Causal Inference

Mixtape Session

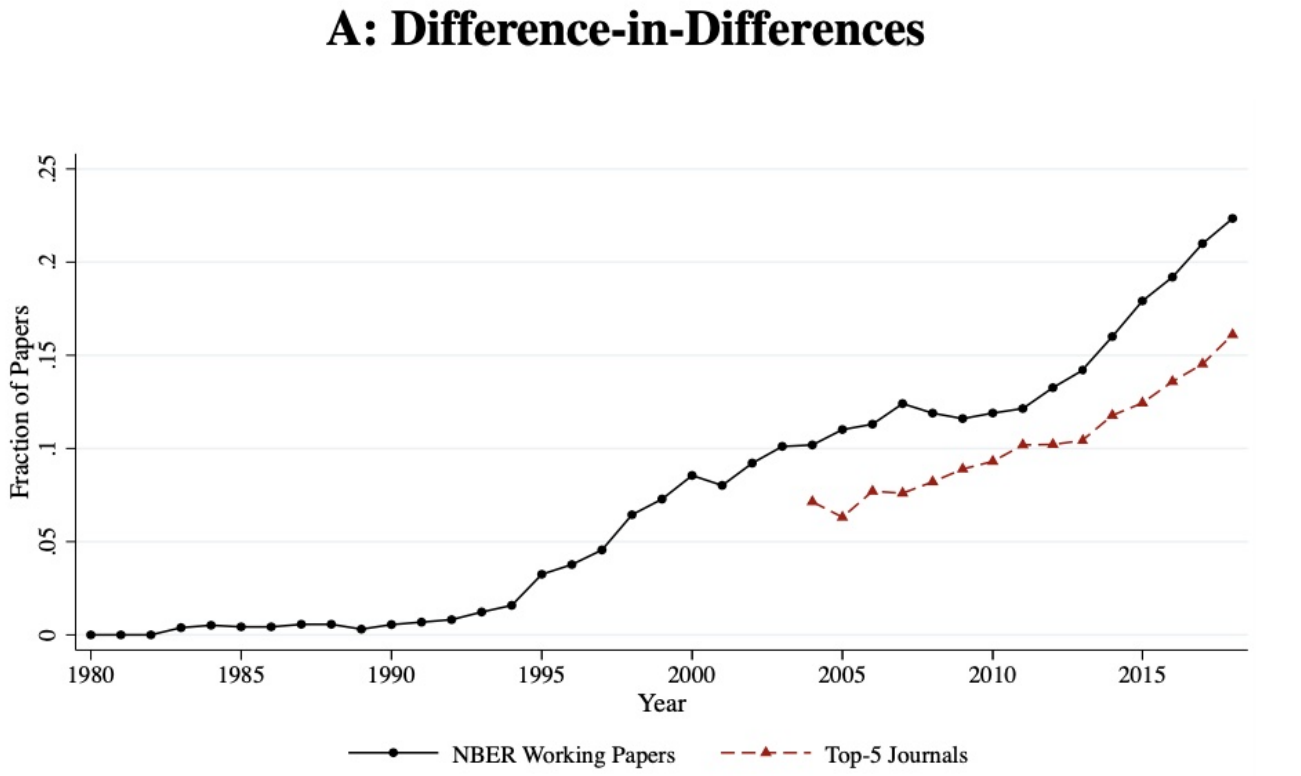
Roadmap

- Difference-in-differences
  - Two group case
  - Verifying assumptions
- Differential timing
  - Twoway Fixed Effects
  - Strict exogeneity
  - Simulation
- Robust DiD with differential timing
  - Implicit imputation
  - Event studies
  - Explicit imputation
- Concluding Remarks

What is difference-in-differences (DiD)

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

Figure: Currie, et al. (2020)



Treatment effect definitions

Individual treatment effect

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

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

Conditional Average Treatment Effects

Average Treatment Effect on the Treated (ATT)

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

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

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

Identification without randomization

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

DiD equation

I call this the DiD equation, but Goodman-Bacon calls it the “2x2”

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

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

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

Potential outcomes and the switching equation

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

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

Parallel trends bias

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

Identification

Parallel trends

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

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

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

Contrast PT with independence

Recall identification of treatment effects with randomized treatment assignment

Independence assumption

Treatment is independent of potential outcomes

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

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

How the science works

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

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

Sample averages

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

Population expectations

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

Potential outcomes and the switching equation

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

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

Parallel trends bias

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

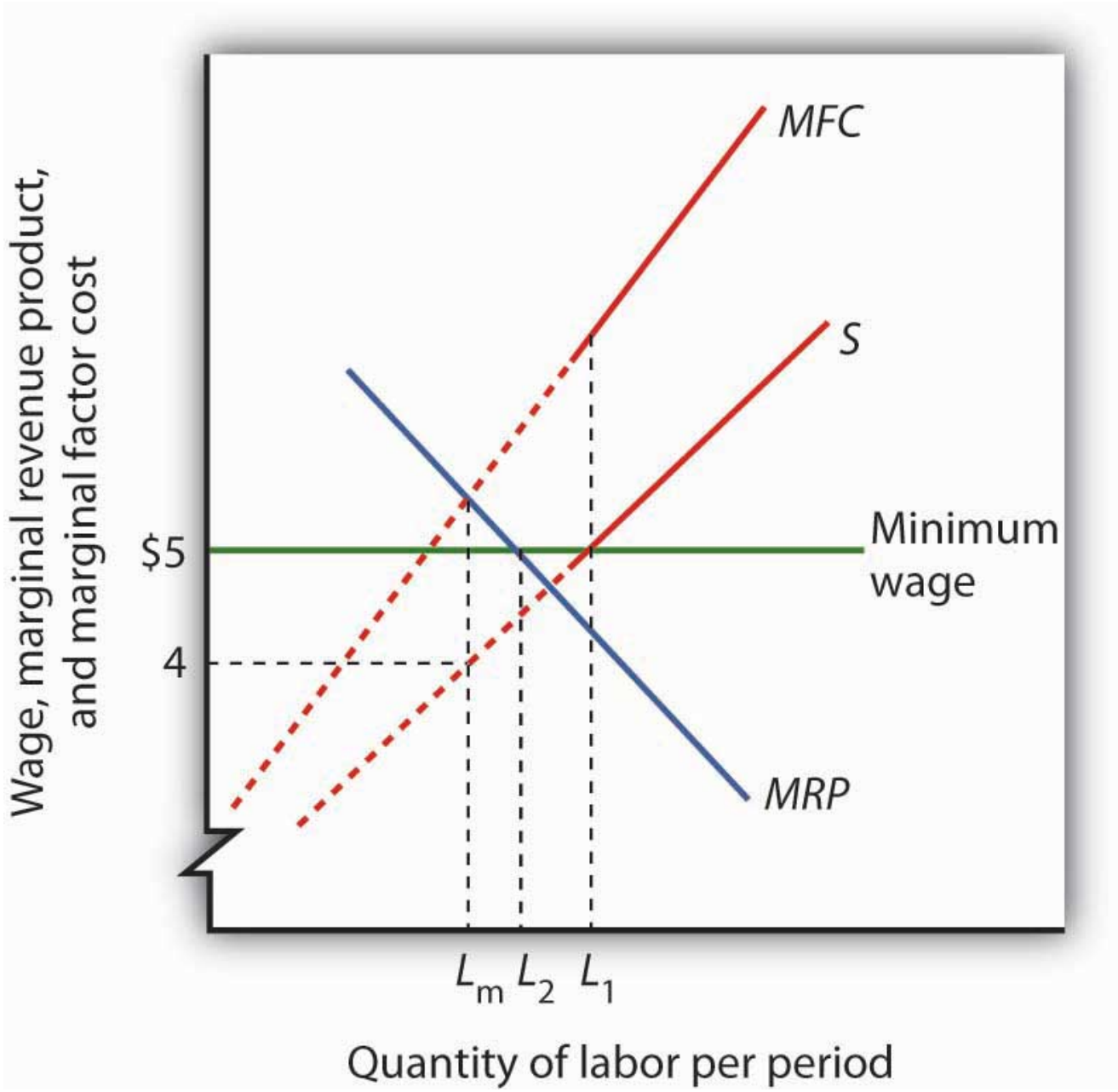
Another famous DD study

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

Competitive vs noncompetitive markets

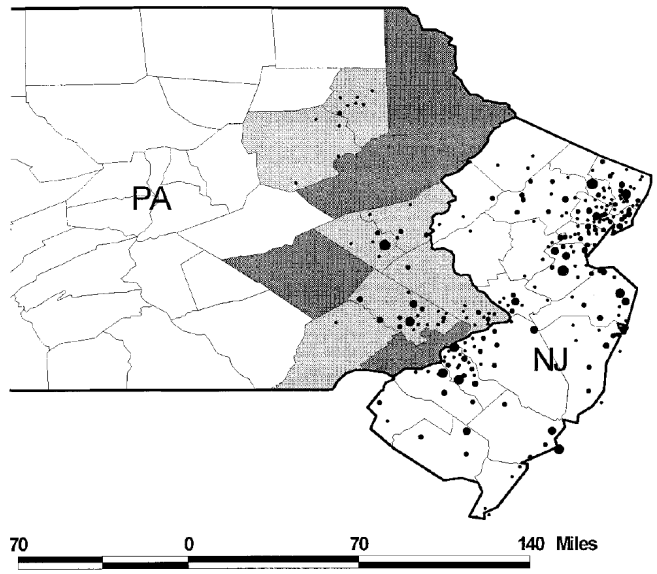
- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In a competitive input market, increases in the minimum wage would move us up a downward sloping labor demand curve $\rightarrow$ employment would fall
- Monopsony (imperfect labor markets)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



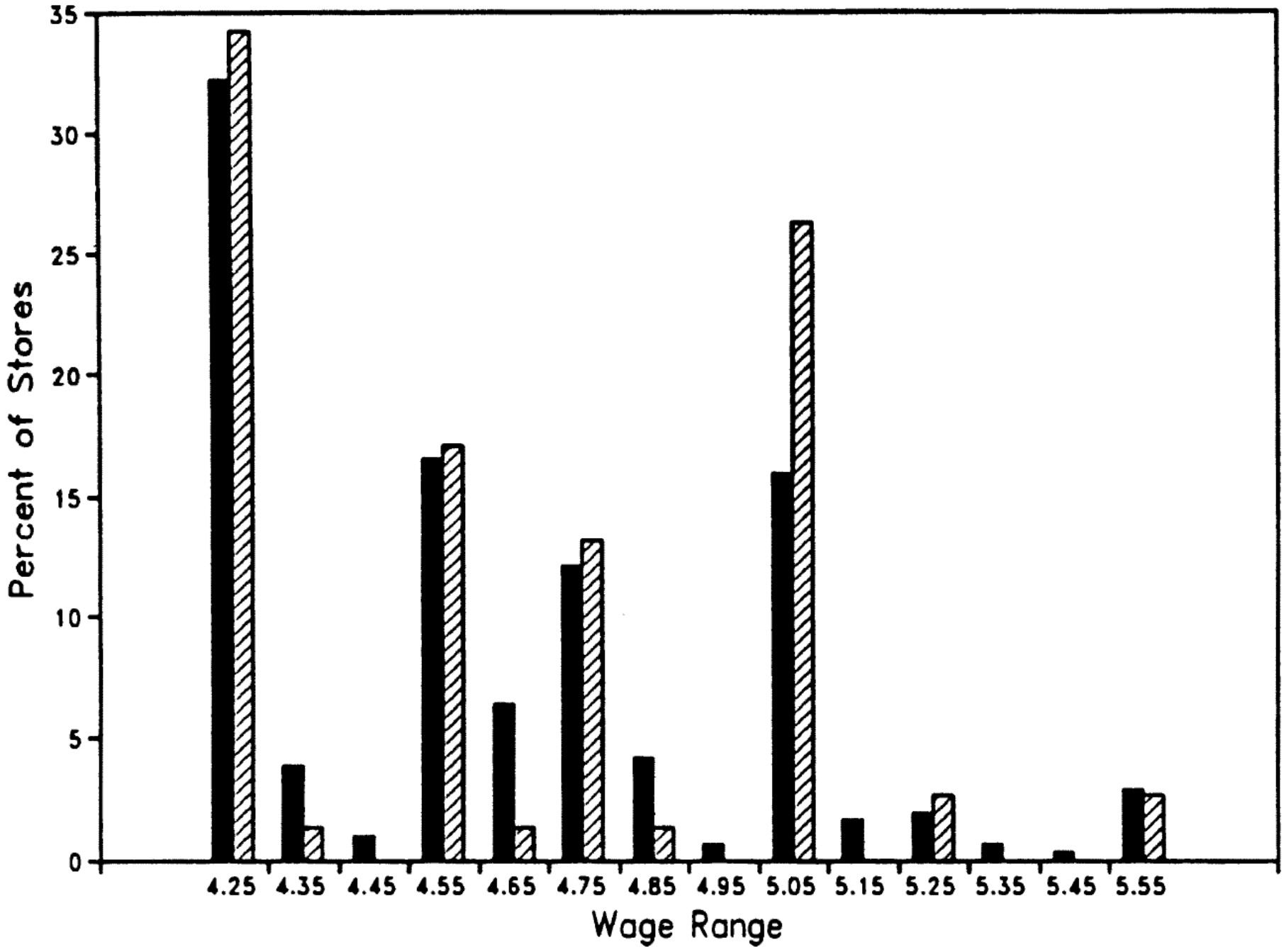
- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey - shoeleather!

