two way and force early leads into one coefficient

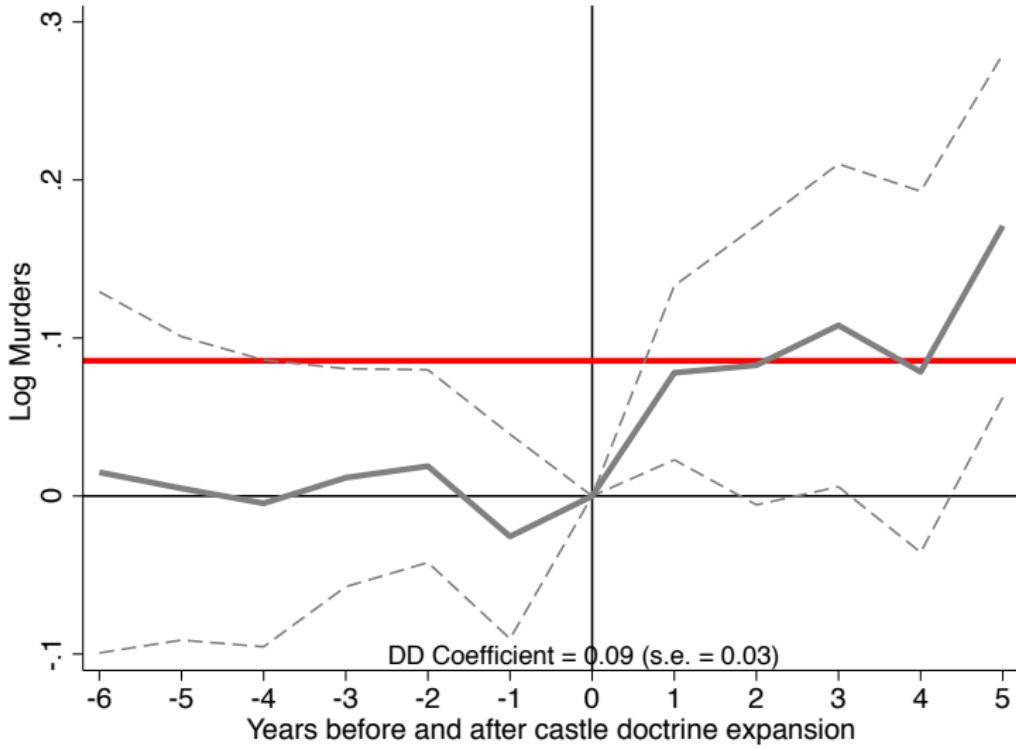


Figure: Homicide event study plots using two-way dropping imbalanced states

## 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

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

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  $2 \times 2$  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  $\widehat{\delta}_{ab}^{2 \times 2, j}$  as the canonical  $2 \times 2$  DD estimator for groups  $a$  and  $b$  where  $j$  is the treatment group
- So what are the possible  $2 \times 2$  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

