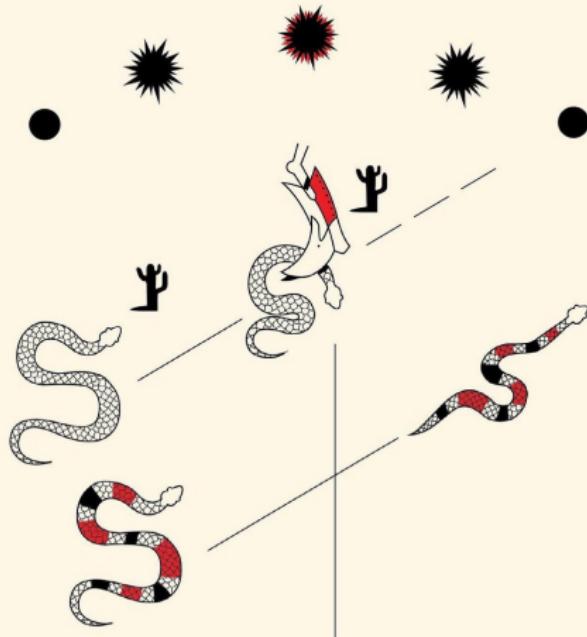


# Difference-in-differences workshop

Presenter: Scott Cunningham

## CODE-CHELLA



## Workshop outline

- Introduction to DiD basics
  - Potential outcomes review
  - DiD formula
  - Covariates
- Differential timing
  - Heterogeneity
  - TWFE bias in estimation of overall and dynamic ATT

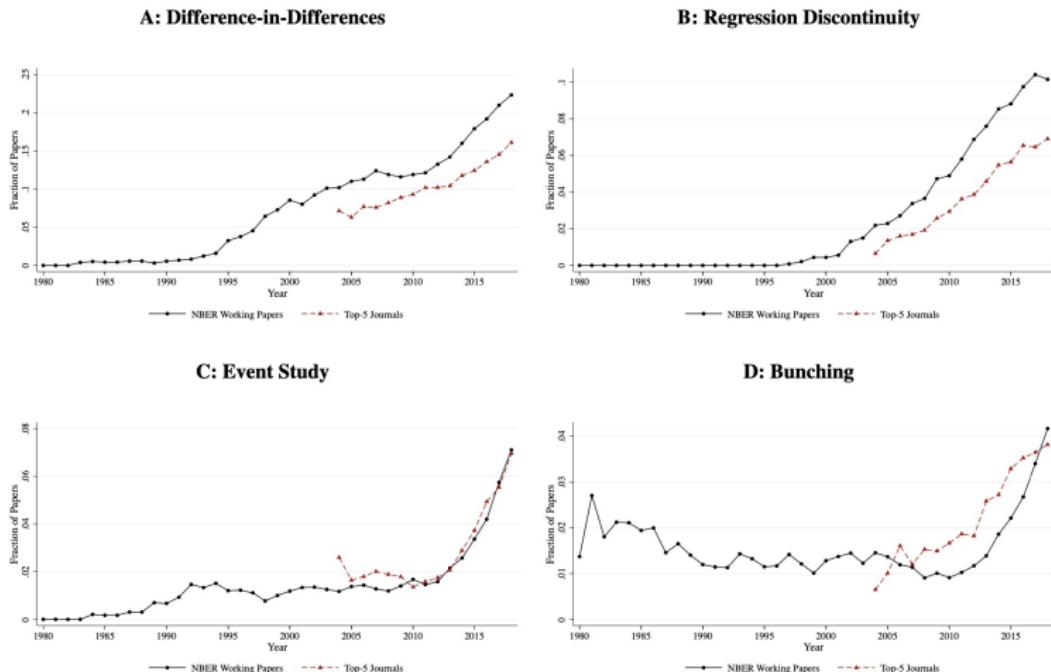
## Workshop outline

- Three types of solutions
  - Aggregated group-time ATT
  - Stacked regression
  - Explicit Imputation
- Continuous treatments
- Fuzzy difference-in-differences

## What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- Early usage in several 19th century health policy debates
- Brought into labor economics with Orley Ashenfelter (1978), LaLonde (1986), Card and Krueger (1994)
- Now the most widely used quasi-experimental method

Figure IV: Quasi-Experimental Methods



Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show 5-year moving averages.

## Why an entire workshop on DiD?

- **Research advantages:** DiD is often one of the only ways to study large social policies (e.g., decriminalized prostitution [Cunningham and Shah 2018])
- **Worrisome news:** Many new papers suggest canonical methods are biased, maybe severely biased
- **Good news:** Many new solutions and widely available code in both R and Stata
- **Econometrics:** It's always fun to learn econometrics

## Pedagogy of the seminar

- Emphasis on assumptions and authors
- It can feel like drinking from a firehose to learn so many papers
- I can't really advise you on how these are connected to one another, as that level of depth I'm still working on myself

## Potential outcomes review

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

## Introduction to Counterfactuals and Causality

- Aliens come and orbit earth, see sick people in hospitals and conclude “these ‘hospitals’ are hurting people”
- Motivated by anger and compassion, they kill the doctors to save the patients
- Sounds stupid, but earthlings do this too - all the time
- Let’s look at the challenges of making causality synonymous with correlations

## #1: Correlation and causality are very different concepts

These are not the same thing:

- Causal question: “If a doctor puts a person with Covid on a ventilator ( $D$ ), will her health ( $Y$ ) improve?”
- Correlation question:

$$\frac{Cov(D, Y)}{\sqrt{Var_D} \sqrt{Var_Y}}$$

## #2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- If the feral cat had killed the rooster the sun would have still risen, so coming first must not be enough
- *Post hoc ergo propter hoc*: “after this, therefore, because of this”



### #3: No correlation does not mean no causality!

- A sailor sails her sailboat across a lake
- Wind blows, and she perfectly counters by turning the rudder
- The same aliens observe from space and say “Look at the way she’s moving that rudder back and forth but going in a straight line. That rudder is broken.” So they send her a new rudder
- They’re wrong but why are they wrong? There is, after all, no correlation
- Question: What if she had been moving the rudder by flipping coins?

## Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 & \text{if hospitalized at time } t \\ 0 & \text{if not hospitalized at time } t \end{cases}$$

where  $i$  indexes an individual observation, such as a person

- Potential outcomes:

$$Y_{i,t}^j = \begin{cases} 1 & \text{health if hospitalized at time } t \\ 0 & \text{health if not hospitalized at time } t \end{cases}$$

where  $j$  indexes a counterfactual state of the world

- I'll drop  $t$  subscript, but note – these are potential outcomes for the same person at the exact same moment in time

## Moving between worlds

- A potential outcome  $Y^1$  and a historical outcome  $Y$  are neither conceptually nor notationally the same thing
- Potential outcomes are *hypothetical* possibilities describing states of the world but historical outcomes actually occurred
- We choose among potential outcomes by selecting the treatment

## Important definitions

### Definition 1: Individual treatment effect

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

### Definition 3: Switching equation

An individual's observed health outcomes,  $Y$ , is determined by treatment assignment,  $D_i$ , and corresponding potential outcomes:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

### Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all  $i$  individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

## So what's the problem?

Definition 4: Fundamental problem of causal inference

If you need both potential outcomes to know causality with certainty, then since it is impossible to observe both  $Y_i^1$  and  $Y_i^0$  for the same individual,  $\delta_i$ , is *unknowable*.

## Conditional Average Treatment Effects

### Definition 5: 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:

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

### Definition 6: Average Treatment Effect on the Untreated (ATU)

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

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

|                                 |                            |
|---------------------------------|----------------------------|
| Basics                          |                            |
| Covariates                      |                            |
| Weighted Group-Time ATT         | Simple case, no covariates |
| Stacking                        | IPW                        |
| Imputation DiD                  | DRDiD                      |
| Alternative estimators          |                            |
| Basic suggestions going forward |                            |

## John Snow and cholera

- John Snow, epidemiologist in 19th century, usually credited with first use of DiD
- Believed cholera was spread through the Thames water supply which contradicted dominant theory about “dirty air” transmission
- Grand experiment: Lambeth moves its pipe between 1849 and 1854; Southwark and Vauxhall delay
- How can he use this event to test his hypothesis? Three ways: simple comparisons, interrupted time series or the difference in differences (DiD)

## Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

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

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

## Interrupted time series design

Table: Lambeth, 1849 and 1854

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

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

## Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

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

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

## Sample averages

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

## 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

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

## Parallel trends bias

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

## OLS Specification

- Properly specified OLS model will also identify the ATT when there is only two groups and no covariates
- Often preferred because
  - OLS estimates the ATT under parallel trends
  - Easy to calculate the standard errors
  - Easy to include multiple periods
- But some issues emerge with differential timing, covariates and continuous treatments

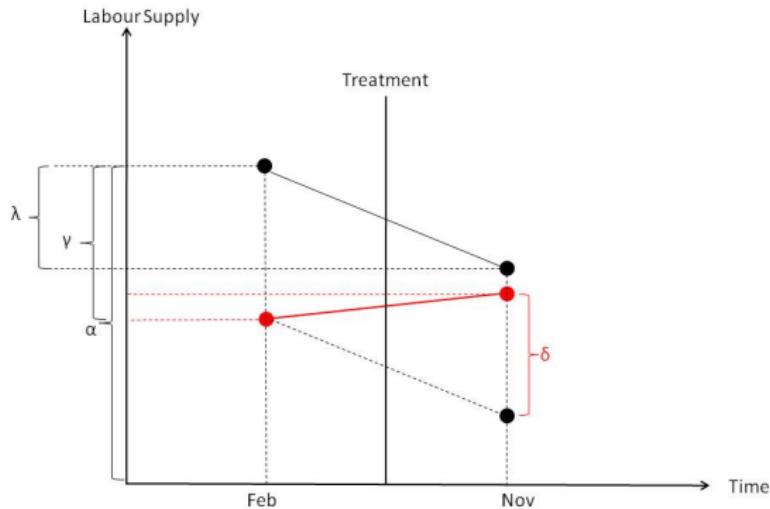
## Regression DiD - Card and Krueger

- The equivalent regression includes time and group fixed effects:

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

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

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



## **OLS with two way fixed effects**

Under parallel trends, OLS estimates the ATT. Researchers often will use OLS with time-varying covariates, but this is not advised as it is only unbiased under more restrictive assumptions which we discuss next

Basics  
**Covariates**  
Weighted Group-Time ATT  
Stacking  
Imputation DiD  
Alternative estimators  
Basic suggestions going forward