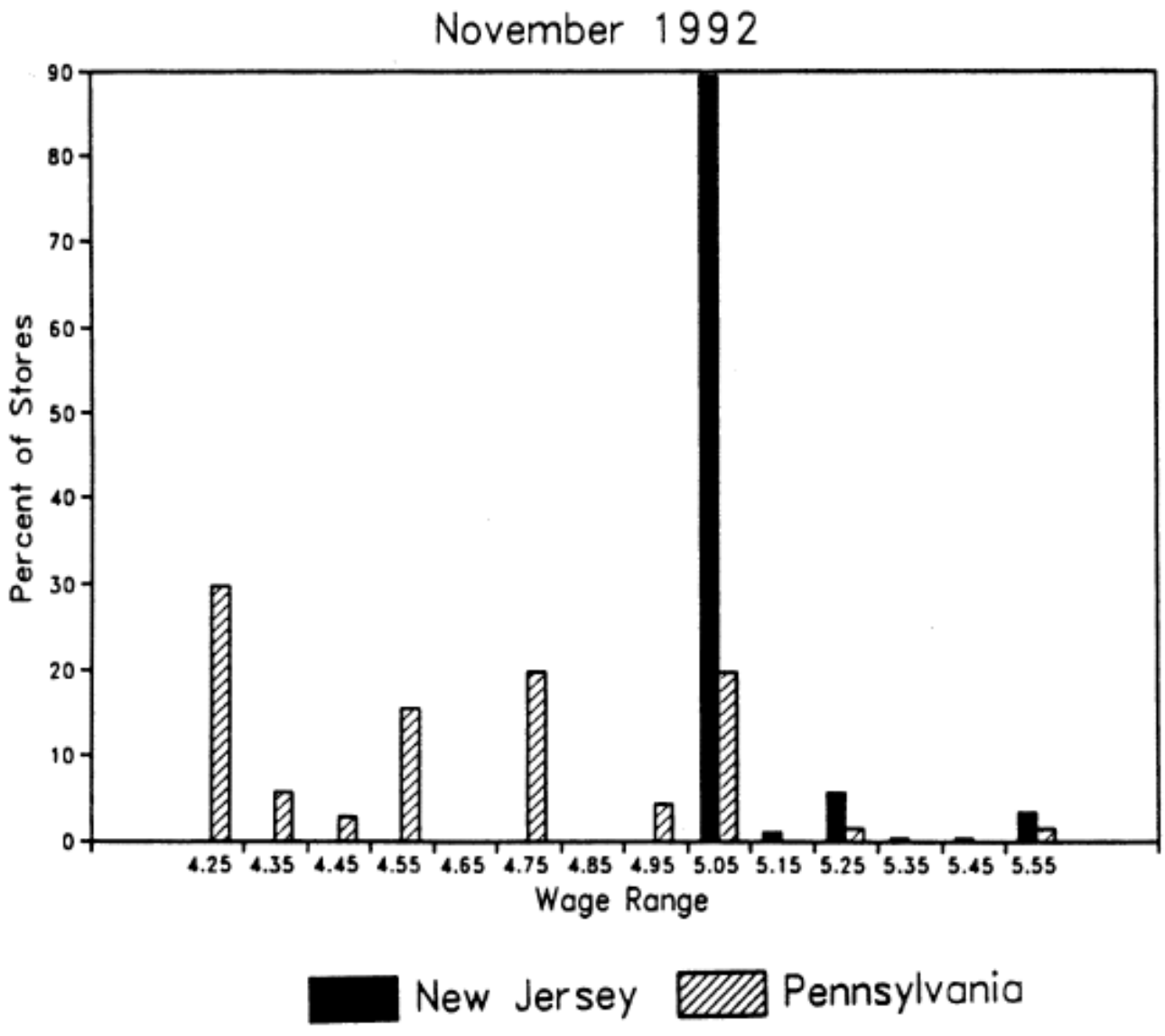
Parallel trends assumption

- Key identifying assumption is the "parallel trends" assumption

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

- 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
Percent of Stores
0
5
10
15
20
25
30
35
4.25
4.35
4.45
4.55
4.65
4.75
4.85
4.95
5.05
5.15
5.25
5.35
5.45
5.55
Wage Range


November 1992
Percent of Stores
90
80
70
60
50
40
30
20
10
0
4.25
4.35
4.45
4.55
4.65
4.75
4.85
4.95
5.05
5.15
5.25
5.35
5.45
5.55
Wage Range
New Jersey
Pennsylvania

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

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

Regression DD

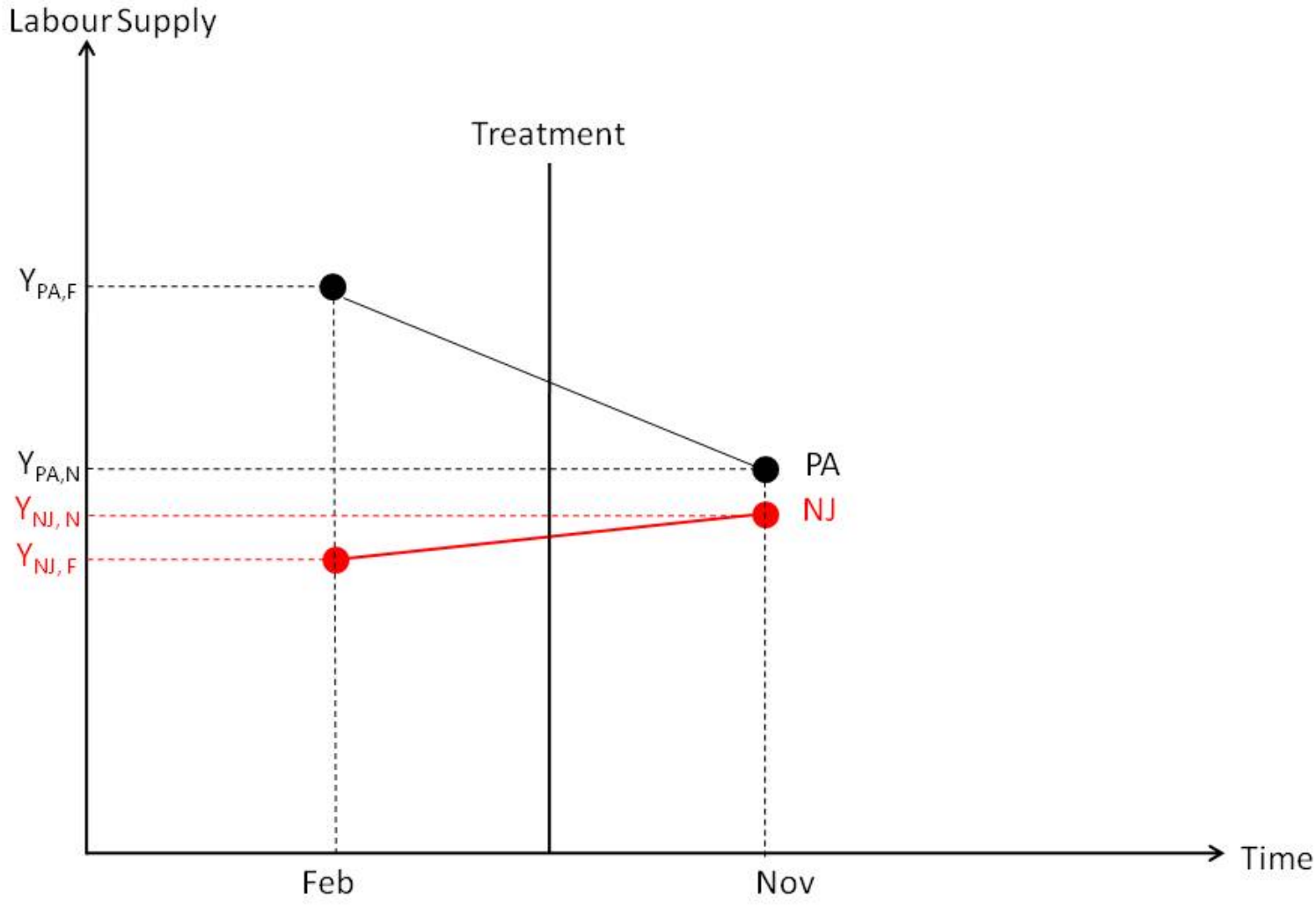
- There are several good reasons to use TWFE
  - → It estimates the ATT under parallel trends
  - → It's easy to calculate the standard errors
  - → It's easy to include multiple periods
  - → We can study treatments with different treatment intensity. (e.g., varying increases in the minimum wage for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

Regression DD - Card and Krueger

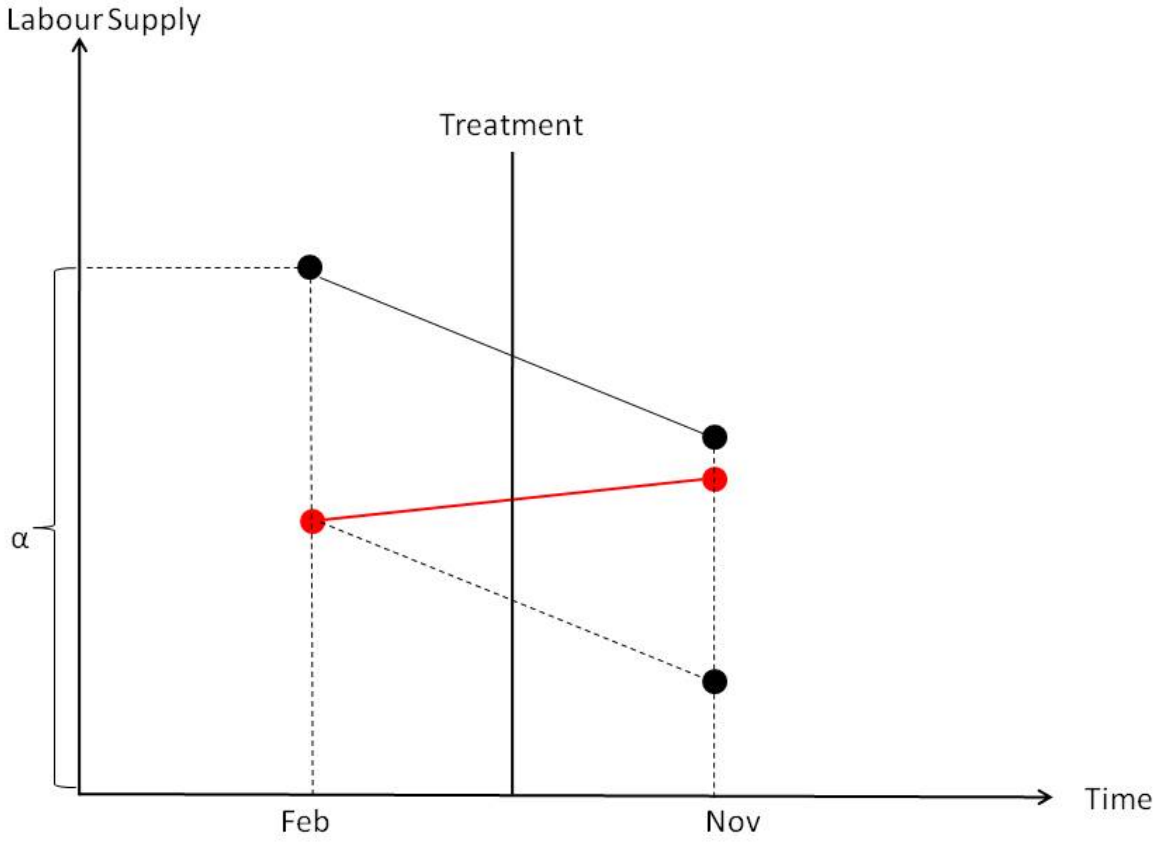
- In the Card and Krueger case, the equivalent regression would be:

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

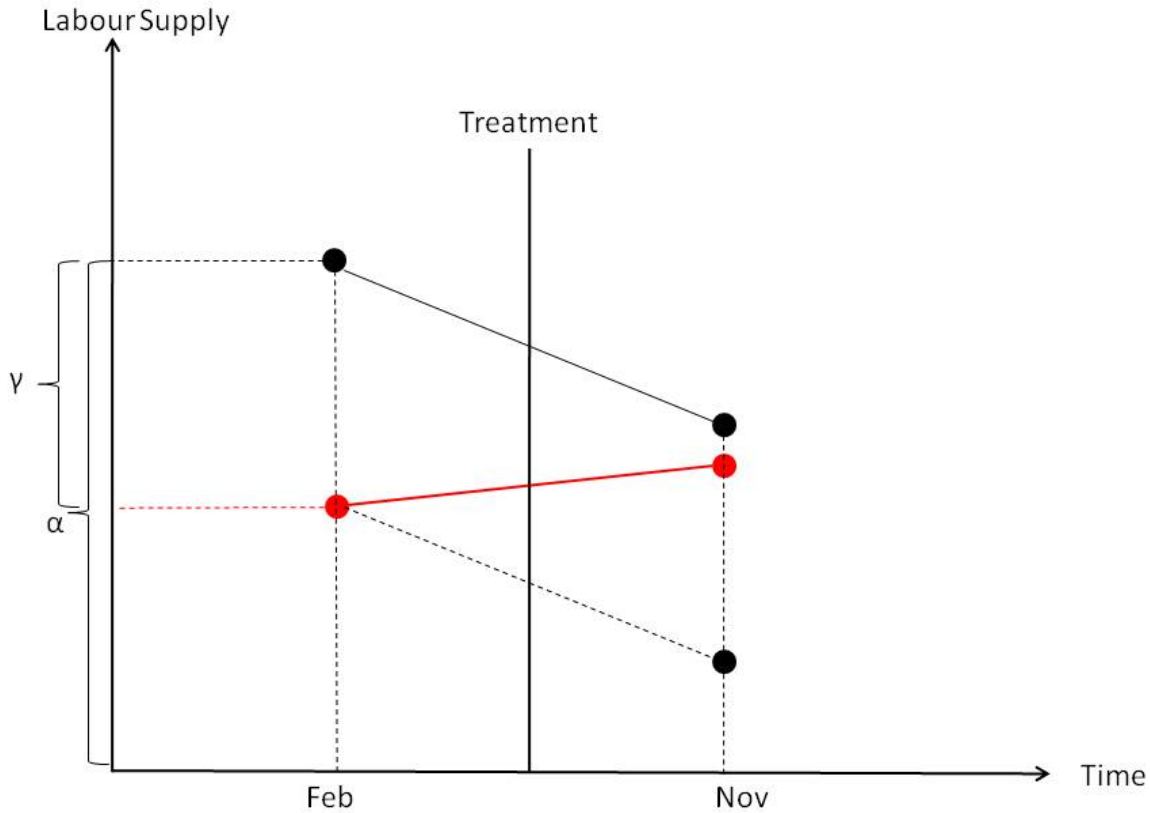
  - → NJ is a dummy equal to 1 if the observation is from NJ
  - → d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
  - → PA Pre: α
  - → PA Post: $\alpha + \lambda$
  - → NJ Pre: $\alpha + \gamma$
  - → NJ Post: $\alpha + \gamma +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


Labour Supply
Treatment
$Y_{PA,F}$
$Y_{PA,N}$
$Y_{NJ,N}$
$Y_{NJ,F}$
PA
NJ
Feb
Nov
Time

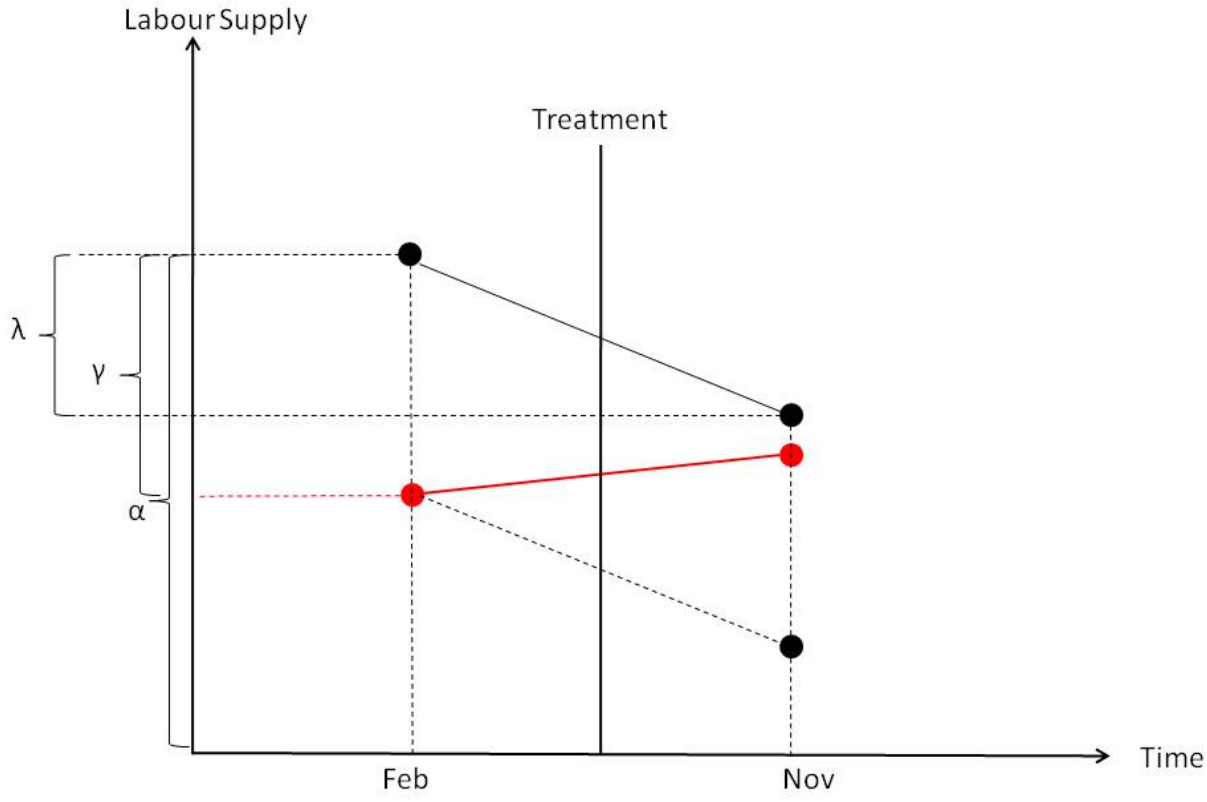
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



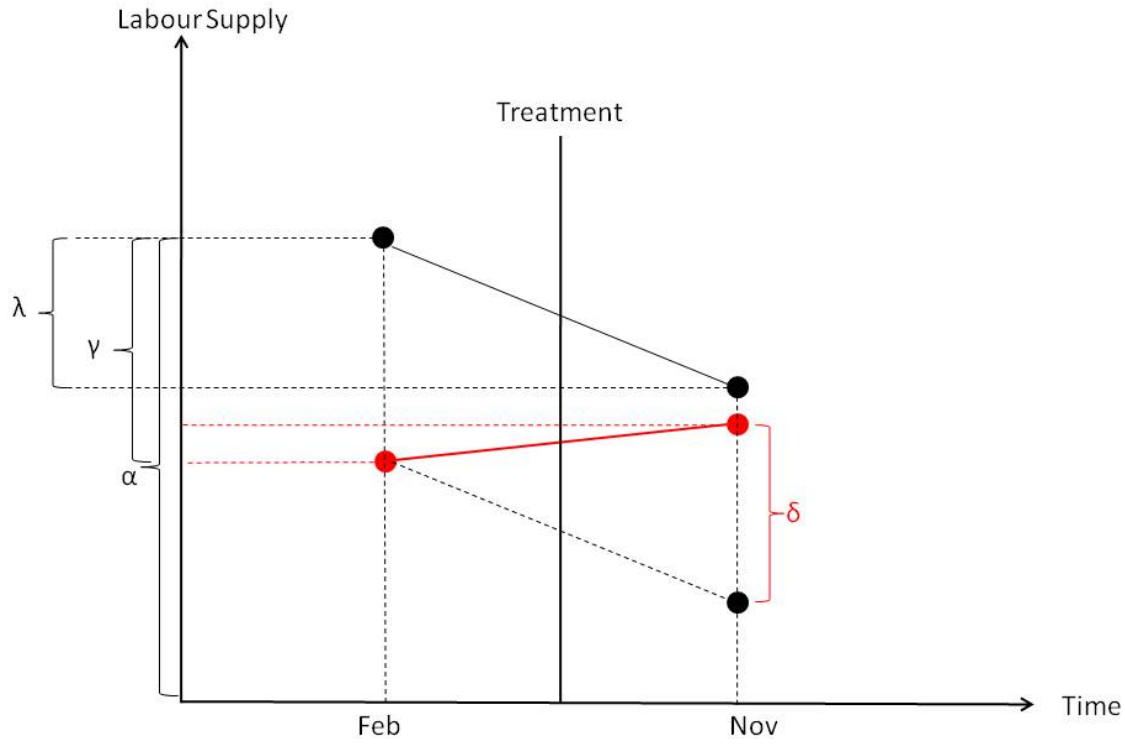
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



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



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



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 trend for the treatment group

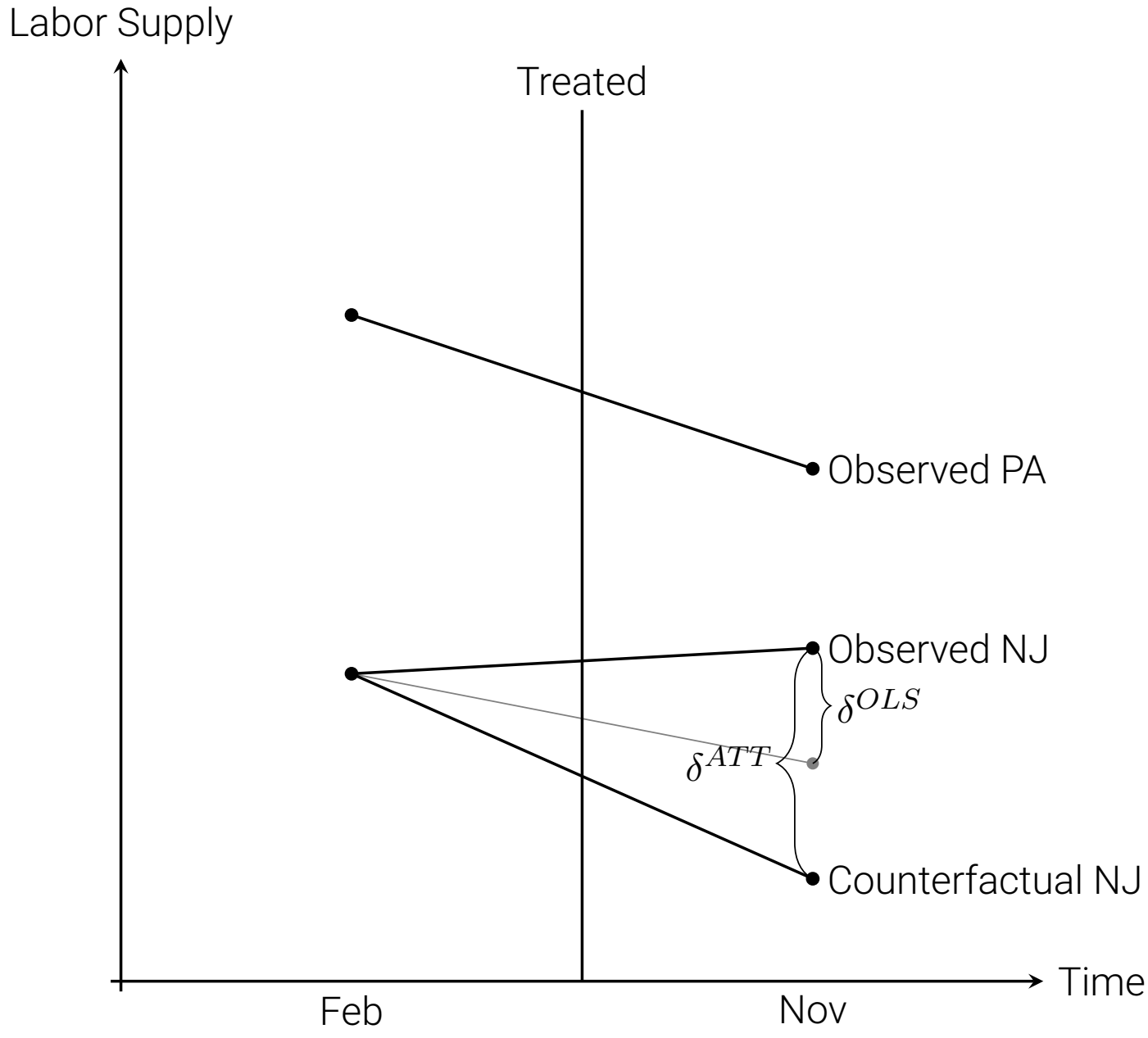


Figure: DD regression diagram without parallel trends

Compositional differences violate parallel trends

- One of the risks of a repeated cross-section is that the composition of the sample may have changed between the pre and post period
- Hong (2011) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napter (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