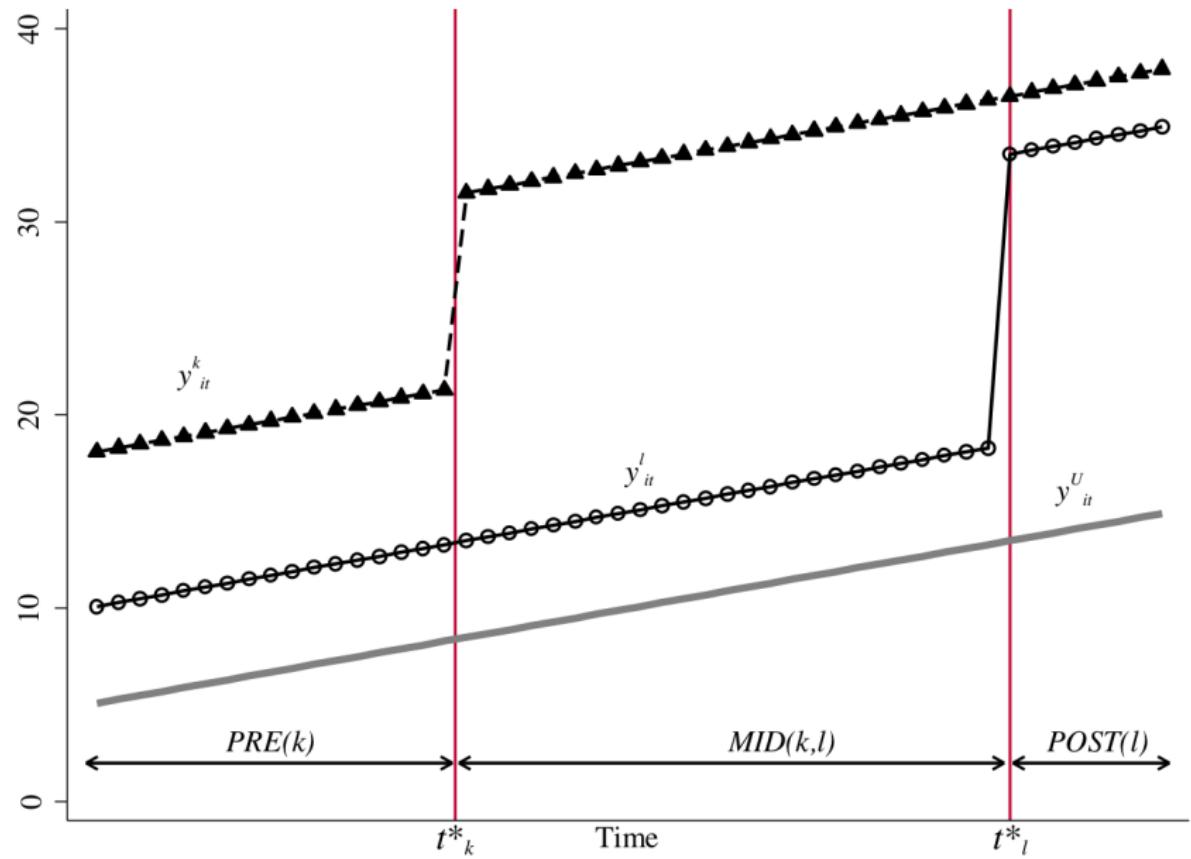
$K^2$  distinct DDs

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”