Simple case, no covariates  
IPW  
DRDiD

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

## Inverse probability weighting DiD

- Abadie (2005) proposed a DiD estimator that could incorporate covariates and get an unbiased estimate of the ATT
- Researcher needs treatment and comparison group observed before and after treatment
- If treatment group units are selected based on their (observed) covariates, then baseline covariates are also needed
- No randomization is needed; just another version of parallel trends called conditional parallel trends

## Time varying versus time invariant covariates

- In a DiD, we may need to control for  $X$  because treatment is only conditional on  $X$
- But in TWFE, all time invariant covariates are absorbed by the unit fixed effects – only time varying covariates will survive TWFE
- But time varying covariates place restrictions, as we will see, on the DGP and run the threat of conditioning on outcomes if they were changed by the treatment
- Abadie proposes using only the covariates at baseline to form weights in the simple DiD formula

## Three step method

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

You can have heterogeneous treatment effects, but not differential timing

## Terms

- $t$  is year of treatment which doesn't vary across units (so no differential timing)
- $Y^1$  and  $Y^0$  are potential outcomes (counterfactual versus actual)
- $D$  is 1 or 0 based on group and time
- $b$  is the “baseline” which is similar to CS using  $g$  as the one year pre-treatment
- $X$  are “baseline” covariates **only** – they do not vary over time, which means propensity scores are estimated off the  $b$  period **only**

## Assumptions

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

- ① Conditional parallel trends

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

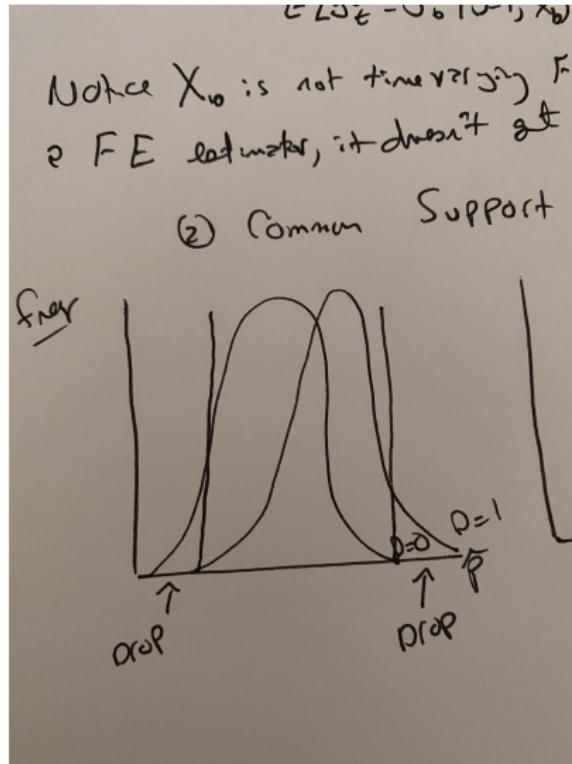
(Notice the  $b$  subscript. What is that you think?)

- ② Common support

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

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

## Trimming the propensity score to get common support



## Definition and estimation

Defining the ATT parameter of interest

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

Abadie's estimator

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

These are also using the “Hajek” (non-normalized) weights from the inverse probability weighting literature

## Propensity scores

- Paper is titled “Semi-parametric DiD” because Abadie imposes structure on the polynomials used to construct the propensity score
- You can use OLS linear probability models or series logit estimation

## Estimating propensity scores

It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions

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

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

## Stata

Stata command is called `absdid`

You need treatment (varname),  $X$  variables (can be a list), the order in which the variables occur (weird, but results change if the order changes), and the exact estimator (LPM or logit)

Why not try it yourselves using the LaLonde NSW job trainings program data?

[https://github.com/scunning1975/mixtape/raw/master/nsw\\_mixtape.dta](https://github.com/scunning1975/mixtape/raw/master/nsw_mixtape.dta)

|                                 |                            |
|---------------------------------|----------------------------|
| Basics                          |                            |
| Covariates                      |                            |
| Weighted Group-Time ATT         | Simple case, no covariates |
| Stacking                        | IPW                        |
| Imputation DiD                  | DRDiD                      |
| Alternative estimators          |                            |
| Basic suggestions going forward |                            |

## Doubly Robust Difference-in-differences

- DR models control for covariates twice – once using the propensity score, once using outcomes adjusted by regression – and are unbiased so long as:
  - The regression specification for the outcome is correctly specified
  - The propensity score specification is correctly specified
- Sant'Anna and Zhao (2020) incorporated DR into DiD by combining inverse probability weighting and outcome regression into a single DiD model
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so it merits close study
- One of my favorite lesser known of the new DiD papers

## Defining the target parameter – the ATT

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

## Basic assumptions of DiD

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

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

## Basic assumptions of DD

Assumption 2: Conditional parallel trends

Counterfactual trends for the treatment group are the same as the control group for all values of  $X$

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

## Basic assumptions of DD

Assumption 3: Common support or overlap

For some  $e > 0$ , the probability of being in the treatment group is greater than  $e$  and the probability of being in the treatment group conditional on  $X$  is  $\leq 1 - e$ .

Intuition of assumption 3: Called overlap or common support.  
Means there is at least a small fraction of the population that is treated and that for every value of the covariates  $X$  there is at least a small chance that the unit is not treated. It's called common support when it's a propensity score but it's just about the distribution of treatment and control across values of  $X$ .

## Estimating DD with Assumptions 1-3

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

## Outcome regression

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

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

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

## Inverse probability weighting

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

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

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

## These models cannot be ranked

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

## TWFE

Here's the TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose three *more* assumptions

## Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

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

Conditional parallel trends implies

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

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

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

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

Last line from the switching equation. This gives us:

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

## Collecting terms

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

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

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

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

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

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

#### **Assumption 4: Homogeneous treatment effects in $X$**

TWFE requires homogenous treatment effects in  $X$  (i.e., the treatment effect is the same for all  $X$ )

If  $X$  is sex, then effects are the same for males and females.

If  $X$  is continuous, like income, then the effect is the same whether someone makes \$1 or \$1 million.

## X-specific trends

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

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

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

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

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

## X-specific trends

Now take the DiD formula:

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

Eliminating terms, we get:

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

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

## **Assumption 5 and 6**

We need “no X-specific trends” for the treatment group (assumption 5) and comparison group (assumption 6)

**Intuition:** No X-specific trends means the evolution of potential outcome  $Y^0$  is the same regardless of  $X$ . This would mean you cannot allow rich people to be on a different trend than poor people, for instance.

Without these six, in general TWFE will not identify ATT.

## Why not both?

- Let's review the problem. What if you claim you need  $X$  for conditional parallel trends?
- You have three options:
  - Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
  - Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
  - TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one
- I'm going to only stick to the panel data expressions bc all repeated cross-section does is add in some terms (and I've not written up semiparametric bounds yet)

## Notation

$p(x)$  : propensity score model

$$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}(X)$ , where  $\mu(X)$  is a model for

$$m_{d,t} = E[Y_t | D = d, X = x]$$

So that means  $\mu_{0,\Delta}$  is just the control group's change in average  $Y$  for each  $X = x$

## Population DR DiD model for panel data

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

Notice how the model controls for  $X$ : you're weighting the adjusted outcomes using the propensity score

The reason you control for  $X$  twice is because you don't know which model is right. DR DiD frees you from making a choice without making you pay too much for it

## Efficiency

- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DiD estimator is also DR for inference
- Let's skip to Monte Carlos

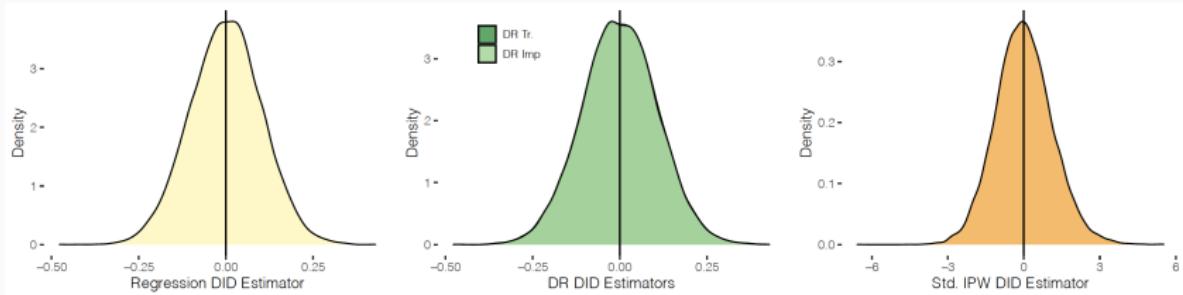
## Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

**Table:** Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

|      | Bias     | RMSE    | SE     | Coverage | CI length |
|------|----------|---------|--------|----------|-----------|
| TWFE | -20.9518 | 21.1227 | 2.5271 | 0.000    | 9.9061    |
| OR   | -0.0012  | 0.1005  | 0.1010 | 0.9500   | 0.3960    |
| IPW  | 0.0257   | 2.7743  | 2.6636 | 0.9518   | 10.4412   |
| DR   | -0.0014  | 0.1059  | 0.1052 | 0.9473   | 0.4124    |

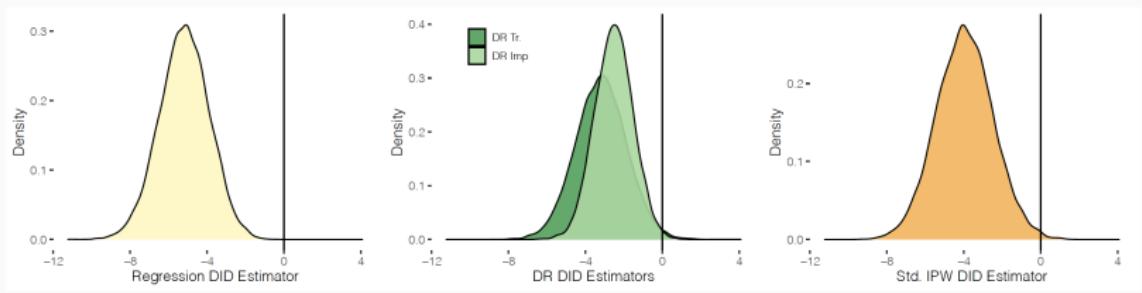
**Figure 1:** Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified