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

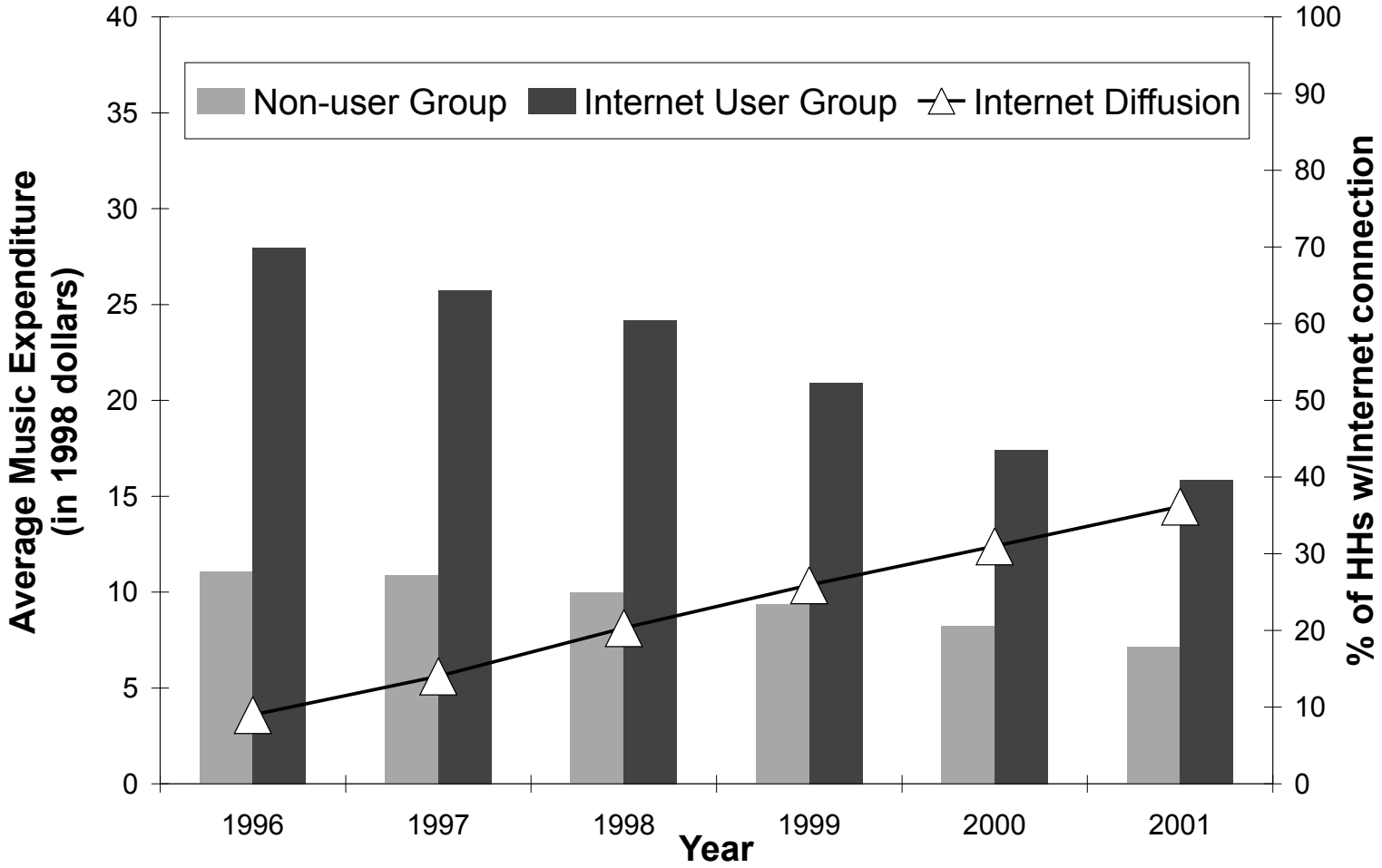


Table 1: Descriptive Statistics for Internet User and Non-user Groups[a]

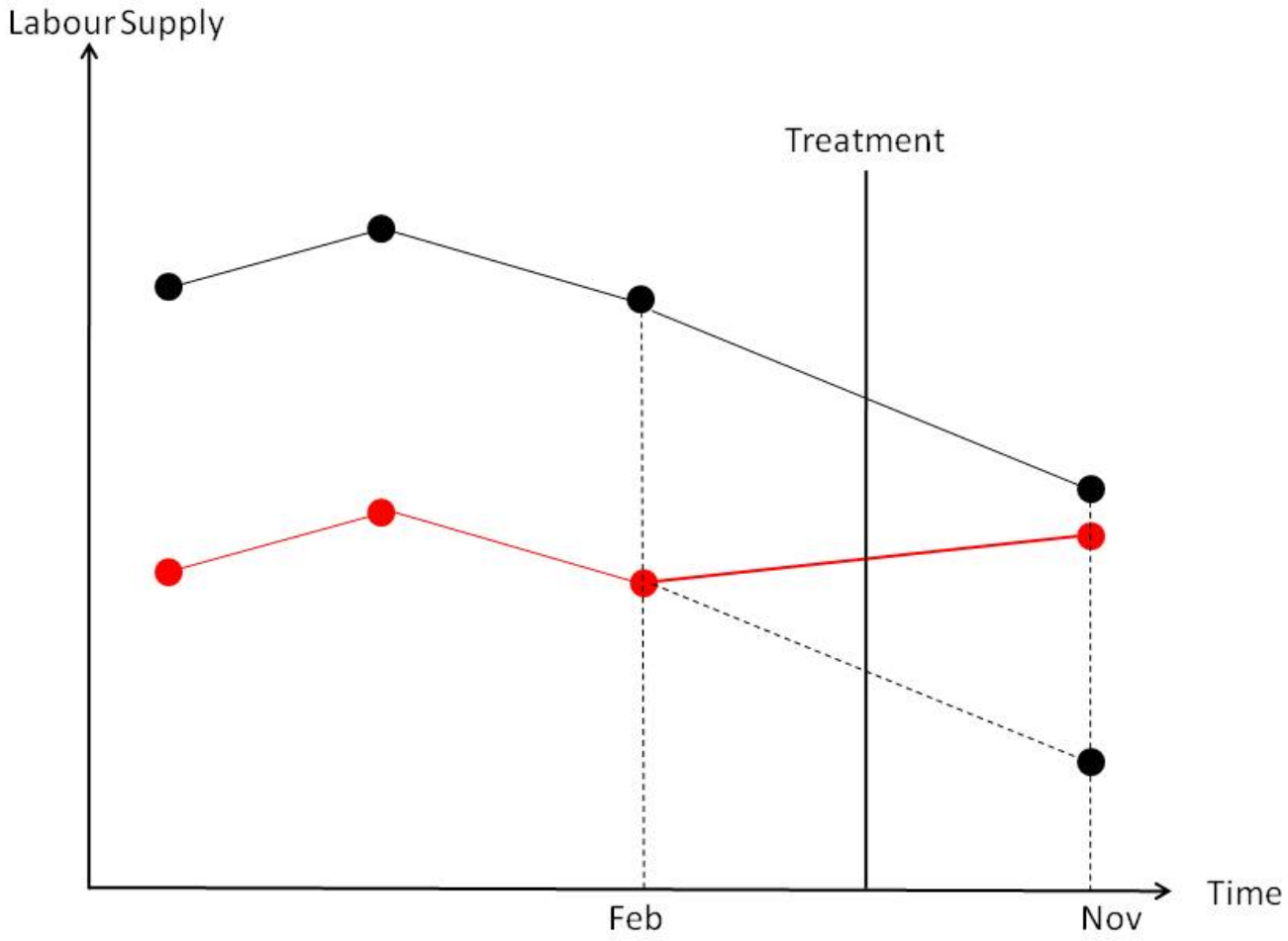
Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	$25.73	$10.90	$24.18	$9.97	$20.92	$9.37
Entertainment	$195.03	$96.71	$193.38	$84.92	$182.42	$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	$52,887	$30,459	$51,995	$28,169	$49,970	$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

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

Evidence for parallel trends: pre-trends

- The identifying assumption for all DD designs is parallel trends
- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check a hunch for parallel trends using pre-trends
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

Plot the raw data when there's only two groups



Event study regression

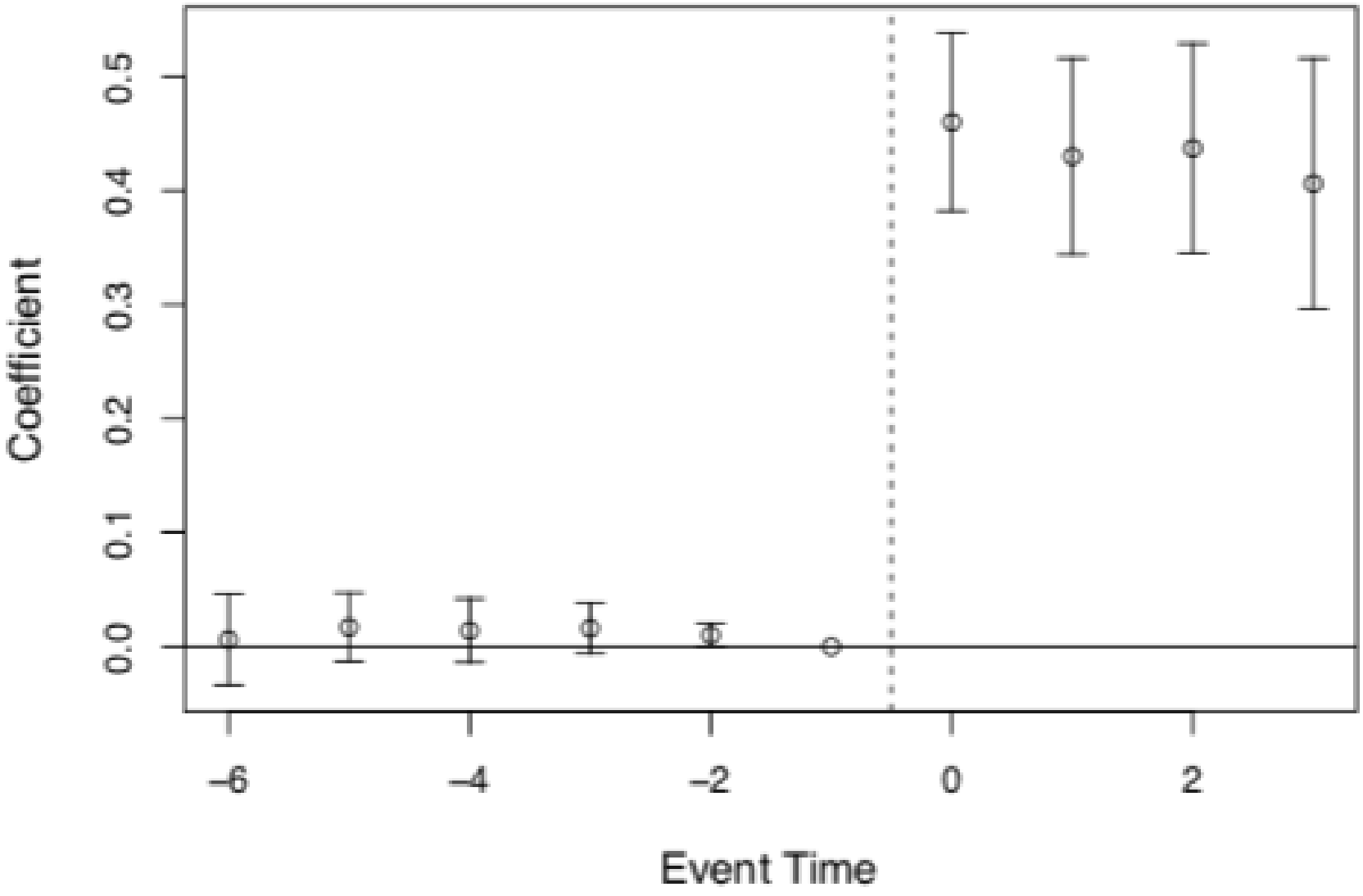
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