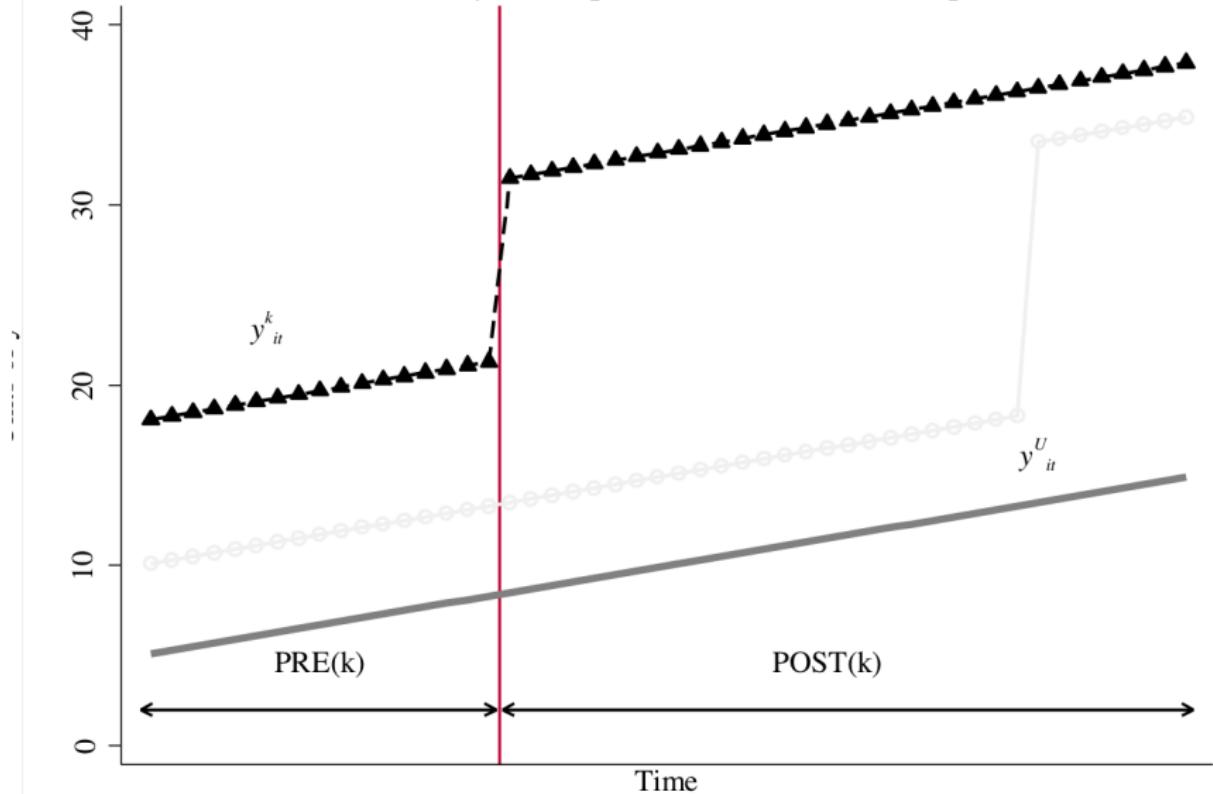
## Three important 2x2 DDs

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

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.

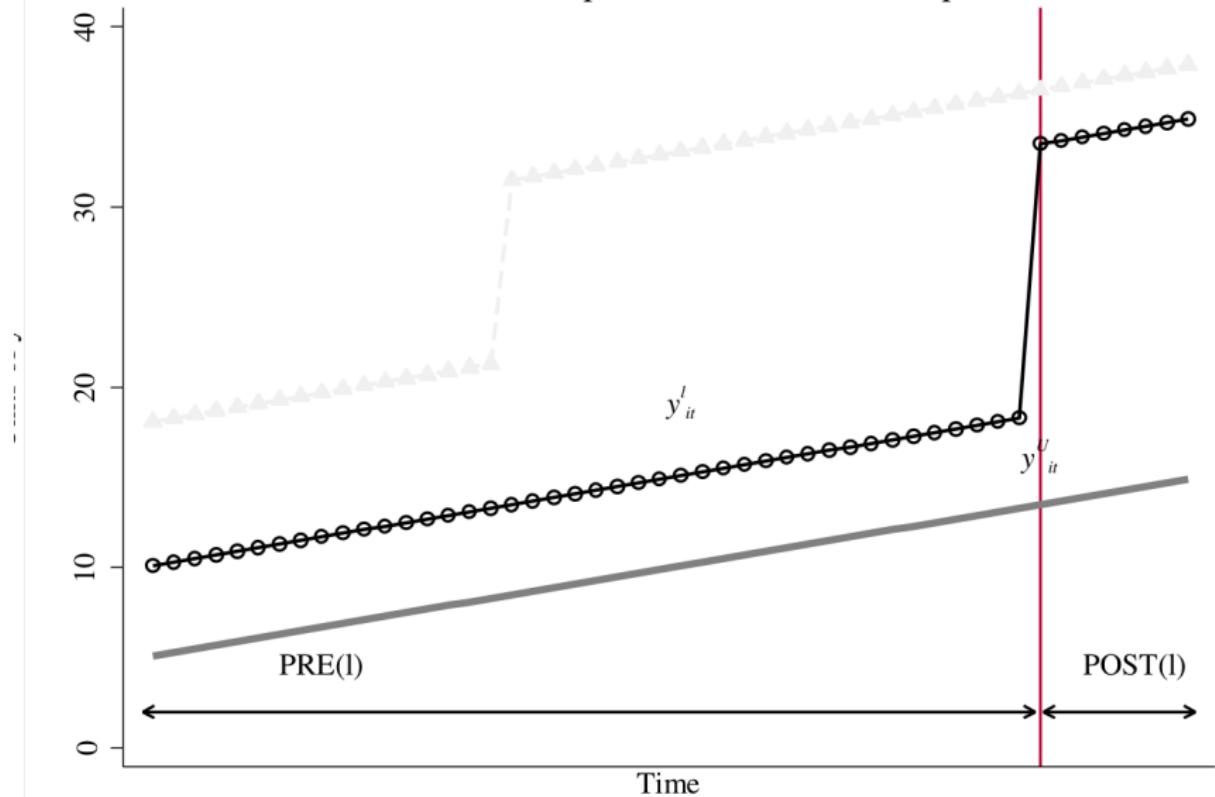
$$\widehat{\delta}_{kU}^{2 \times 2} = \left( \bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left( \bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