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

|      | Bias     | RMSE    | SE     | Coverage | CI length |
|------|----------|---------|--------|----------|-----------|
| TWFE | -16.3846 | 16.5383 | 3.6268 | 0.000    | 14.2169   |
| OR   | -5.2045  | 5.3641  | 1.2890 | 0.0145   | 5.0531    |
| IPW  | -1.0846  | 2.6557  | 2.3746 | 0.9487   | 9.3084    |
| DR   | -3.1878  | 3.4544  | 1.2946 | 0.3076   | 5.0749    |

**Figure 4:** Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



## Code

There is code in R and Stata

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

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

## Concluding remarks

- These two papers mark a different approach than is often the case for applied researchers who simply estimate regression models and hope they recover “reasonably weighted” causal effects
- These new DiD start with target parameter and identification then build estimation
- TWFE, as it turns out, is not mostly harmless

## Differential timing

- We covered mostly the simple two group case
- In the two group case, we can estimate the ATT under parallel trends using OLS with unit and time fixed effects
- If we have covariates, then we can use TWFE under restrictive assumptions, or we have other options (OR, IPW, DR)
- Now let's move to a more common scenario where we have more than two groups who get treated at various times

## 2x2 versus differential timing

- For this next part, similar to how we did with Sant'Anna and Zhao (2020), we will decompose TWFE to understand what it needs for unbiasedness under differential timing
- All of this is from Goodman-Bacon (2021, forthcoming) though the expression of the weights is from 2018 for personal preference
- Goodman-Bacon (2021, forthcoming) shows that parallel trends is **not enough** for TWFE to be unbiased when treatment adoption is described by differential timing
- TWFE with differential timing uses treated groups as controls – not all estimators do – and this can introduce bias

## Decomposition Preview

- TWFE estimates a parameter that is a weighted average over all 2x2 in your sample
- TWFE assigns weights that are a function of sample sizes of each “group” and the variance of the treatment dummies for those groups

## Decomposition (cont.)

- TWFE needs two assumptions: that the variance weighted parallel trends are zero (far more parallel trends iow) and no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs

$K^2$  distinct DDs

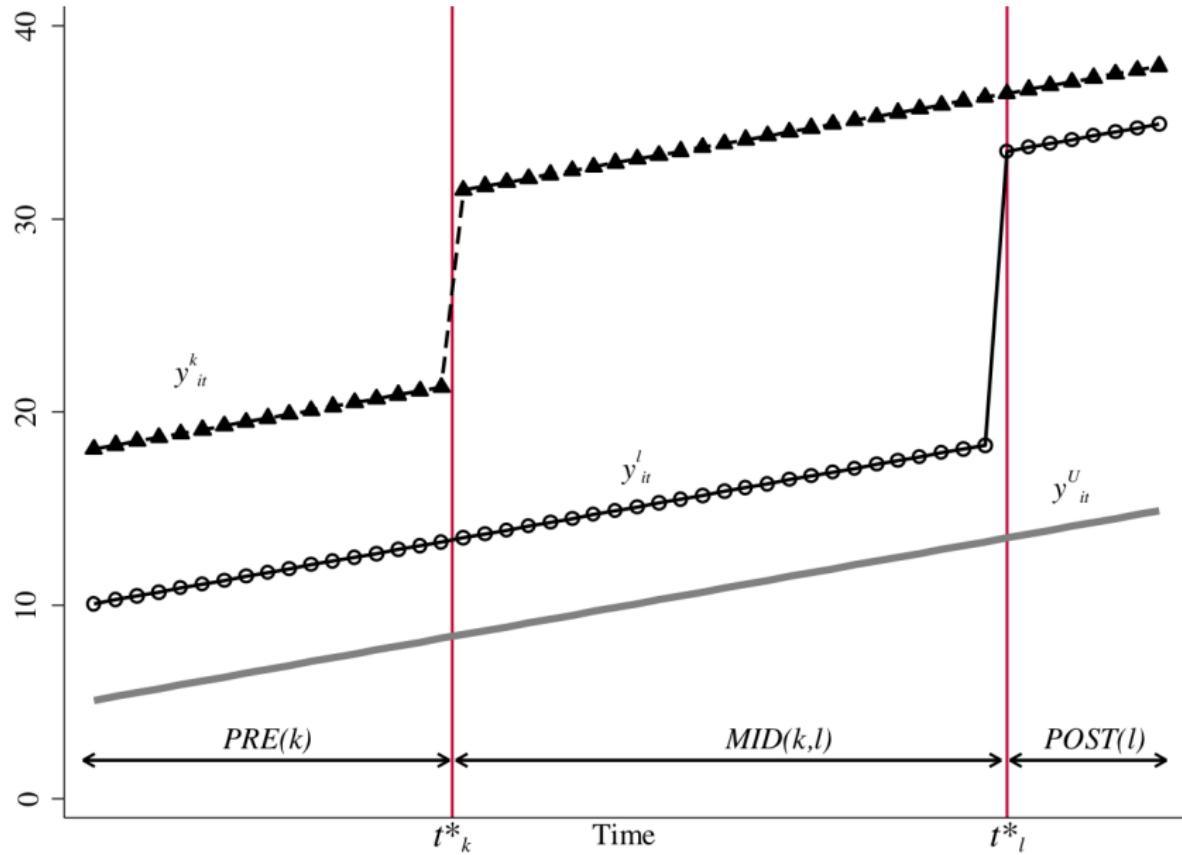
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

|        |        |        |
|--------|--------|--------|
| 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 |

Let's return to a simpler example with only two groups – a  $k$  group treated at  $t_k^*$  and an  $l$  treated at  $t_l^*$  plus an never-treated group called the  $U$  untreated group

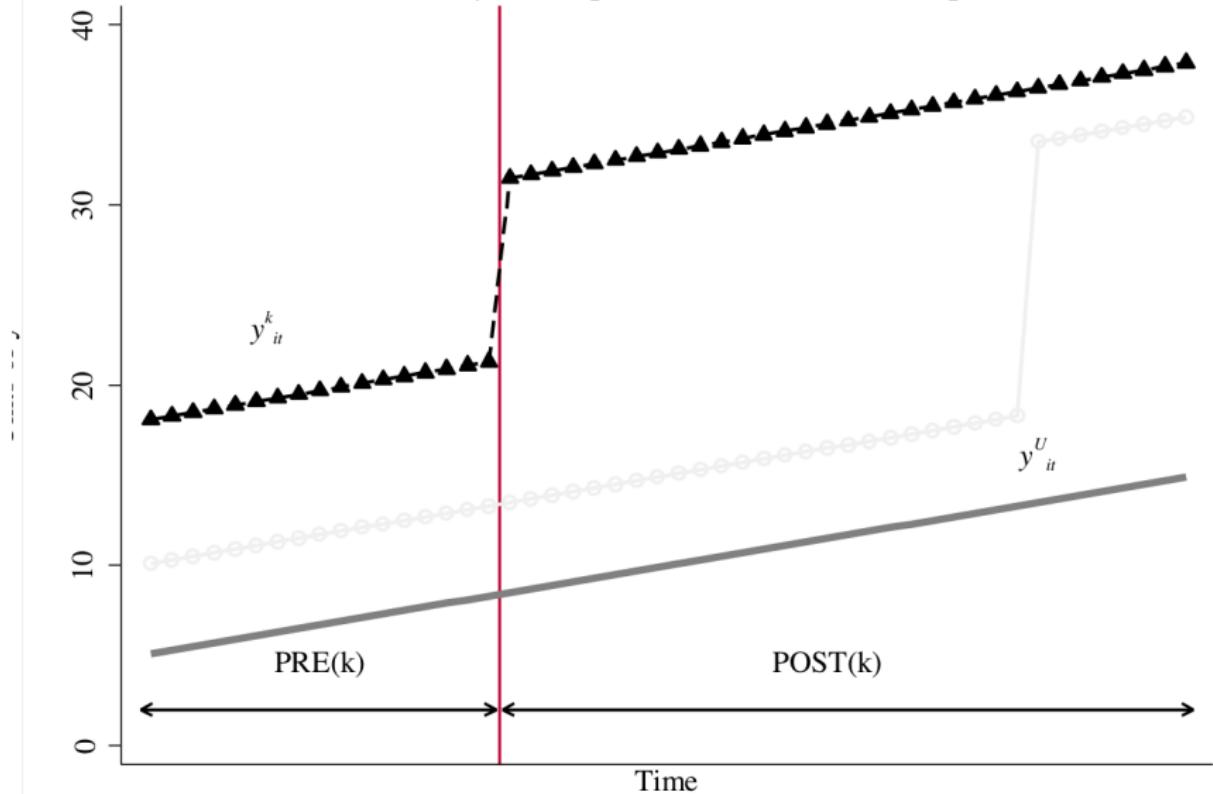
## Terms and notation

- Let there be 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 earlier than  $l$
- Denote  $\bar{D}_k$  as the share of time each group spends in treatment status
- Denote  $\widehat{\delta}_{jb}^{2 \times 2}$  as the canonical  $2 \times 2$  DD estimator for groups  $j$  and  $b$  where  $j$  is the treatment group and  $b$  is the comparison group