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

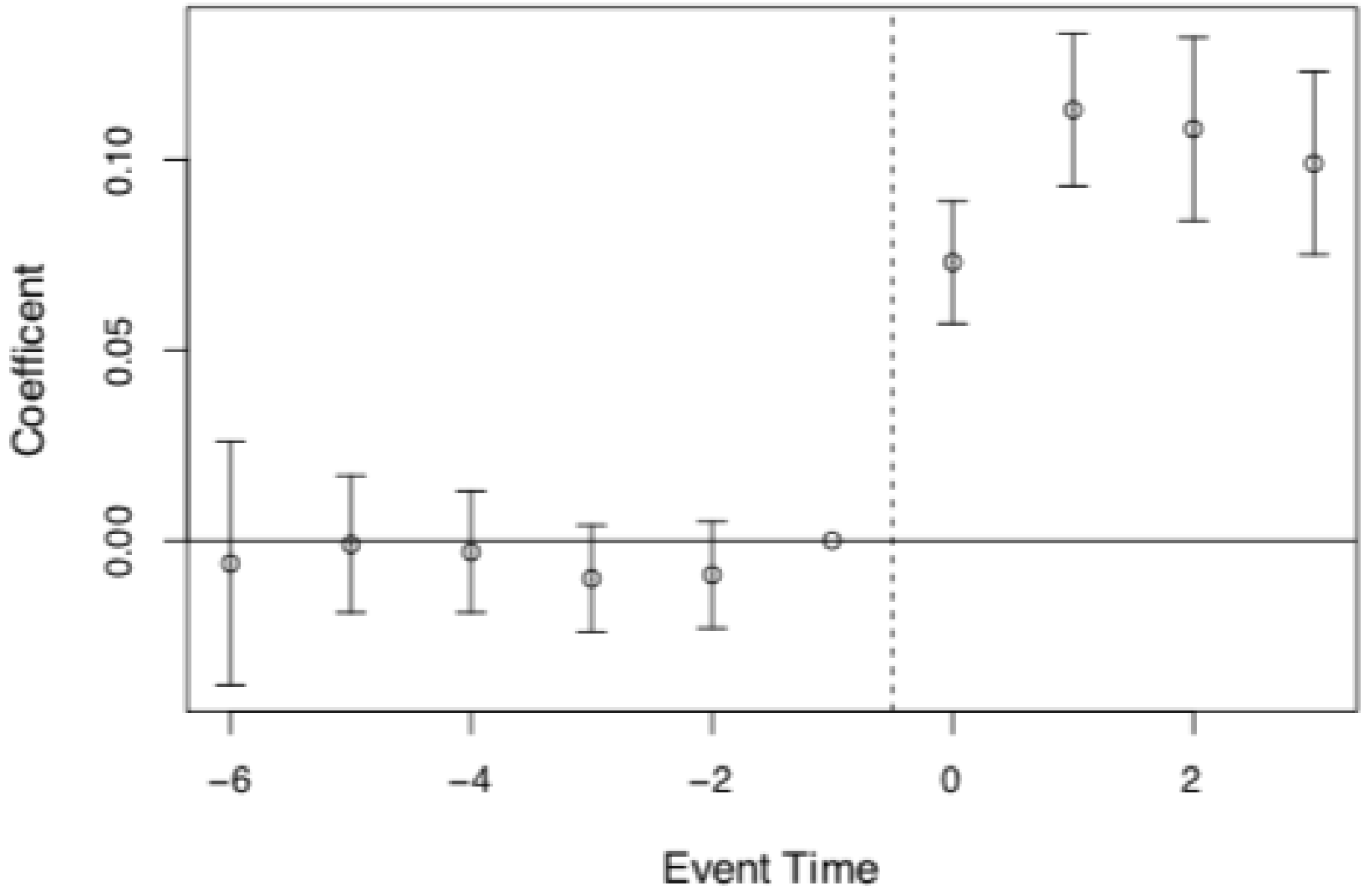
→ Treatment occurs in year 0
→ Includes q leads or anticipatory effects
→ Includes m leads or post treatment effects

Medicaid and Affordable Care Act example

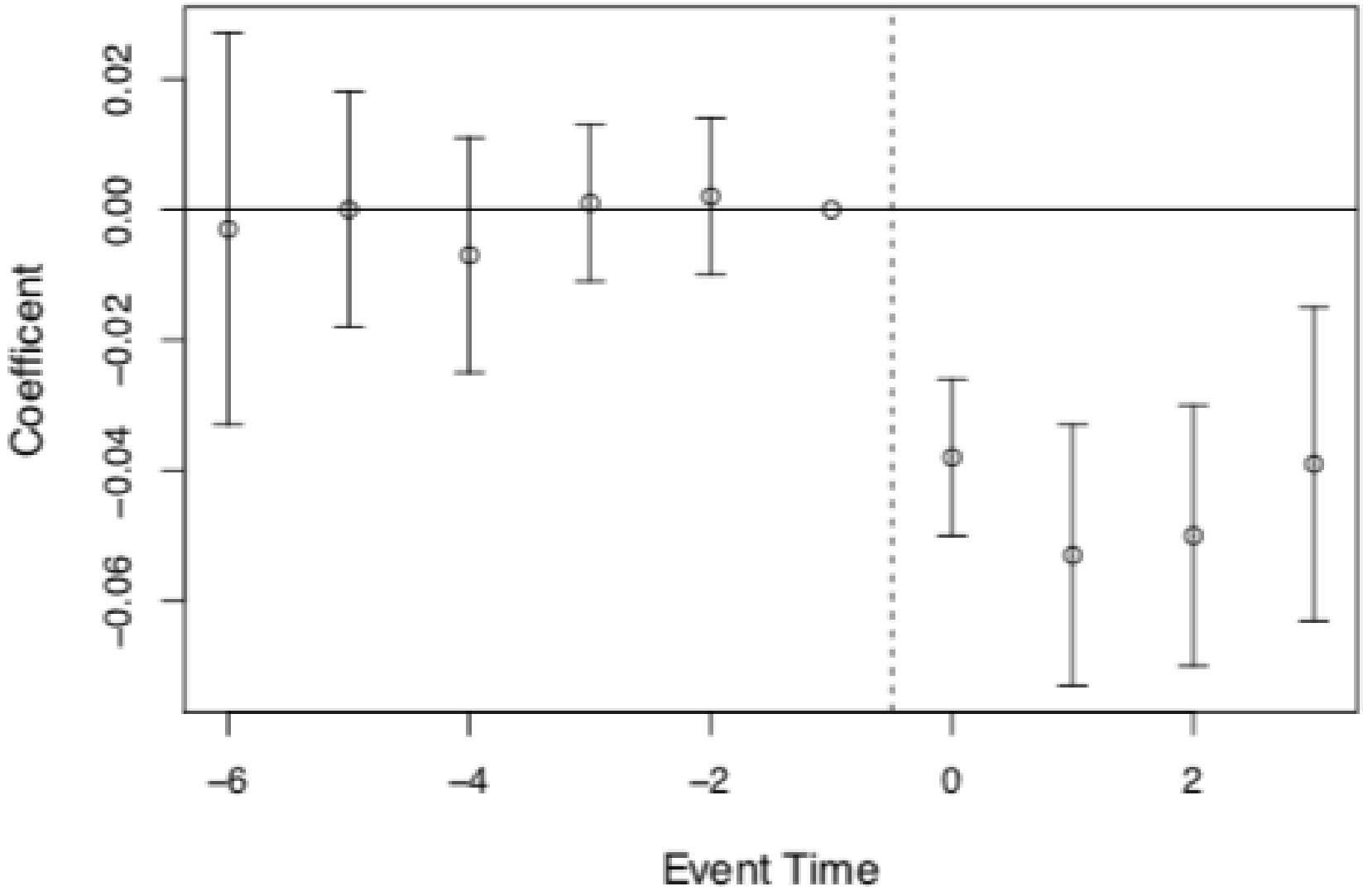
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

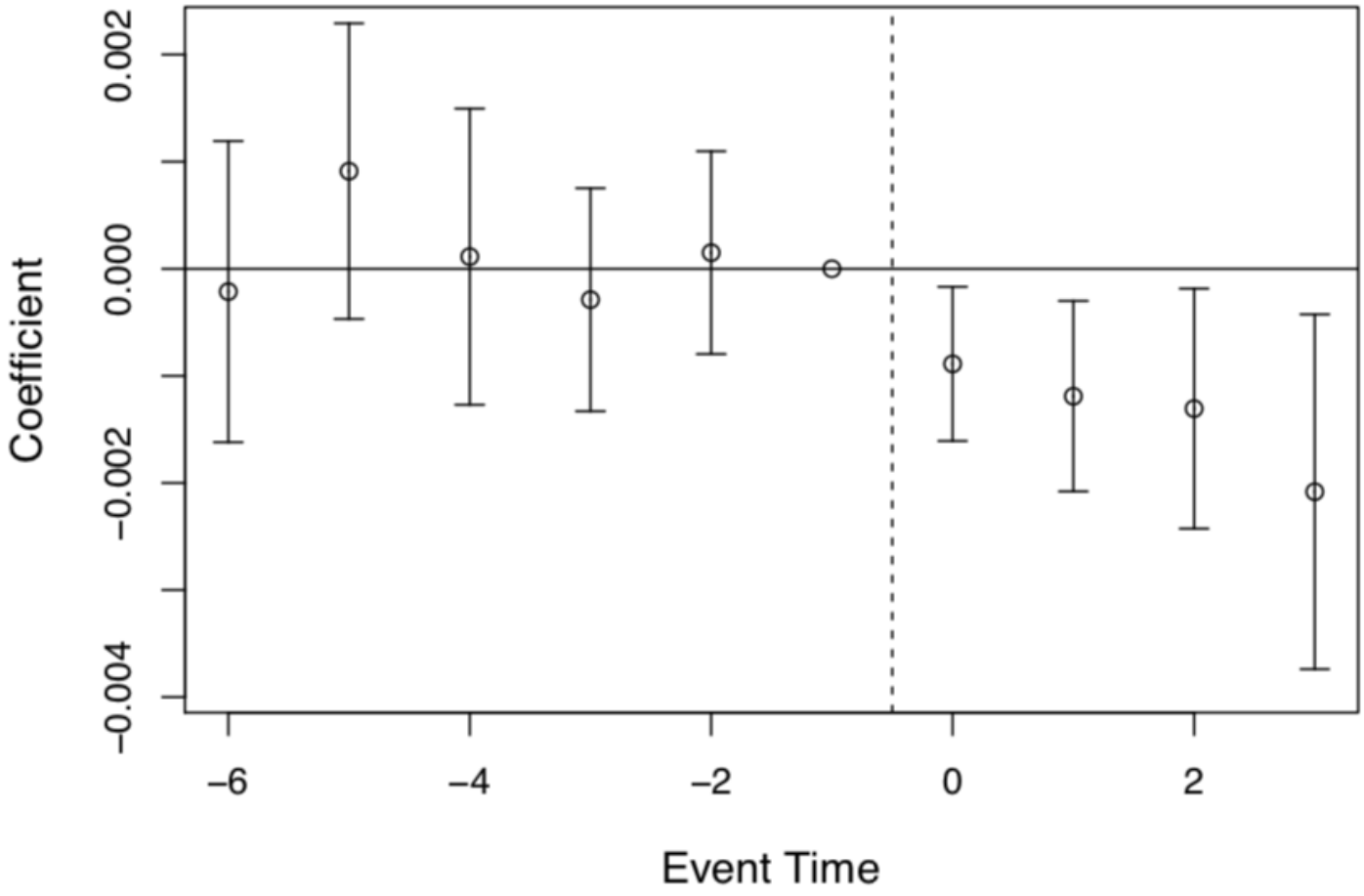


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

Standard errors in DD strategies

- Many paper 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 understate the standard deviation of the estimators – standard errors are biased downward (i.e., too small, over reject)

Standard errors in DD – practical solutions

- Bertrand, Duflo and Mullainathan (2004) propose the following solutions:
  1. Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
  2. Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)
  3. Aggregating the data into one pre and one post period. Literally works if there is only one treatment data. With staggered treatment dates one should adopt the following procedure:
     - Regress Y_{st} onto state FE, year FE and relevant covariates
     - Obtain residuals from the treatment states only and divide them into 2 groups: pre and post treatment
     - Then regress the two groups of residuals onto a post dummy

Additional identification things

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

Table: Difference-in-Difference-in-differences

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

What is our identifying assumption?