A. Early Group vs. Untreated Group

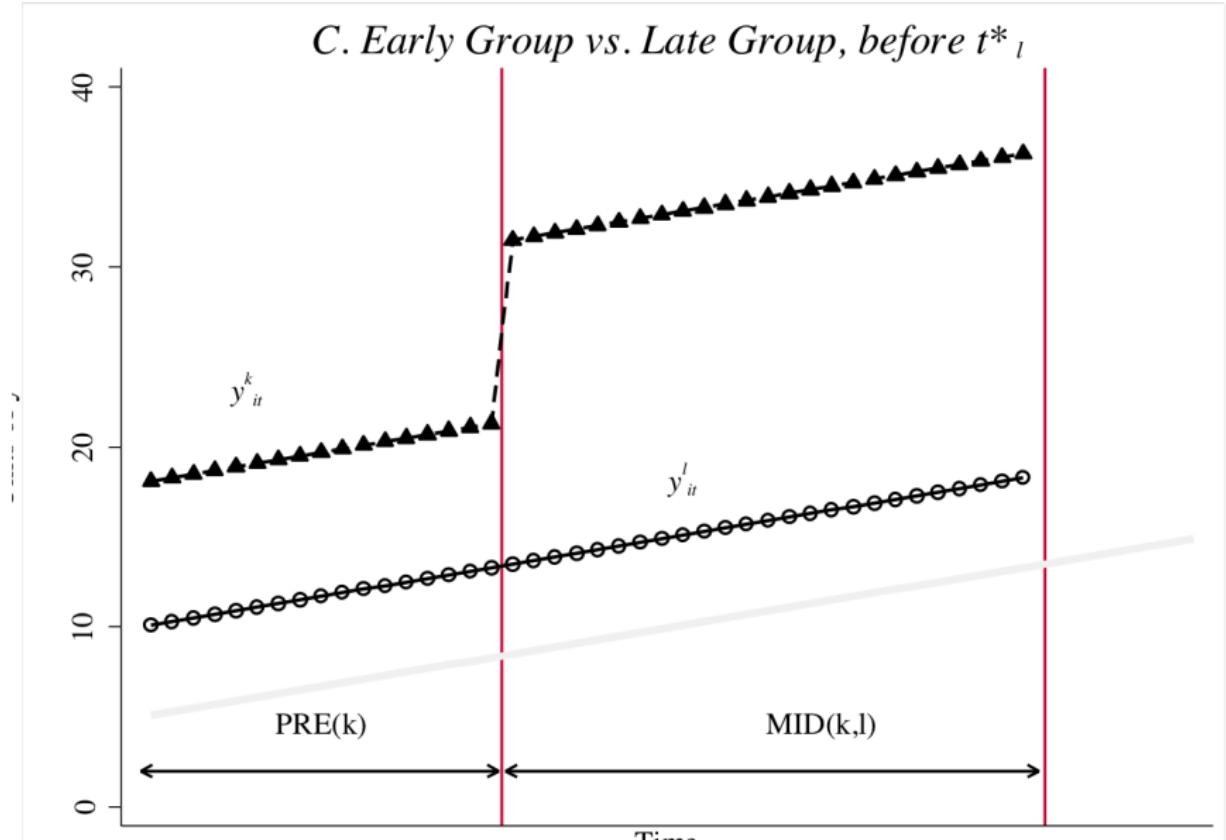


$$\widehat{\delta}_{IU}^{2 \times 2} = \left( \bar{y}_I^{post(I)} - \bar{y}_I^{pre(I)} \right) - \left( \bar{y}_U^{post(I)} - \bar{y}_U^{pre(I)} \right)$$