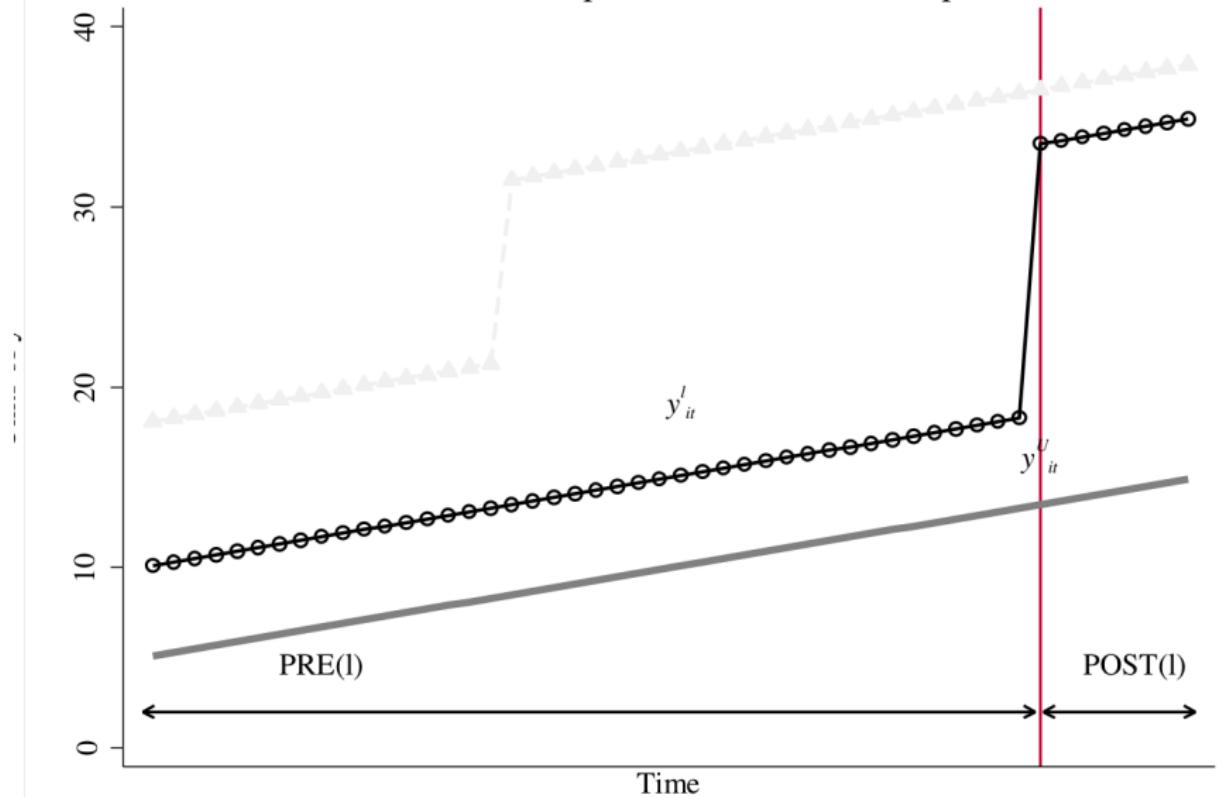
$$\widehat{\delta}_{kU}^{2 \times 2} = \left( \bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left( \bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

A. Early Group vs. Untreated Group

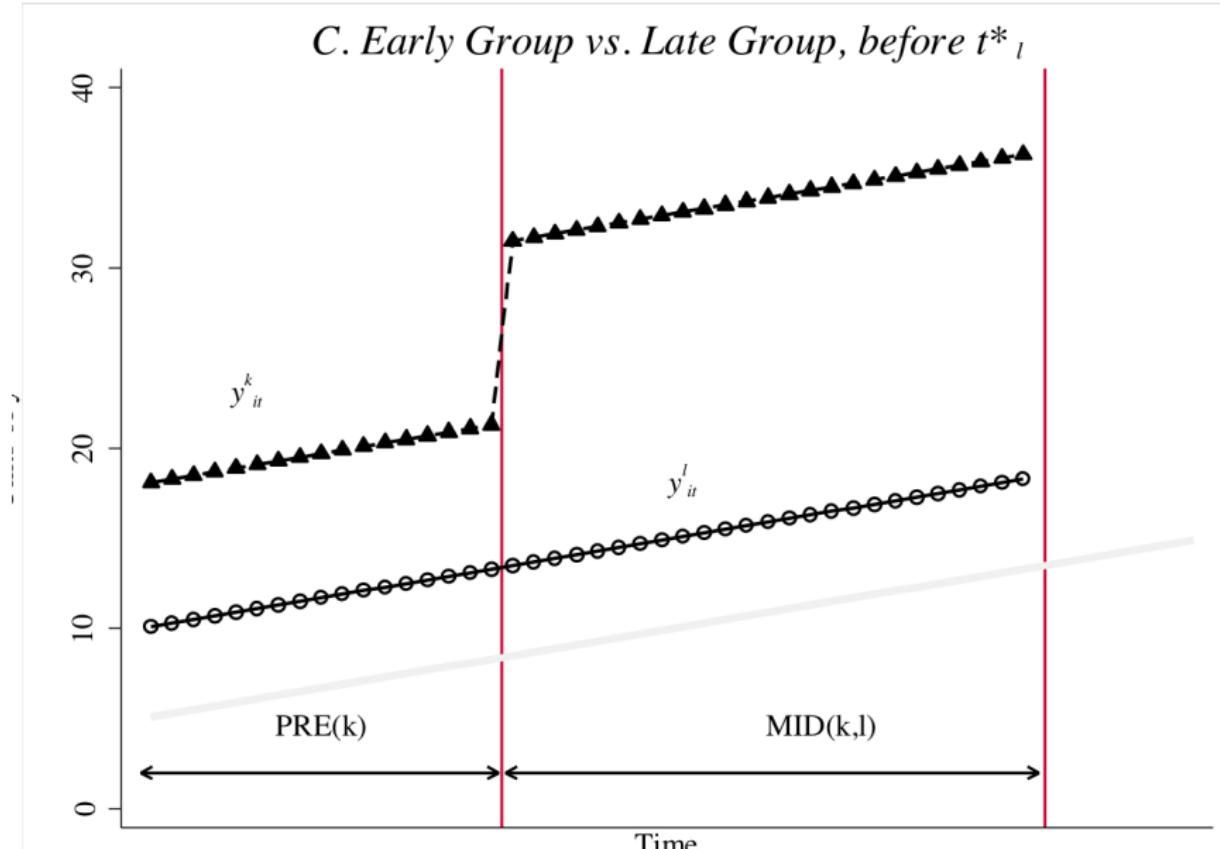


$$\widehat{\delta}_{IU}^{2 \times 2} = \left( \bar{y}_I^{post(I)} - \bar{y}_I^{pre(I)} \right) - \left( \bar{y}_U^{post(I)} - \bar{y}_U^{pre(I)} \right)$$

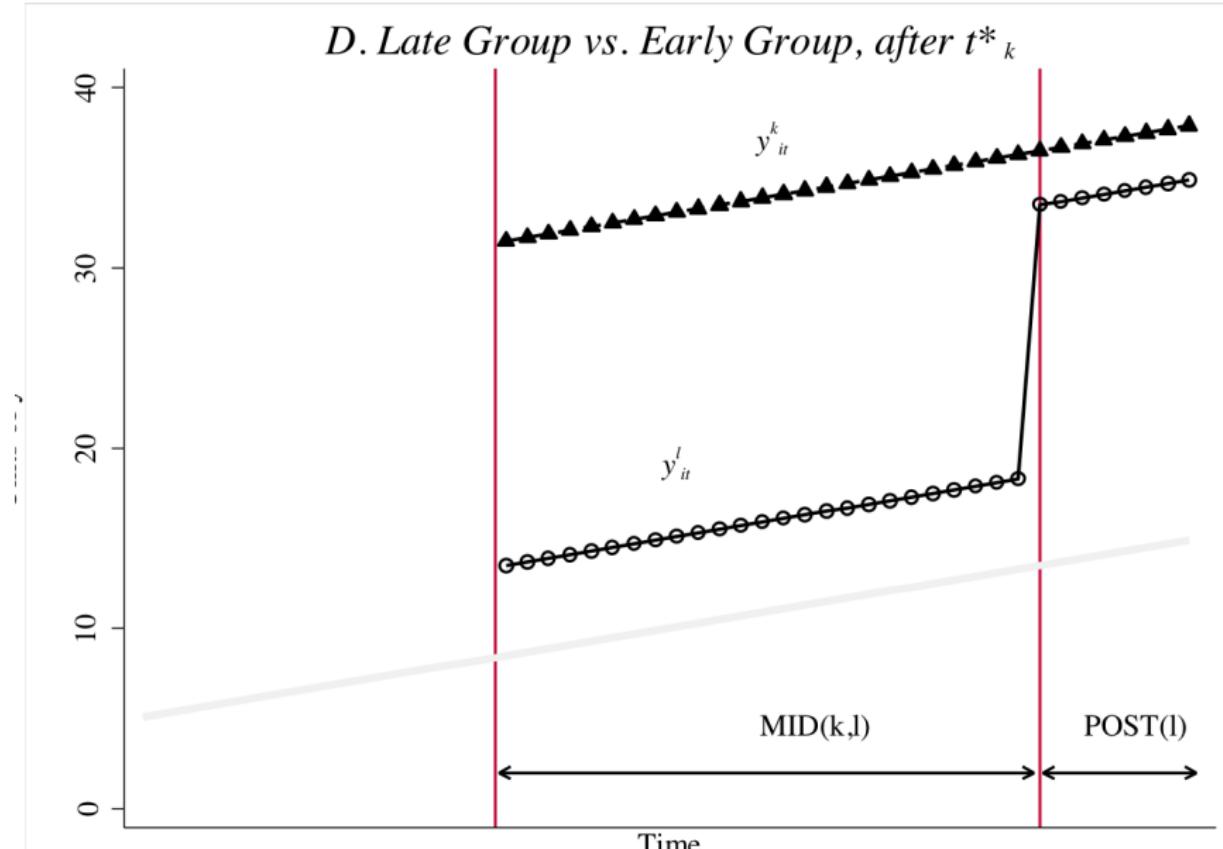
B. Late Group vs. Untreated Group



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



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



## Bacon decomposition

TWFE 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 combines the k compared to U and the l 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 to note:
  - More units in a group, the bigger its 2x2 weight is
  - Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the  $s_{ku}$  weights.
  - $\bar{D} = 0.1$ . Then  $0.1 \times 0.9 = 0.09$
  - $\bar{D} = 0.4$ . Then  $0.4 \times 0.6 = 0.24$
  - $\bar{D} = 0.5$ . Then  $0.5 \times 0.5 = 0.25$
  - $\bar{D} = 0.6$ . Then  $0.6 \times 0.4 = 0.24$
- This means 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 (i.e.,  $\bar{D}_k - \bar{D}_I$ )
- 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  $0.52 \times 0.48 = 0.2496$ .

## Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

## Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

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

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 we get:

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

**Substitute all this stuff into the decomposition formula**

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

where we will make these substitutions

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

Notice all those potential sources of biases!

## Potential Outcome Notation

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

- 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 parallel trends

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

There are  $K^2$  parallel trends inside the weights. Their weighted average must equal zero.