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

DDD Example by Gruber

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

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

DDD in Regression

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

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

Falsification test with alternative outcome

- The within-group control group (DDD) is a form of placebo analysis using the same *outcome*
- But there are also placebos using a *different* outcome – but you need a hypothesis of mechanisms to figure out what is in fact a *different outcome*
- Figure out what those are, and test them – finding no effect raises the epistemological credibility of the first result, interestingly
- Cheng and Hoekstra (2013) examine the effect of castle doctrine gun laws on non-gun related offenses like grand theft auto and find no evidence of an effect

Rational addiction as a placebo critique

Sometimes, an empirical literature may be criticized using nothing more than placebo analysis

> *"A majority of [our] respondents believe the literature is a success story that demonstrates the power of economic reasoning. At the same time, they also believe the empirical evidence is weak, and they disagree both on the type of evidence that would validate the theory and the policy implications. Taken together, this points to an interesting gap. On the one hand, most of the respondents claim that the theory has valuable real world implications. On the other hand, they do not believe the theory has received empirical support."*

Placebo as critique of empirical rational addiction

- Auld and Grootendorst (2004) estimated standard “rational addiction” models (Becker and Murphy 1988) on data with milk, eggs, oranges and apples.
- They find these plausibly non-addictive goods are addictive, which casts doubt on the empirical rational addiction models.

Placebo as critique of peer effects

- Several studies found evidence for “peer effects” involving inter-peer transmission of smoking, alcohol use and happiness tendencies
- Christakis and Fowler (2007) found significant network effects on outcomes like obesity
- Cohen-Cole and Fletcher (2008) use similar models and data and find similar network “effects” for things that *aren’t* contagious like acne, height and headaches
- Ockham’s razor - given social interaction endogeneity (Manski 1993), homophily more likely explanation

Roadmap

- Difference-in-differences
  - Two group case
  - Verifying assumptions
- Differential timing
  - Twoway Fixed Effects
  - Strict exogeneity
  - Simulation
- Robust DiD with differential timing
  - Implicit imputation
  - Event studies
  - Explicit imputation
- Concluding Remarks

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

Edit profile

Estimation

You can use the simple DiD equation or you can use the following OLS specification: you get the same answer.

$$Y_{it} = \alpha + \gamma D_k + \lambda Post_t + \delta(D_k \times Post_t) + \varepsilon_{it}$$

If parallel trends holds, then $\widehat{\delta}_{OLS} = \delta$, which is the ATT. (See Heckman, et al. 1997; Abadie 2005; Sant'Anna and Zhao 2020 for including covariates)

But when units select into treatment at different points in time, the above specification won't work

Differential timing

- When treatments are adopted by different units at different points in time, canonical OLS specifications included panel unit i and time t fixed effects hence “twoway fixed effects”
- Standard assumption needed to identify coefficient on treatment status in panel literature was the strict exogeneity assumption (errors in all periods independence of treatment status in all periods)
- Turned out to be more complicated than just parallel trends

TWFE Specification

Let's define an aggregate treatment effect, δ, with heterogeneity across panel units and time (Gardner 2021)

$$E\left[Y_{gpit}|g,p,D_{gp}\right] = \underbrace{\lambda_g + \gamma_p}_{\text{Parallel trends}} + \delta_{gp}D_{gp} \quad (1)$$

$$E\left[\delta_{gp}|D_{gp}=1\right] = \underbrace{E\left[Y^1_{gpit} - Y^0_{gpit}|D_{gp}=1\right]}_{\text{ATT}} \quad (2)$$

i: panel units; t: Calendar time

$g \in \{0,1,\ldots,G\}$ – groups of panel units with same treatment dates

$p \in \{0,1,\ldots,P\}$ – relative time or "periods"

Strict exogeneity violation

Substitute (2) into (1):

$$E\left[Y_{gpit}|g,p,D_{gp}\right] = \lambda_g + \gamma_p + E\left[\delta_{gp}|D_{gp}=1\right]D_{gp} + \left[\delta_{gp} - E(\delta_{gp}|D_{gp}=1)\right]D_{gp} \quad (3)$$

- New composite error term may not be mean-zero conditional on group, period and treatment status (a differential timing problem)
- Strict exogeneity was violated with differential timing and heterogenous treatment effects (I was never taught this)
- Therefore TWFE is biased with differential timing and certain forms of heterogeneity

Twoway fixed effects

$$Y_{st} = \alpha + \delta D_{st} + \sigma_s + \tau_t + \varepsilon_{st}$$

Estimated with OLS, we call this the model the twoway fixed effects (TWFE) model where σ_s is one fixed effect (state) and τ_t is the second fixed effect (time).

Historically, we ran this model and *hoped* that $\widehat{\delta}$ was some reasonably weighted average of treatment effects, but now we know better because heterogeneity creates potential biases under differential timing

TWFE decomposition

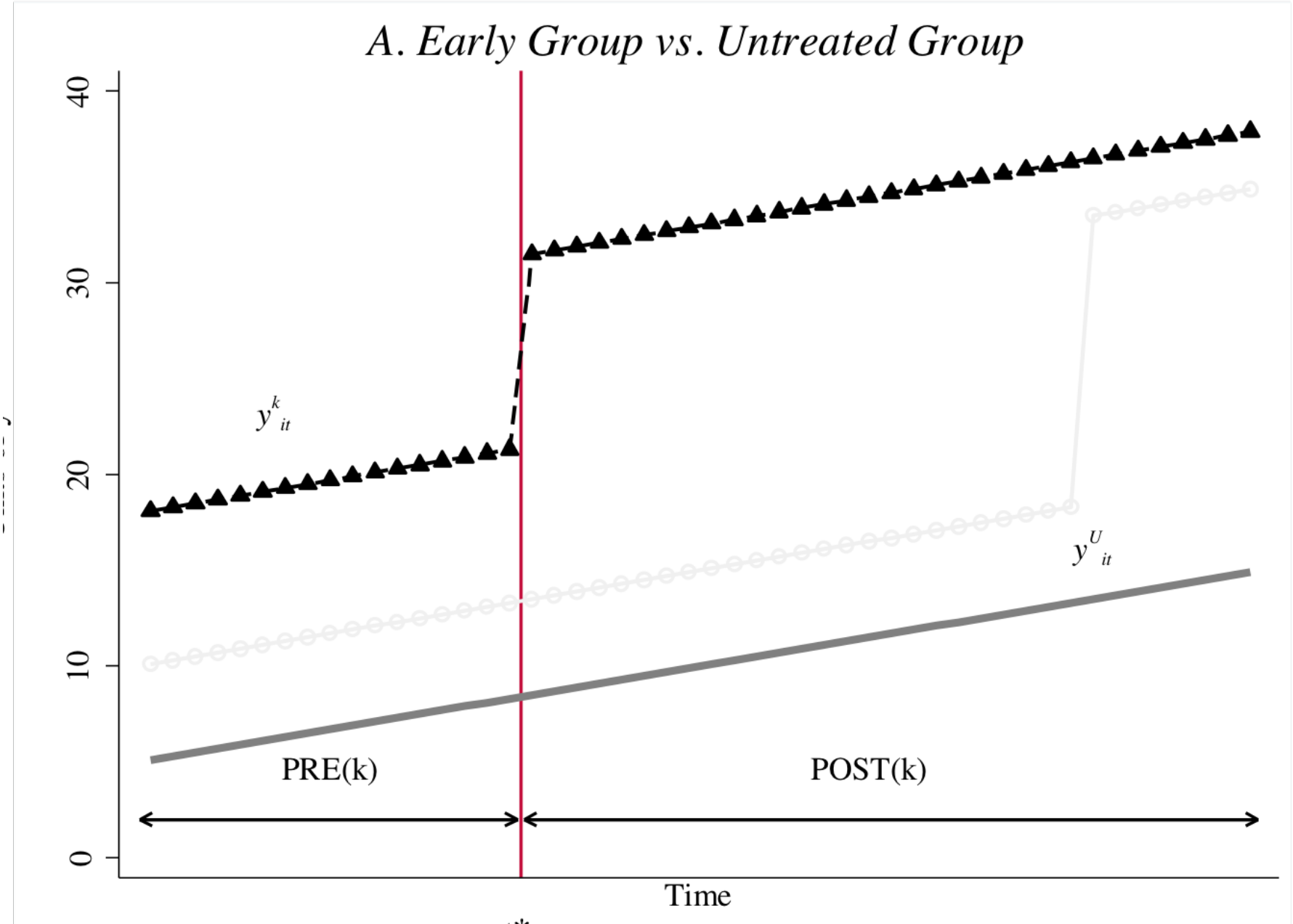
Theoretical decompositions reveal possibly negative weights on treatment effects, but numerical decompositions show positive weights on data itself

1. **Numerical decomposition of TWFE coefficient**: what numbers add up to equal the TWFE coefficient? Goodman-Bacon 2021
2. **Theoretical decomposition**: what theoretically does TWFE coefficient "mean" in terms of causality and bias? Gardner 2021, de Chaisemartin and d'Haultfoeille 2020, Borusyak and Jaravel 2016