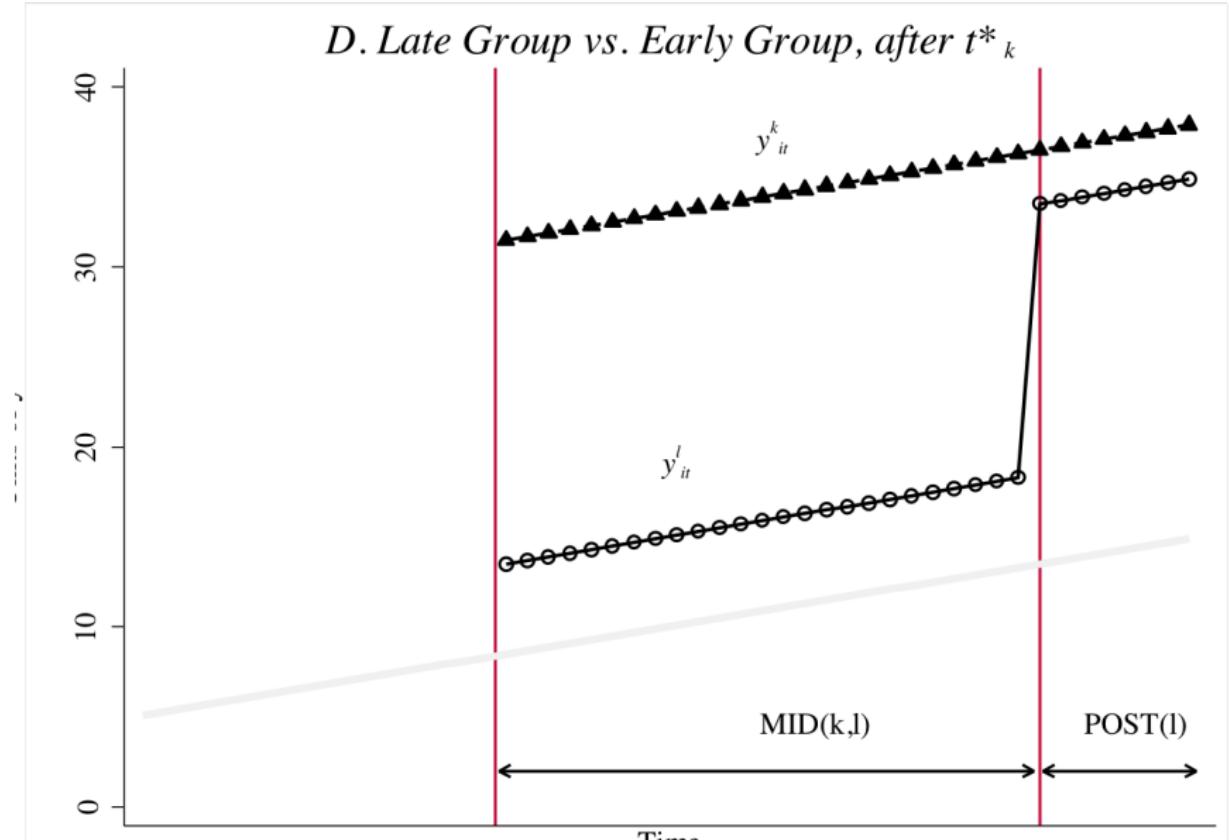
B. Late Group vs. Untreated Group



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



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



## 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)

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

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

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

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  $2 \times 2$  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  $\bar{D}_k - \bar{D}_I$  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  $t_I^* = 0.67$ . Then  $\bar{D}_k - \bar{D}_I = 0.52$ . 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