## 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

## Callaway and Sant'Anna 2020

- New papers are coming out focused on the issues that we are seeing with TWFE
- I'll discuss one though by Callaway and Sant'anna (2020) due to time constraints (call it CS)
- If we have time, I'll run through a simulation illustrating both the bias of TWFE and the unbiased estimation of this CS estimator
- Interesting ancestry – CS is a descendent of Abadie (2005) from earlier

## Preliminary

CS considers identification, aggregation, estimation and inference procedures for ATT in DD designs with

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

## When might you use this estimator

Probably in the very situations describing your own study

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

**Group-time ATT is the parameter of interest in CS**

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

## Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- CS will calculate an ATT per group/time which will be the sum of all  $T - t_k$  for all groups (i.e., a lot)
- Group-time ATT estimates are not determined by the estimation method one adopts (first difference or FE) bc they are simple differences in means
- 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 ATT
- Inference is the bootstrap

## 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” (can be relaxed though to “not yet treated”)
  - Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = \Pr(G_g = 1 | X, G_c + C = 1)$$

## Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

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

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

## CS Estimator (the IPW version)

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

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. Notice how CS doesn't use already-treated as controls.

## Staggered adoption (i.e., 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”
- Inference from a bootstrap

## **Stata simulation**

Let's now review a simulation in Stata which can be downloaded from my github repo called `baker.do`.

## Pedro Sant'Anna for the win

- Now a word from a good friend – Pedro Sant'Anna. Legend!
- He'll be discussing deChaisemartin and D'Haultfoeiller (2020) because:
  - He's a good guy
  - He's a great presenter
  - I fell behind

|                                 |       |
|---------------------------------|-------|
| Basics                          |       |
| Covariates                      |       |
| <b>Weighted Group-Time ATT</b>  | Bacon |
| Stacking                        | CS    |
| Imputation DiD                  | dCH   |
| Alternative estimators          | SA    |
| Basic suggestions going forward |       |

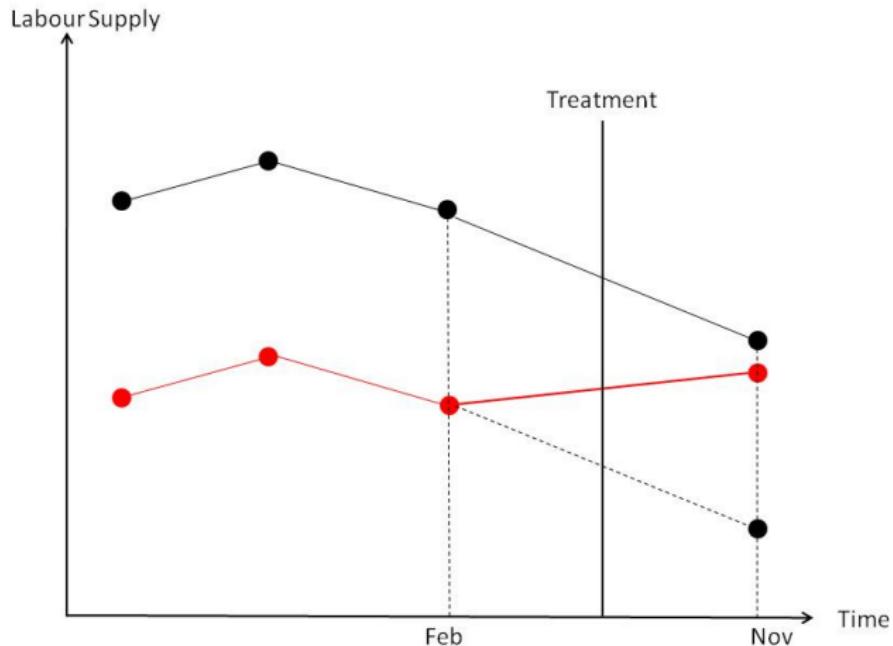
## de Chaisemartin and D'Haultfoeuille

Now a word from our sponsor (Pedro San'tAnna)

## 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

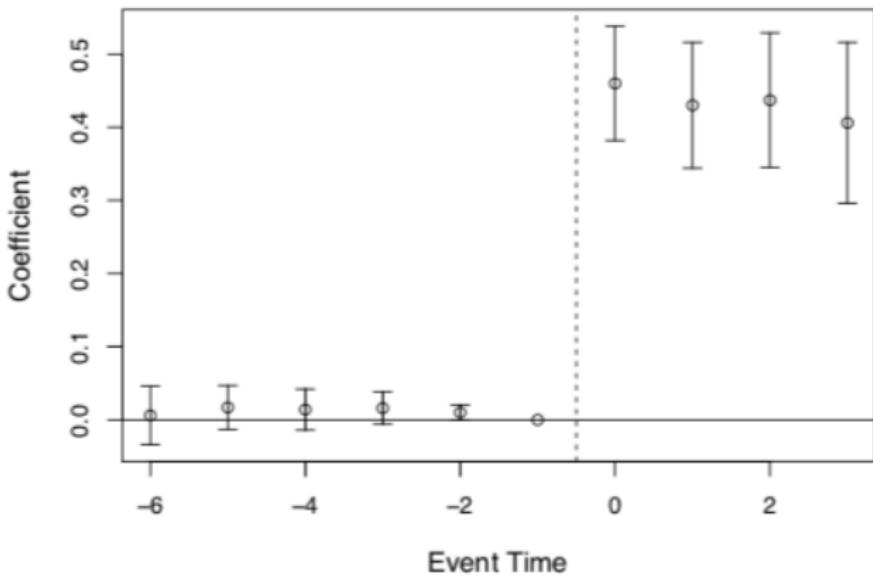
- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

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

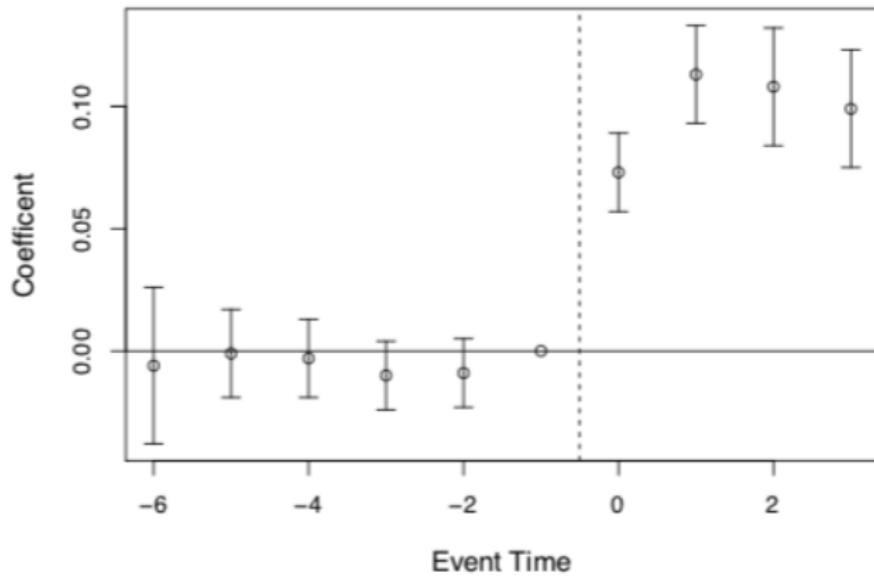
- Treatment occurs in year 0
- Includes  $q$  leads or anticipatory effects
- Includes  $m$  leads or post treatment effects

## **Medicaid and Affordable Care Act example**

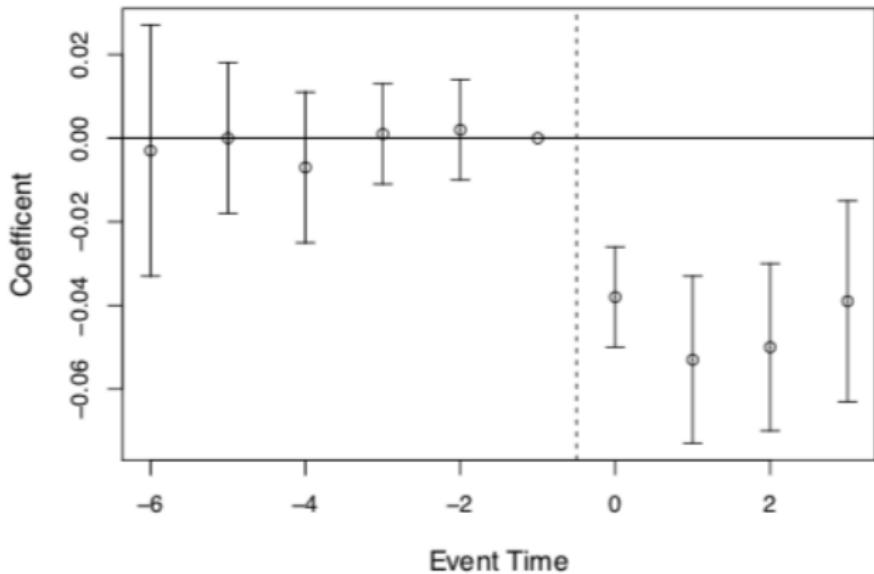
- Miller, et al. (2019) examine a rollout of Medicaid under the Affordable Care Act
- They link large-scale survey data with administrative death records
- 9.3 reduction in annual mortality caused by Medicaid expansion
- Driven by a reduction in disease-related deaths which grows over time