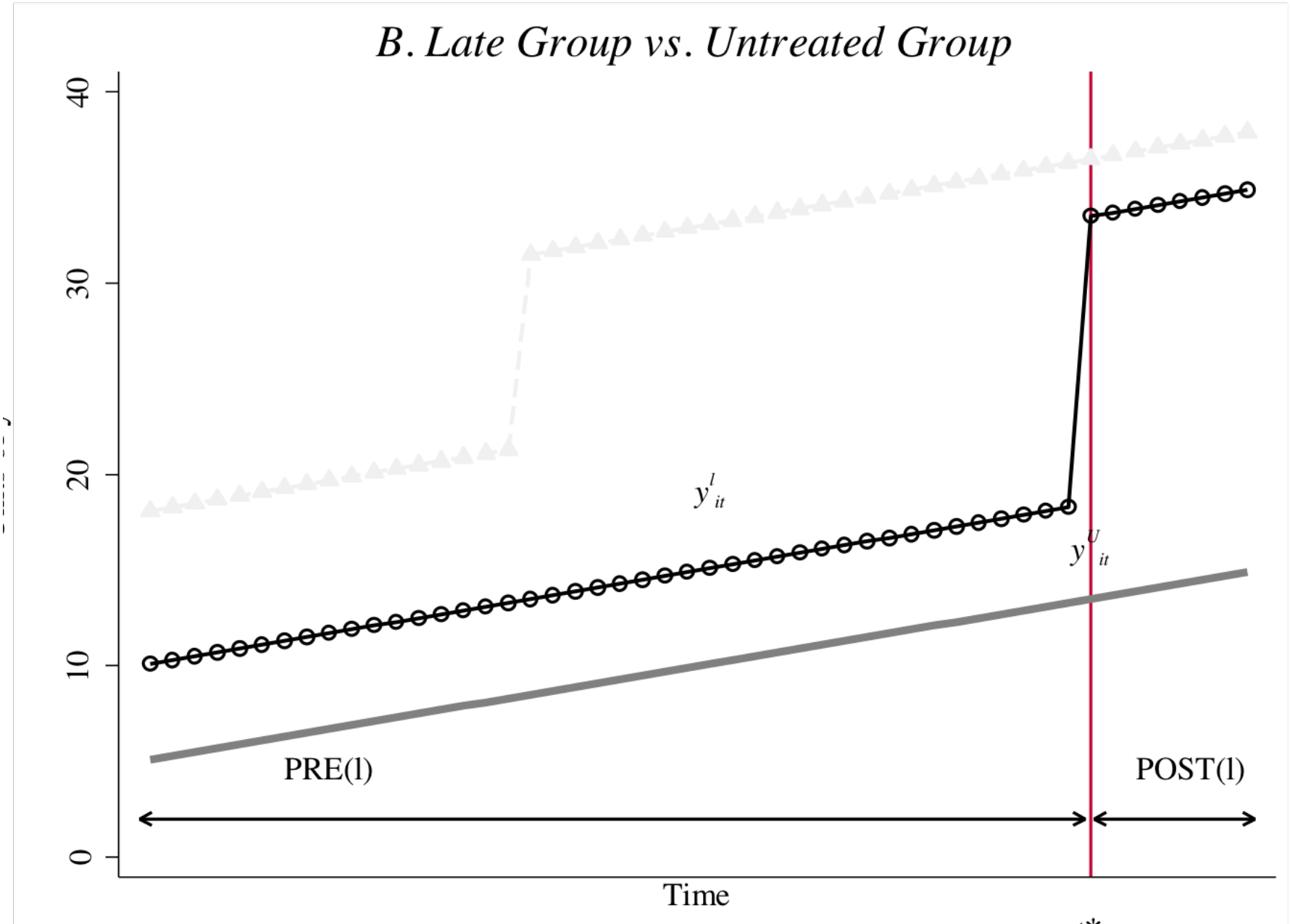
Terms and notation

- Now there will be two treatment groups (k,l) and one untreated group (U)
- k,l are defined by their treatment date with k receiving their treatment earlier than l
- Weights, s_{jb}, are based on variance of treatment and group size
- Denote $\widehat{\delta}_{jb}^{2x2}$ as the canonical 2×2 DD estimator for groups j and b where j is the treatment group and b is the comparison group

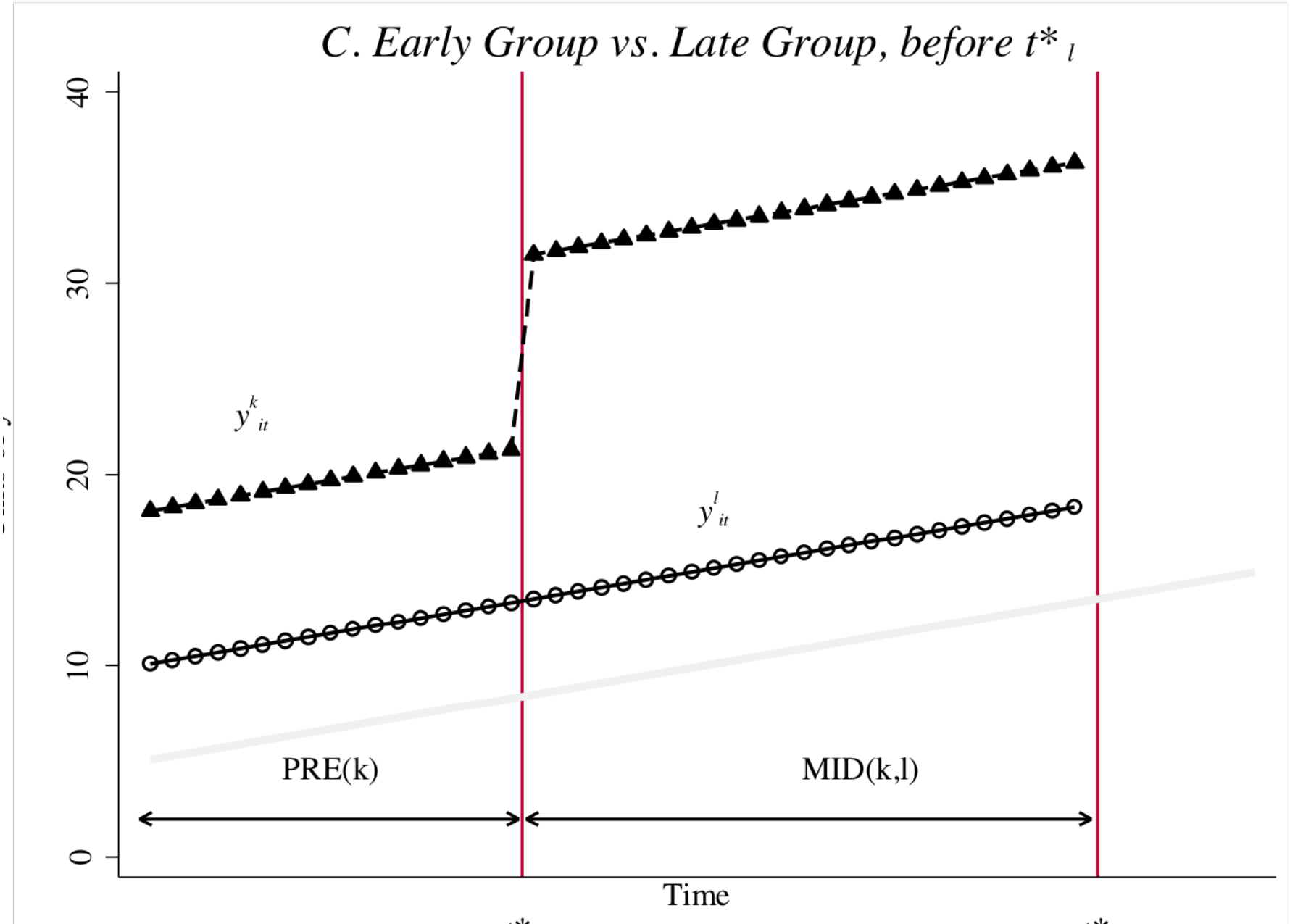
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$



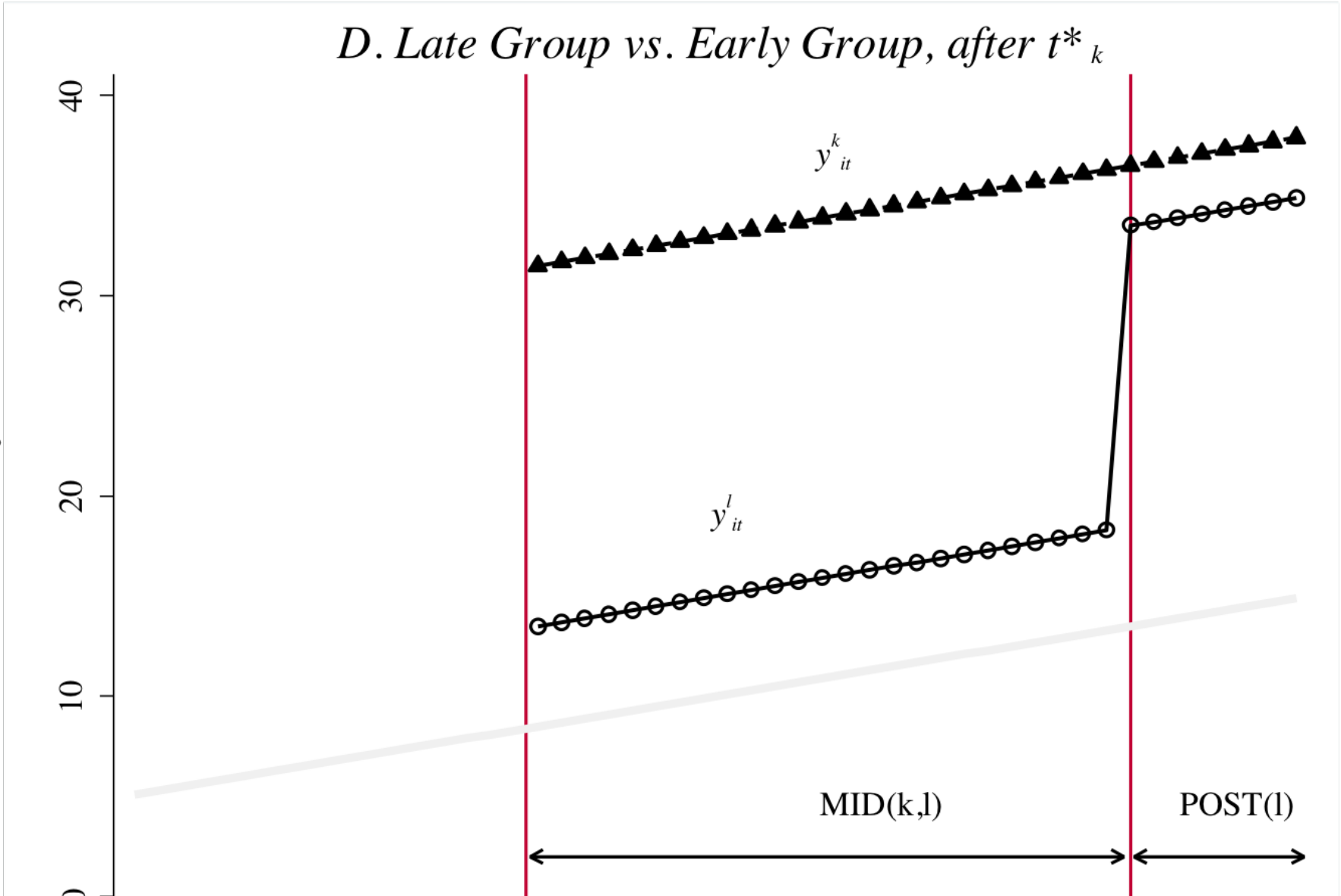
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



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



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



Substitute causality and bias into Bacon decomposition

Bacon decomposition

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

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

Three types of 2x2s (the last is "forbidden") :

$$\begin{aligned}
\widehat{\delta}^{2x2}_{kU} &= ATT_k(Post) + \Delta Y^0_l(Post, Pre) - \Delta Y^0_U(Post, Pre) \\
\widehat{\delta}^{2x2,k}_{kl} &= ATT_k(Mid) + \Delta Y^0_l(Mid, Pre) - \Delta Y^0_l(Mid, Pre) \\
\widehat{\delta}^{2x2,l}_{lk} &= ATT_l Post(l) + \Delta Y^0_l(Post(l), MID) - \Delta Y^0_k(Post(l), MID) \\
&\quad - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Dynamic treatment effect bias!}}
\end{aligned}$$

Theoretical decomposition

$$p\,lim\,\widehat{\delta}^{TWFE}_{n\to\infty} \quad = \quad VWATT + VWPT - \Delta ATT$$

TWFE does not satisfy a “no sign flip” property because if ΔATT is larger in absolute value than $VWATT$, sign can flip

Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- $E[Y^0]$ satisfies “strong parallel trends” (stronger than necessary)

$$Y_{ist}^0 = \alpha_i + \gamma_t + \varepsilon_{ist}$$

- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

- Heterogenous treatment effects across time and across groups
- Cells are called “group-time ATT” (Callaway and Sant’anna 2020) or “cohort ATT” (Sun and Abraham 2020)
- ATT is weighted average of all cells and $+82$ with uniform weights $1/60$

Estimation

Estimate the following equation using OLS:

$$Y_{ist} = \alpha_i + \gamma_t + \delta D_{it} + \varepsilon_{ist}$$

Table: Estimating ATT with different models

	Truth	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{ATT}$	82	-6.69***			

The sign flipped. Why? Because of *extreme* dynamics (i.e., $-\Delta ATT$)

Bacon decomposition

Table: Bacon Decomposition (TWFE $= -6.69$)

DD Comparison	**Weight**	**Avg DD Est**
Earlier T vs. Later C	0.500	51.800
Later T vs. Earlier C	0.500	-65.180

T $=$ Treatment; C$=$ Comparison

$(0.5 * 51.8) + (0.5 * -65.180) = -6.69$

While large weight on the “late to early 2x2” is *suggestive* of an issue, these would appear even if we had constant treatment effects