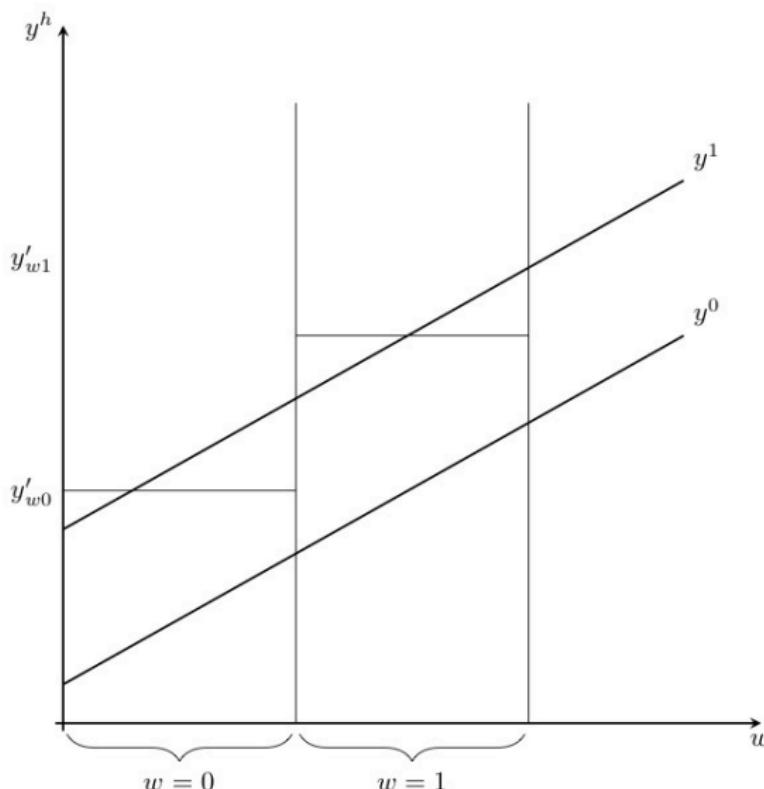


Figure: With trends, differences in mean potential outcomes is non-zero

## From 2x2 to ATT

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

## Potential outcomes

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

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

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

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:

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

## Heterogeneity bias?

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

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

- 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

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

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  $\sigma$  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.

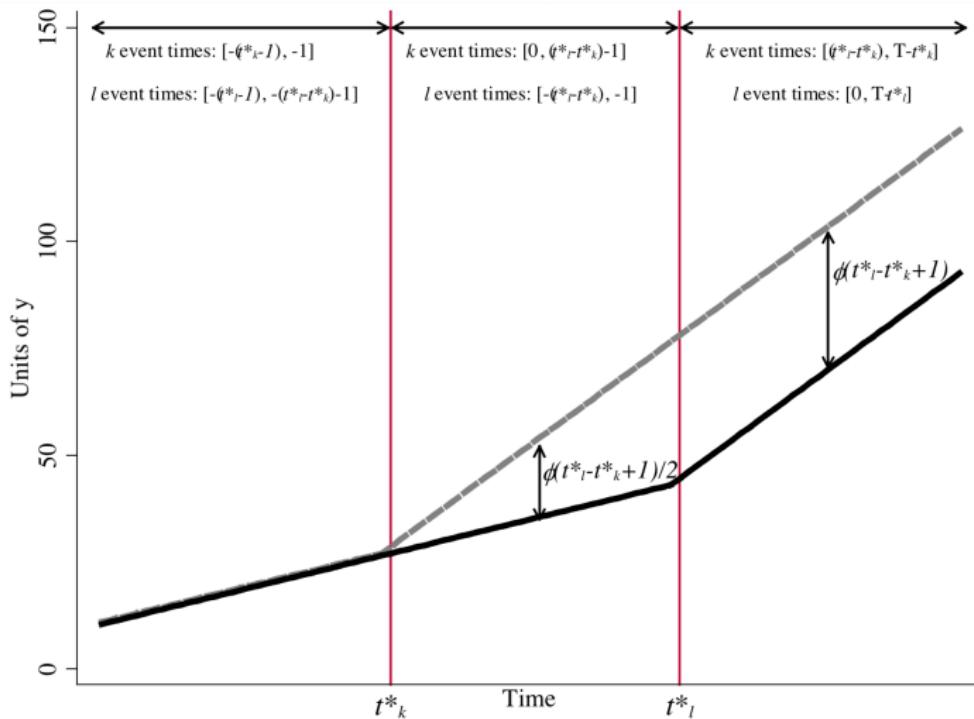
## Case 1 cont.