(a) Medicaid Eligibility



(b) Medicaid Coverage



(c) Uninsured

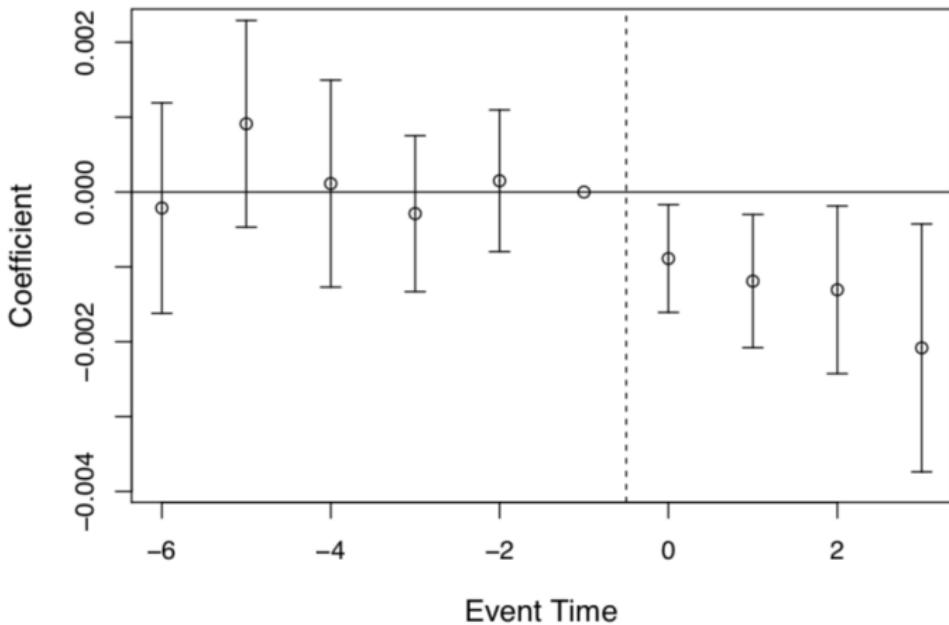


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on on annual mortality

## Differential timing complicates plotting sample averages

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

## Early efforts at event studies

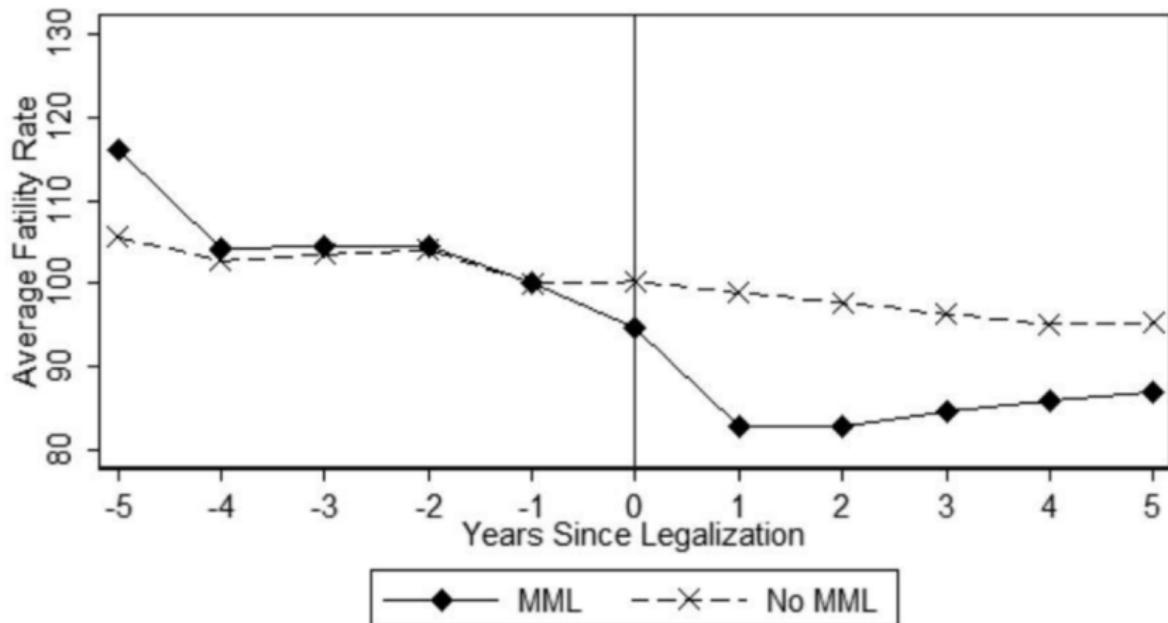


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

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

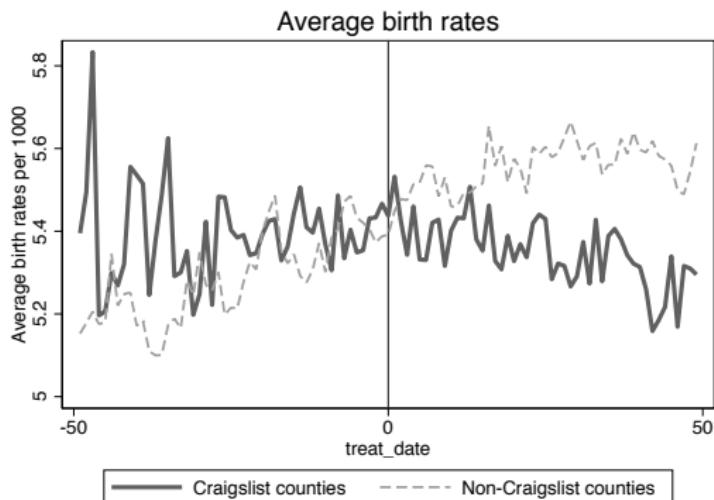
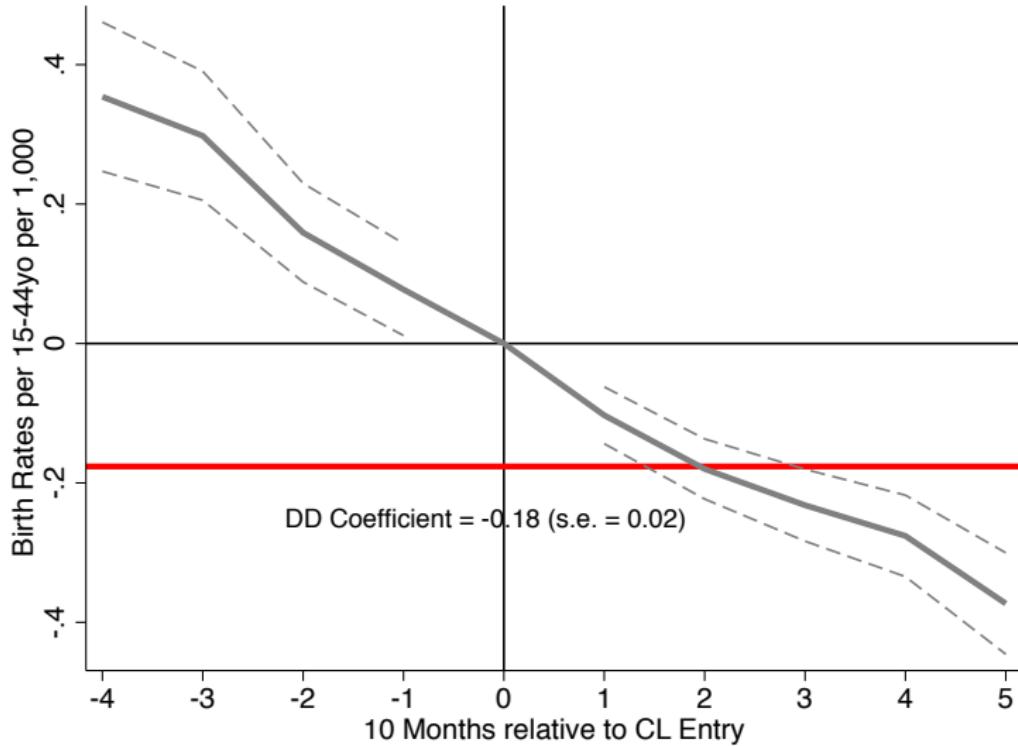


Figure: From one of my studies. Looks decent right?



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

## Sun and Abraham 2020

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

## Summarizing

- Goodman-Bacon (2021, forthcoming) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

## Summarizing (cont.)

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

## Summarizing (cont.)

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

## Summarizing (cont.)

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

## Difficult notation sadly

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

## Difficult notation (cont.)

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

## Difficult notation (cont.)

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

## More difficult notation (cont.)

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

## **Definition 1**

**Definition 1:** The cohort-specific ATT / periods from initial treatment date  $e$  is:

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

## Identifying assumption 1

**Assumption 1: Parallel trends in baseline outcomes:**

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

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

## Assumption 2

**Assumption 2: No anticipator behavior in pre-treatment periods:** There is a set of pre-treatment periods such that  $E[Y_{i,e+I}^e - Y_{i,e+I}^\infty | E_i = e] = 0$  for all possible leads.

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

### Assumption 3

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

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

## Treatment effect heterogeneity

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

## TWFE Regression

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

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

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

## Specifying the leads and lags

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

- ① Static specification:

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

- ② Dynamic specification:

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

## Multicollinearity

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

## Trimming and binning

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

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

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

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

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

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

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

## Estimating the weights

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

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

## Interpretation of coefficients under parallel trends only

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

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

## Comment on Proposition 2

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

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

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

Weights can help us gauge the severity of this problem.

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

## Interpretation under parallel trends and no anticipation

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

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

### Proposition 3 comment

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

## Proposition 4

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

$$\begin{aligned}\mu_g &= \sum_{I \in g} w_I^g ATT_I \\ &+ \sum_{g' \neq g} \sum_{I' \in g'} w_{I'}^{g'} ATT_{I'} \\ &+ \sum_{I' \in g^{excl}} w_{I'}^{g'} ATT_{I'}\end{aligned}$$

## Proposition 4 comment

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

$$D'_{i,t} = \alpha_i + \lambda_t + \sum_{g \in G} w_I^g \cdot 1\{t - E_i \in g\} + u_{i,t}$$

which regresses  $D'_{i,t}$  on all bin indicators and TWFE

## On binning

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

## Intuition for contamination

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

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

## Toy example

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

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

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

## Interaction-weighted estimator

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

## Interaction-weighted estimator

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

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

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

## Interaction-weighted estimator

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

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

## IW estimator

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

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

## Consistency and Inference

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

## DD Estimator of CATT

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

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

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

## Software

- **Stata:** eventstudyinteract (can be installed from ssc)
- **R:** did2s (see  
[https://asjadnaqvi.github.io/DiD/docs/02\\_R/](https://asjadnaqvi.github.io/DiD/docs/02_R/))

## Conclusion of SA

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

## Keep TWFE but avoid differential timing problems

- Problem with TWFE occurred when we used already-treated as controls
- CS and SA used the never-treated (CS and SA), not-yet-treated (CS) or last-treated (SA) as controls
- Each identified group-time ATT (“cohort-specific ATT”), then via weighting based on group shares, aggregate treatment parameters (e.g., ATT)
- Could we still use TWFE, but avoid the use of already-treated as controls? How? Interpretation?
- Cengiz, et al. (2019) call this “clean controls” but now it’s called “stacking”

## Stacking minimum wages

- Discuss the Cengiz, et al. (2019) as an example of stacking, even though it is a very small part of the paper
- It's a good paper at illustrating an argument built up through tables, figures and various forms of intuitive reasoning that guides estimation
- Data is 1979 to 2016 US state-level panel, 138 “prominent” state-level minimum wage change events
- Main model specification is TWFE, but in a robustness section (Appendix D) they introduce a stacking alternative

## Background

- Theory: minimum wages should reduce employment in perfectly competitive labor markets
- Theory: minimum wages will increase employment in monopsony labor markets
- Prior research: “new minimum wage” studies starting with Card and Krueger (1995) found no effect of minimum wages on employment, but others (often authored by David Neumark) found reductions
- Focus had often been on aggregate employment or teens
- Controversial, contentious and unsettled

## Data

- Hourly wage data from 1976-2016 NBER out-rotation group of the Current Population Survey broken into wage bins (by \$0.25) from \$0 to \$30

*"We use the individual-level NBER Merged Outgoing Rotation Group of the CPS for 1979-2016 to calculate quarterly, state-level distributions of hourly wages. For hourly workers, we use the reported hourly wage, and for other workers, we define the hourly wage to be their usual weekly earnings divided by usual weekly hours. We do not use any observations with imputed wage data to minimize the role of measurement error."*

- Wages are deflated to 2016 dollars to get “real wage”
- There are 117 wage bins that are then collapsed into quarterly, state-level employment counts  $E_{swt}$  using person-level sampling weights

## Minimum wage data

- Quarterly max of the state-level daily minimum wage series from Vaghul and Zipperer (2016)
- 138 minimum wage events, of which 8.6% of workers were below the minimum wage the year before the event
- Focus will be on “missing” vs “excess” jobs which is a break in the wage frequency just above and just below the minimum wage cutoff

## Bite

Constant reference to “bite”. Footnote 2.