Roadmap

- Difference-in-differences
  - Two group case
  - Verifying assumptions
- Differential timing
  - Twoway Fixed Effects
  - Strict exogeneity
  - Simulation
- Robust DiD with differential timing
  - Implicit imputation
  - Event studies
  - Explicit imputation
- Concluding Remarks

Imputation

- “At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others.” – Imbens and Rubin (2015)
- Causal inference *always* imputes missing counterfactuals, just not always easiest to see it (e.g., RCTs)
- New solutions can be thought of as either “implicit” vs “explicit” imputation methods

Assumptions

1. Panel or repeated cross section data (modularity)
2. Conditional parallel trends (if covariates)
3. Common support (if covariates)
4. No anticipation (zero treatment effects before treatment)
5. Irreversible treatment (treatment cannot switch back and forth)

Callaway and Sant'anna 2020 with IPW

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

CS avoids mortal sin of late-to-early comparison by only using the never or not-yet treated as controls C through subsetting dataset

Group-time ATT

Truth				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			
Feasible ATT	68.3333333			

CS estimates				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1981	-0.0548	0.0191	0.0578	0
1986	10.0258	-0.0128	-0.0382	0
1987	20.0439	0.0349	-0.0105	0
1988	30.0028	-0.0516	-0.0055	0
1989	40.0201	0.0257	0.0313	0
1990	50.0249	0.0285	-0.0284	0
1991	60.0172	-0.0395	0.0335	0
1992	69.9961	8.013	0	0
1993	80.0155	16.0117	0.0105	0
1994	89.9912	24.0149	0.0185	0
1995	99.9757	32.0219	-0.0505	0
1996	110.0465	40.0186	0.0344	0
1997	120.0222	48.0338	-0.0101	0
1998	129.9164	56.0051	6.027	0
1999	139.9235	63.9884	11.969	0
2000	150.0087	71.9924	18.0152	0
2001	159.9702	80.0152	23.9656	0
2002	169.9857	88.0745	29.9757	0
2003	179.981	96.0161	36.013	0
2004				
2005				
2006				
2007				
2008				
2009				
Total ATT	n/a			
Feasible ATT	68.33718056			

Question: Why didn't CS estimate all ATT(g,t)? What is "feasible ATT"?

Reporting results

Table: Estimating ATT using only pre-2004 data

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
$\widehat{Feasible\ ATT}$	68.33	26.81 ***	68.34***		

TWFE is no longer negative, interestingly, once we eliminate the last group (giving us a never-treated group), but is still suffering from attenuation bias.

Event studies

- Remember Don Rubin: randomization gives us confidence because "we know how the science works" but **there is no science of parallel trends**, so we must support it somehow with indirect evidence
- Because the identifying assumption is parallel trends, the falsifications that ultimately were chosen were tests on estimated lead coefficients in event study design
- Event study estimation with misspecified TWFE models under differential timing were also problematic (Borusyak and Jaravel 2016; Sun and Abraham 2020; Borusyak, Jaravel and Speiss 2021)

Notation and terms

- Units treated at the same time are considered part of the same "group" or "cohort" e (Sun and Abraham 2020 notation)
- If we bin g leads and lags l, then g becomes our focus (i.e., $l \in g$), otherwise l is our focus
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ which was each cell in the simulation data

Notation and terms

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

Assumptions

1. Parallel trends
2. No anticipation
3. Treatment effect homogeneity

TWFE will be unbiased estimate of each population regression coefficient lead and lag

TWFE Regression

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

We estimate this μ_g population regression coefficient leads and lags using TWFE and get $\widehat{\mu_g}$.

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

Weight ($w^g_{e,l}$) summation cheat sheet

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

Intuition for contamination

- Each population regression coefficient is the sum of three things (one good, two bad)
  1. CATT from that period
  2. CATT from the omitted period
  3. CATT from all other relative periods
- When all three assumptions hold, only the lead/lag's CATT remains (all others vanish)
- This vanishing of other period leads and lag CATT happens either bc CATT=0 (no anticipation), or because of the weighting scheme (with homogenous treatment effects)

Toy example

Simple example: A balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$. Estimated coefficient on μ_{-2} is:

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

Robust event study estimation

- All the robust estimators under differential timing have solutions and they all skip over forbidden contrasts.
- Sun and Abraham (2020) propose a 3-step interacted weighted estimator (IW) using last treated group as control group
- Callaway and Sant'anna (2020) estimate group-time ATT which can be a weighted average over relative time periods too but uses "not-yet-treated" as control

Reporting results

Table: Estimating ATT

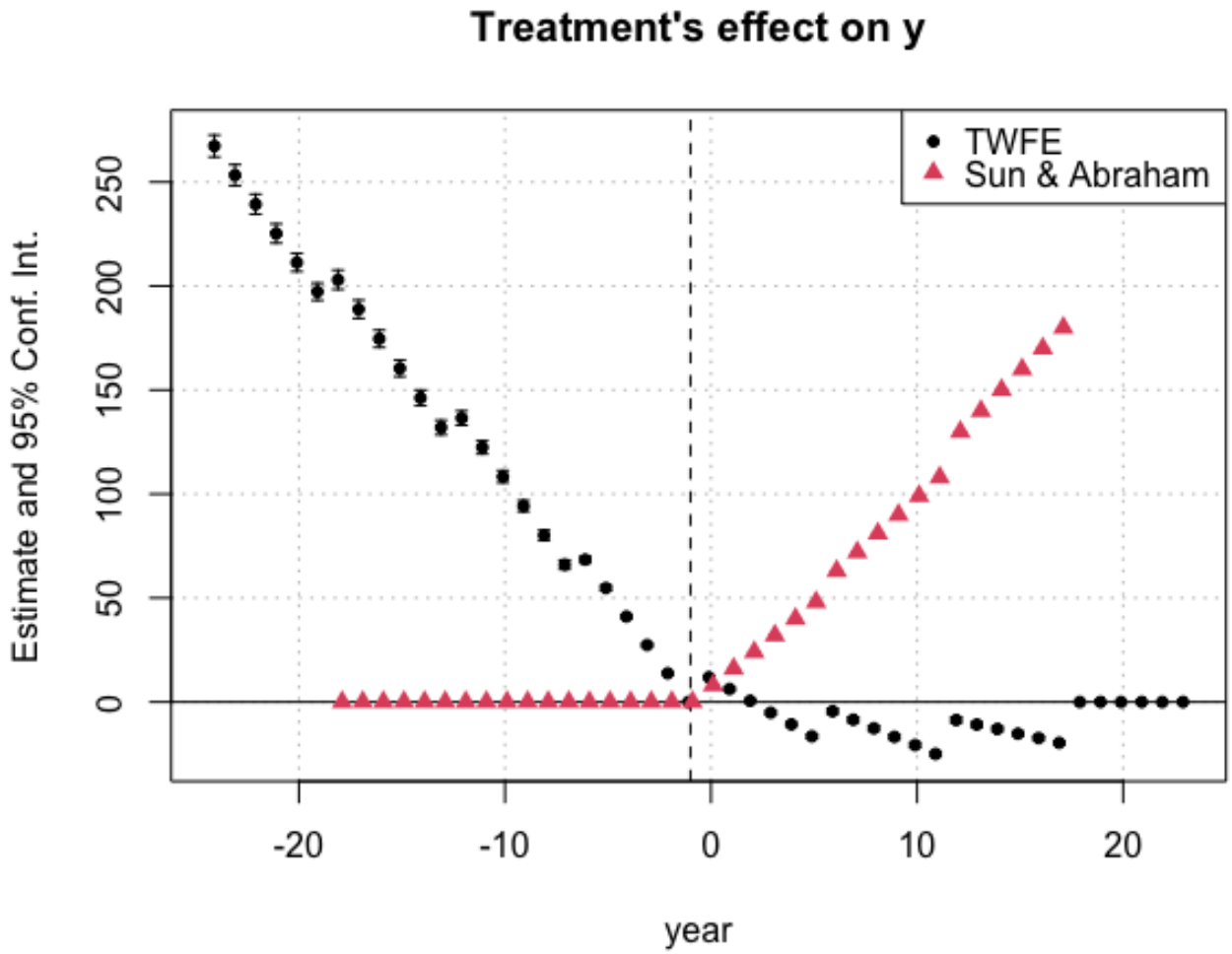
	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	

Computing relative event time leads and lags

Truth						Relative time coefficients		
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)		Leads	Truth	SA
1980	0	0	0	0		t-2	0	0.02
1986	10	0	0	0	(10+8+6)/3 = 8	t	8	8.01
1987	20	0	0	0	(20+16+12)/3 = 16	t+1	16	16.00
1988	30	0	0	0		t+2	24	24.00
1989	40	0	0	0		t+3	32	31.99
1990	50	0	0	0		t+4	40	40.00
1991	60	0	0	0		t+5	48	48.01
1992	70	8	0	0		t+6	63	62.99
1993	80	16	0	0		t+7	72	72.00
1994	90	24	0	0		t+8	81	80.99
1995	100	32	0	0		t+9	90	89.98
1996	110	40	0	0		t+10	99	99.06
1997	120	48	0	0		t+11	108	108.01
1998	130	56	6	0		t+12	130	129.92
1999	140	64	12	0		t+13	140	139.92
2000	150	72	18	0		t+14	150	150.01
2001	160	80	24	0		t+15	160	159.97
2002	170	88	30	0		t+16	170	169.99
2003	180	96	36	0		t+17	180	179.98
2004	190	104	42	4				
2005	200	112	48	8				
2006	210	120	54	12				
2007	220	128	60	16				
2008	230	136	66	20				
2009	240	144	72	24				

Two things to notice: (1) there only 17 lags with robust models but will be 24 with TWFE; (2) changing colors mean what?

Comparing TWFE and SA



Question: why is TWFE *falling* pre-treatment? Why is SA rising, but jagged, post-treatment?

Imputation methods

All recent working papers

1. **2SDiD** (Gardner 2021) – imputes Y^0 using estimated fixed effects from the $D = 0$ units, residualizing into $\widehat{Y}$, regressing new $\widehat{Y}$ using GMM
2. **Robust efficient imputation** (Borusyak, et al. 2021) – very similar to 2SDID in that you impute Y^0 using $D = 0$ sample and estimated fixed effects
3. **Mundlak** (Wooldridge 2022) – TWFE with saturated interactions, is equivalent to the above two

I don't count Athey, et al. (2020) matrix completion with nuclear norm regularization because under my definition since it it does not depend on parallel trends (rather it nests synthetic control), it is not a DiD method

Steps for BJS

Target parameter is individual treatment effect, δ_i

1. Estimate expected potential outcomes using OLS and only the untreated observations (this is similar to Gardner 2021)
2. Then calculate $\widehat{\delta}_{it} = Y_{it}^1 - \widehat{Y}_{it}^0$
3. Then estimate target parameters as weighted sums

$$\widehat{\delta}_W = \sum_{it} w_{it} \widehat{\delta}_{it}$$

Why is this working?

- Because we can obtain consistent estimates of the fixed effects, we can extrapolate to the counterfactual units for all $Y(0)_{it}$ that were treated
- This is the same type of trick we see with Heckman, et al. (1997) as well as Gardner (2021)
- As it is still OLS, it's computationally fast and flexible to unit-trends, triple diff, covariates and so forth (with caveats about time-varying covariates requiring more assumptions)
- Wooldridge shows the Mundak estimator maps onto BJS robust model

Reporting results

Table: Estimating ATT

	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	68.33***

Software

1. Callaway and Sant'anna (2020)
   - → **Stata**: csdid
   - → **R**: did
2. Sun and Abraham (2020)
   - → **Stata**: eventstudyinteract
   - → **R**: fixest with subab() option
3. Borusyak, et al. (2022)
   - → **Stata**: dd_imputation
   - → **R**: didimputation
4. Gardner (2021)
   - → **Stata**: did2s
   - → **R**: did2s

Roadmap

- Difference-in-differences
  - Two group case
  - Verifying assumptions
- Differential timing
  - Twoway Fixed Effects
  - Strict exogeneity
  - Simulation
- Robust DiD with differential timing
  - Implicit imputation
  - Event studies
  - Explicit imputation
- Concluding Remarks

Encouragement

- Models give identical results for “strong parallel trends” types of assumptions (though confidence intervals differ)
- Choose estimator based on which parallel trend assumption you feel most comfortable with is my suggestion
- Encourage you to use authors’ own packages as they have the incentive to fix bugs and maintain it