- 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

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

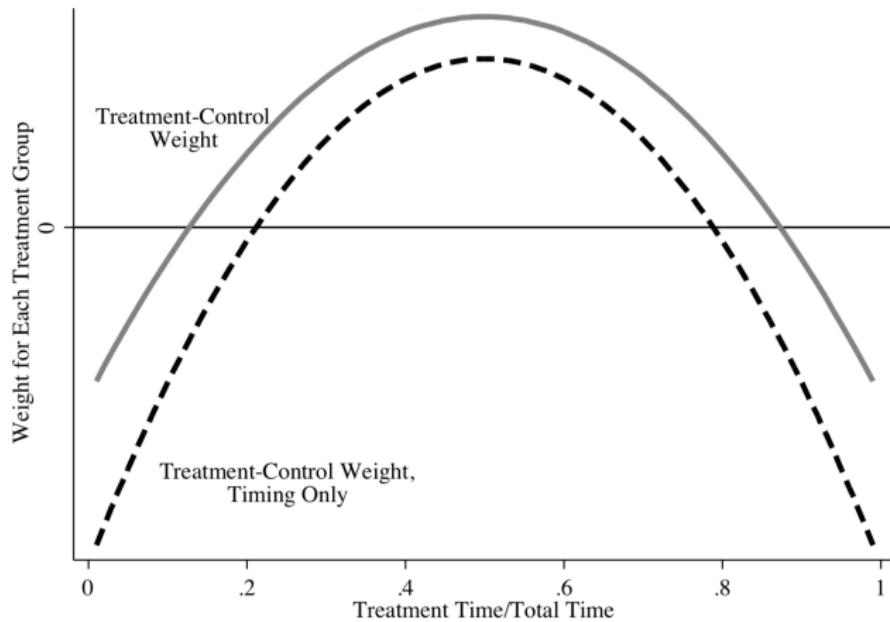
## Variance weighted common trends

- 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

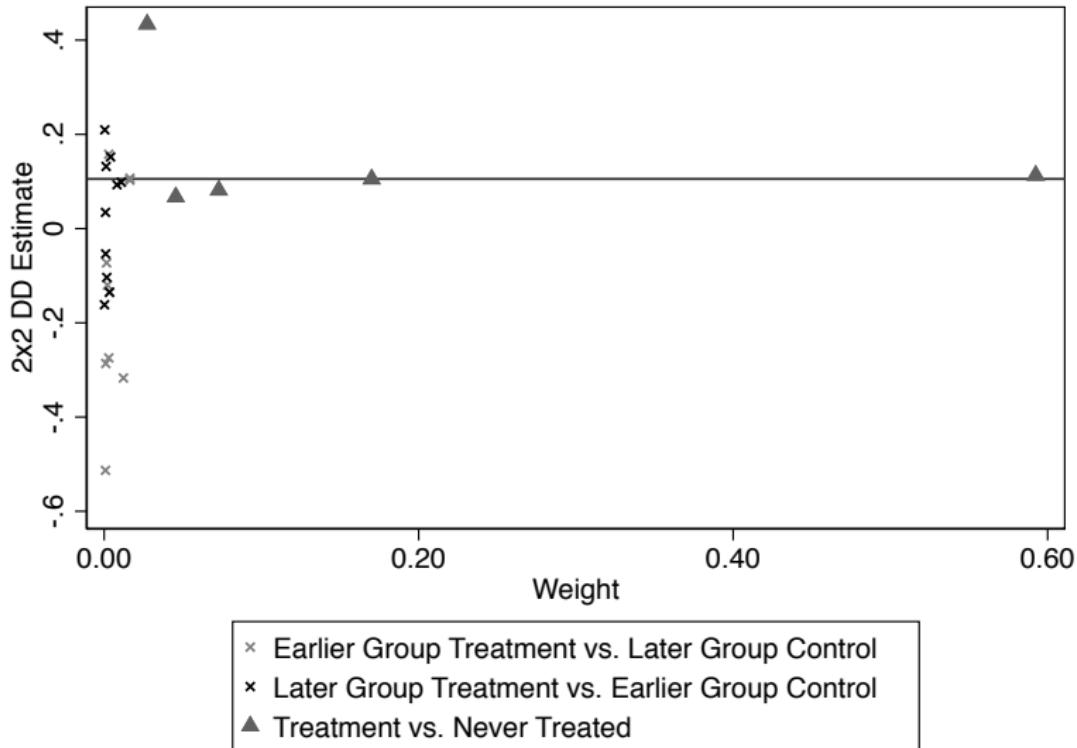
```
. ddtiming l_murder post, i(sid) t(year)
```

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:

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

- 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 shorrun 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

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

## 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

## **Stata example**

See baker.do

## Concluding remarks on DD

- Chances are you are going to write more papers using DD than any other design
- Goodman-Bacon (2018, 2019) is *worth your time* so that you know what you are estimating
- And Callaway and Sant'ann (2019) 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