*“When we refer to the ‘bite’ of the minimum wage, or to the extent to which the minimum wage is ‘binding’, we mean how effective the minimum wage is in raising wages at the bottom. Therefore, the bite is a function of (i) how many workers are earning below the minimum wage, (ii) how many of those workers are legally covered by the policy, and (iii) the extent of compliance.”*

You can sometimes hear it referred to as the first stage, using IV language, but as this is a DiD and not IV they tend to use the word “bite” over and over instead.

The idea is that this paper only makes sense if the policy change has “bite”

## Rhetorical argument

- Key visual showing excess vs missing jobs helps communicate the idea of “bite” which supports all subsequent analysis
- If no bite, no effects shown can be plausibly causal – so notice, the logic of identification comes from convincingly showing bite, not from the model results themselves
- Carefully construct data into bins so that narrowly employment above and below the minimum wage can be measured
- Do the same for employment per population
- Calculate net changes in employment both for total (main results) and by various slices (heterogenous effects)

# Bunching

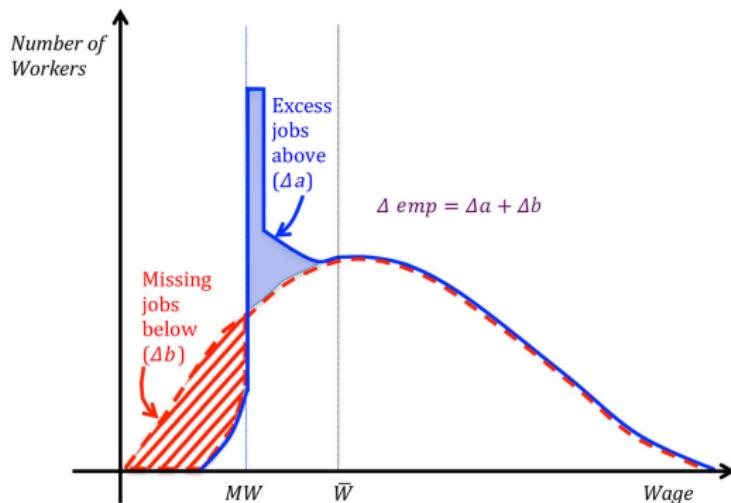


FIGURE I

The Impact of Minimum Wages on the Frequency Distribution of Wages

## Null results

- Results we will now see is essentially a lot of zeroes
- Authors fail to find any evidence of an effect on employment other than the “bite” result which I’ll show
- Empirical equation is estimated using TWFE; differential timing problems,  $\tau$  are event years and  $k$  are dollar bins,  $\mu$  and  $\rho$  are state-by-wage-bin and period-by-wage-bin fixed effects,  $\omega$  are controls for small or federal increases

$$E/N = \sum_{\tau=-3}^4 \sum_{k=-4}^{17} \alpha_{\tau k} I_{sjt}^{\tau k} + \mu_{sj} + \rho_{jt} + \omega_{sjt} + u_{sjt} \quad (5)$$

# Results

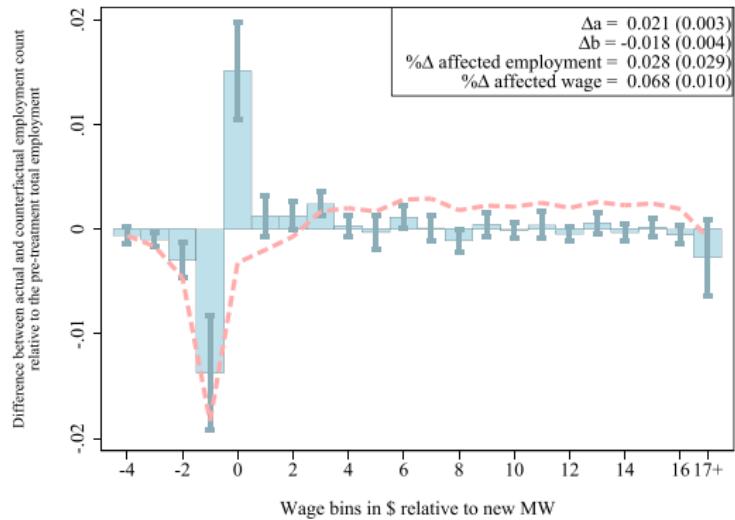
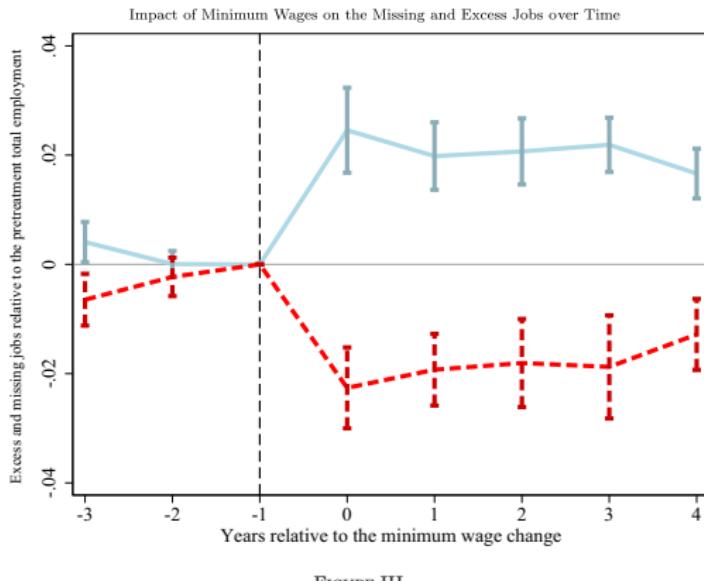


FIGURE II  
Impact of Minimum Wages on the Wage Distribution

# Results



## Results

TABLE I  
IMPACT OF MINIMUM WAGES ON EMPLOYMENT AND WAGES

# Results

TABLE II  
IMPACT OF MINIMUM WAGES ON EMPLOYMENT AND WAGES BY DEMOGRAPHIC GROUPS

|  | (1)                  | (2)                  | (3)                  | (4)                  | (5)                  | (6)                  | (7)                  | (8)                  |
|--|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| Missing jobs below new MW ( $\Delta b$ ) | -0.065***<br>(0.010) | -0.032***<br>(0.007) | -0.114***<br>(0.010) | -0.023***<br>(0.005) | -0.028***<br>(0.008) | -0.094***<br>(0.010) | -0.020***<br>(0.005) | -0.004***<br>(0.001) |
| Excess jobs above new MW ( $\Delta a$ )  | 0.075***<br>(0.011)  | 0.038***<br>(0.006)  | 0.127***<br>(0.020)  | 0.026***<br>(0.004)  | 0.028***<br>(0.006)  | 0.100***<br>(0.012)  | 0.021***<br>(0.003)  | 0.004***<br>(0.001)  |
| % $\Delta$ affected wages                | 0.080***<br>(0.014)  | 0.076***<br>(0.014)  | 0.083***<br>(0.018)  | 0.072***<br>(0.011)  | 0.044***<br>(0.012)  | 0.073***<br>(0.011)  | 0.051***<br>(0.013)  | 0.060***<br>(0.032)  |
| % $\Delta$ affected employment           | 0.038<br>(0.024)     | 0.043<br>(0.030)     | 0.030<br>(0.032)     | 0.025<br>(0.027)     | -0.004<br>(0.044)    | 0.015<br>(0.018)     | 0.015<br>(0.048)     | 0.011<br>(0.055)     |
| Employment elasticity w.r.t. MW          | 0.097<br>(0.061)     | 0.061<br>(0.042)     | 0.125<br>(0.134)     | 0.025<br>(0.027)     | -0.005<br>(0.058)    | 0.052<br>(0.062)     | 0.016<br>(0.049)     | 0.003<br>(0.014)     |
| Emp. elasticity w.r.t. affected wage     | 0.475*<br>(0.268)    | 0.570<br>(0.386)     | 0.356<br>(0.317)     | 0.343<br>(0.362)     | -0.086<br>(1.005)    | 0.206<br>(0.233)     | 0.304<br>(0.904)     | 0.184<br>(0.841)     |
| Jobs below new MW ( $\bar{b}_{-1}$ )     | 0.264                | 0.145                | 0.432                | 0.102                | 0.133                | 0.358                | 0.104                | 0.027                |
| % $\Delta$ MW                            | 0.103                | 0.103                | 0.102                | 0.101                | 0.100                | 0.103                | 0.103                | 0.103                |

# Results

TABLE II  
CONTINUED

|                                 | (1)                      | (2)                    | (3)     | (4)       | (5)                  | (6)                 | (7)                   | (8)                |
|---------------------------------|--------------------------|------------------------|---------|-----------|----------------------|---------------------|-----------------------|--------------------|
| Number of events                | 138                      | 138                    | 138     | 138       | 138                  | 138                 | 138                   | 138                |
| Number of observations          | 847,314                  | 847,314                | 847,314 | 847,314   | 846,729              | 847,314             | 847,314               | 847,314            |
| Number of workers in the sample | 660,771                  | 2,248,711              | 287,484 | 2,277,624 | 781,003              | 469,226             | 1,830,393             | 2,349,485          |
| Sample                          | Less than<br>high school | High school<br>or less | Teen    | Women     | Black or<br>Hispanic | High<br>probability | Medium<br>probability | Low<br>probability |

*Notes.* The table reports effects of a minimum wage increase by demographic groups based on the event study analysis (see equation (1)) exploiting 138 state-level minimum wage changes between 1979 and 2016. The table reports five-year averaged post-treatment estimates on missing jobs up to \$4 below the new minimum wage, excess jobs at and up to \$5 above it, employment, and wages for individuals without a high school degree (column (1)), for individuals with high school degree or less schooling (column (2)), for teens (column (3)), for women (column (4)), for black or Hispanic workers (column (5)). Columns (6)–(8) report the results for groups of workers with differential probability of being exposed to the minimum wage changes. We use the Card and Krueger (1995) demographic predictors to estimate the probability of being exposed (see the text for details). Column 6 shows the results for the workers who have a high probability of being exposed to the minimum wage increase, column (7) for the middle-probability group, and column (8) for the low-probability group. All specifications include wage bin-by-state and wage bin-by-period fixed effects. Regressions are weighted by state-quarter aggregated population of the demographic groups. Robust standard errors in parentheses are clustered by state; significance levels are \*0.10, \*\*0.5, \*\*\*0.01.

The first two rows report the change in number of missing jobs below the new minimum wage ( $\Delta b$ ), and excess jobs above the new minimum wage ( $\Delta a$ ) relative to the pretreatment total employment. The third row, the percentage change in average wages in the affected bins, ( $\% \Delta W$ ), is calculated using equation (2) in Section 2.2. The fourth row, percentage change in employment in the affected bins, is calculated by dividing change in employment by jobs below the new minimum wage ( $\frac{\Delta a + \Delta b}{b_{-1}}$ ). The fifth row, employment elasticity with respect to the minimum wage, is calculated as  $\frac{\Delta a + \Delta b}{\% \Delta M W}$ , whereas the sixth row, employment elasticity with respect to the wage, reports  $\frac{1}{\% \Delta W} \frac{\Delta a + \Delta b}{b_{-1}}$ . The line on the number of observations shows the number of quarter-bin cells used for estimation, while the number of workers refers to the underlying CPS sample used to calculate job counts in these cells.

## Results

**TABLE III**  
**IMPACT OF MINIMUM WAGES ON EMPLOYMENT AND WAGES BY SECTORS (1992–2016)**

# Results

TABLE III  
CONTINUED

|                                 | (1)       | (2)       | (3)          | (4)          | (5)       | (6)         | (7)     | (8)           |
|---------------------------------|-----------|-----------|--------------|--------------|-----------|-------------|---------|---------------|
| Number of events                | 118       | 118       | 118          | 118          | 118       | 118         | 118     | 118           |
| Number of observations          | 554,931   | 554,931   | 554,931      | 554,931      | 554,931   | 554,931     | 554,931 | 554,931       |
| Number of workers in the sample | 2,652,792 | 358,086   | 384,498      | 274,812      | 1,504,643 | 156,634     | 315,397 | 349,749       |
| Sector                          | Overall   | Tradeable | Nontradeable | Construction | Other     | Restaurants | Retail  | Manufacturing |

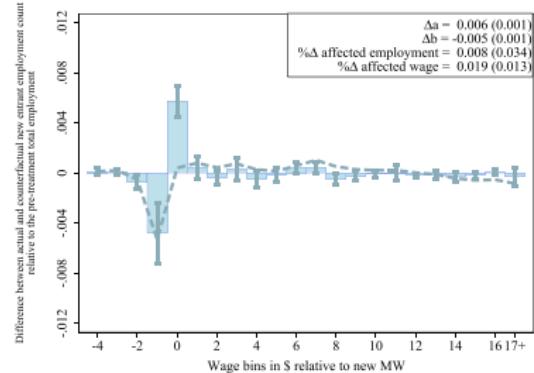
*Notes.* The table reports the effects of a minimum wage increase by industries based on the event study analysis (see [equation \(1\)](#)) exploiting 138 state-level minimum wage changes between 1992 and 2016. The table reports five-year averaged post-treatment estimates on missing jobs up to \$4 below the new minimum wage, excess jobs at and up to \$5 above it, employment, and wages for all sectors (column 1), tradable sectors (column 2), nontradable sectors (column 3), construction (column 4)), other sectors (column 5), restaurants (column 6)), retail (column 7), and manufacturing industries (column 8)). Our classification of tradable, nontradable, construction, and other sectors follows [Mian and Sufi \(2014\)](#) (see [Online Appendix D](#) for the details). Regressions are weighted by state-quarter aggregated population. Robust standard errors in parentheses are clustered by state; significance levels are \*0.10, \*\*0.05, \*\*\*0.01.

The first two rows report the change in number of missing jobs below the new minimum wage ( $\Delta b$ ), and excess jobs above the new minimum wage ( $\Delta a$ ) relative to the pretreatment total employment. The third row, the percentage change in average wages in the affected bins, ( $\% \Delta W$ ), is calculated using [equation \(2\)](#). The fourth row, percentage change in employment in the affected bins, is calculated by dividing change in employment by jobs below the new minimum wage ( $\frac{\Delta a + \Delta b}{b_{-1}}$ ). The fifth row, employment elasticity with respect

to the minimum wage, is calculated as  $\frac{\Delta a + \Delta b}{\% \Delta MW}$ , whereas the sixth row, employment elasticity with respect to the wage, reports  $\frac{1}{\% \Delta W} \frac{\Delta a + \Delta b}{b_{-1}}$ . The line on the number of observations shows the number of quarter-bin cells used for estimation, while the number of workers refers to the underlying CPS sample used to calculate job counts in these cells.

# Results

(A) New entrants



(B) Incumbents

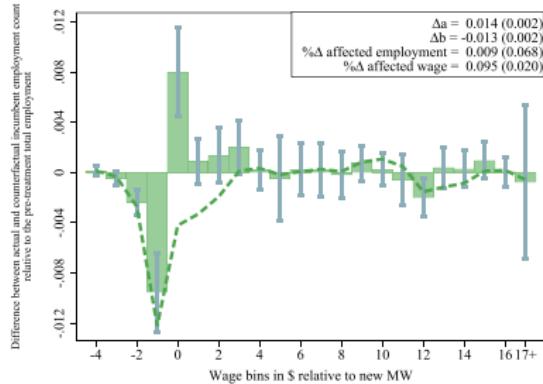


FIGURE IV

Impact of Minimum Wages on the Wage Distribution by Pretreatment Employment Status: New Entrants and Incumbents

## Stacking alternative

- TWFE estimation is biased if there are differential timing and heterogenous treatment effects by cohort (SA 2020)
- They propose their own alternative which they call “clean controls” but which is more commonly called “stacking”
- Estimation models are TWFE; weights were unknown at time of writing (but are now known via Gardner 2021)

## Dataset construction

Clean controls is done one of two ways:

- ① Create 138 datasets, one for each event  $h$  where the treatment group is one state and the control are all other states that did not have a minimum wage increase in eight-year panels around event  $h$ , balanced in calendar time, inference must adjust for heteroskedasticity with only one treatment date (Ferman and Pinto Restat)
- ② “Stack” the 138 datasets, re-centering each treatment date such that data is balanced in “event time” with 3 periods pre-treatment, 4 years post-treatment, and controls are all untreated units from -3 to +4. Several units will appear more than once).

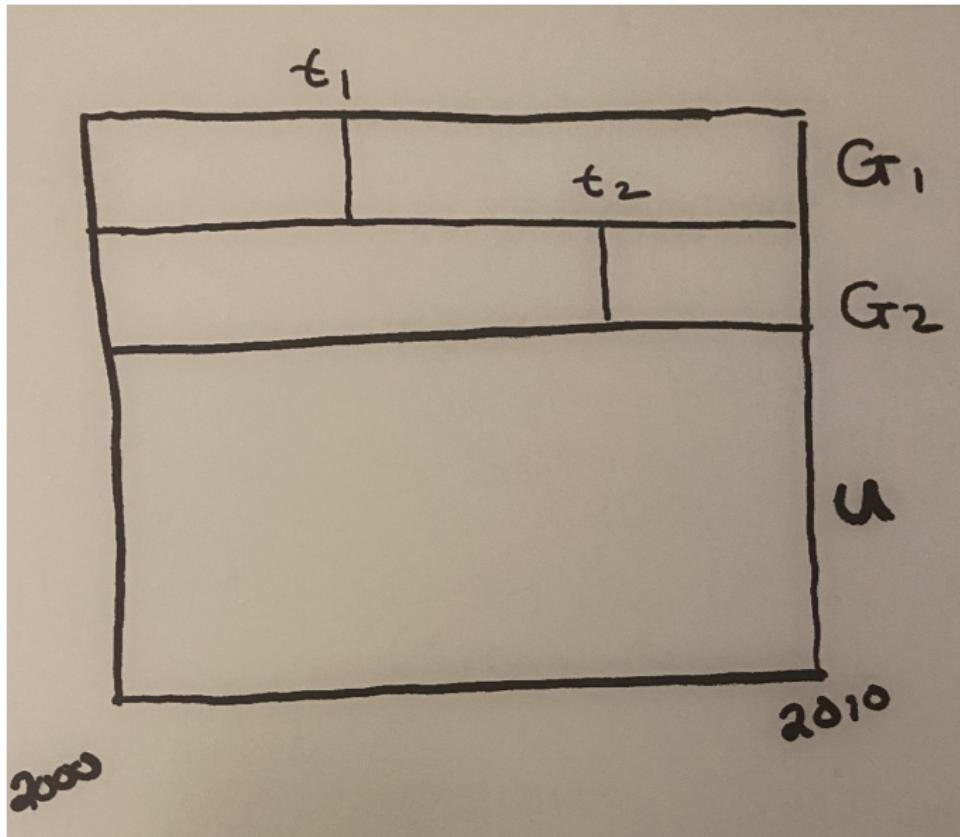
## Steps to stacked regression

- ① Create separate “event by cohort specific” datasets for each policy cohort (e.g., groups who pass minimum wages in the same year)
  - Dataset will consist of the relevant policy cohort **plus** controls
  - Data is structured in “event time” and will be balanced such that panel length is  $h$ , perhaps starting point being 3 years prior to treatment and ending point 4 years after or something like that
  - Each dataset will contain individuals untreated over the  $h$  period defined
- ② Append each dataset (or what people are now calling “stacking”) to one another
  - This necessarily replicates control observations though as they are in each datasets
  - Since the same people are often appearing many many times, you will correct for this in the regression model specification
- ③ Estimate a simple 2x2 model but include “cohort-by-state” fixed effects so as to account for the multiple appearances of observations from the never-treated control states

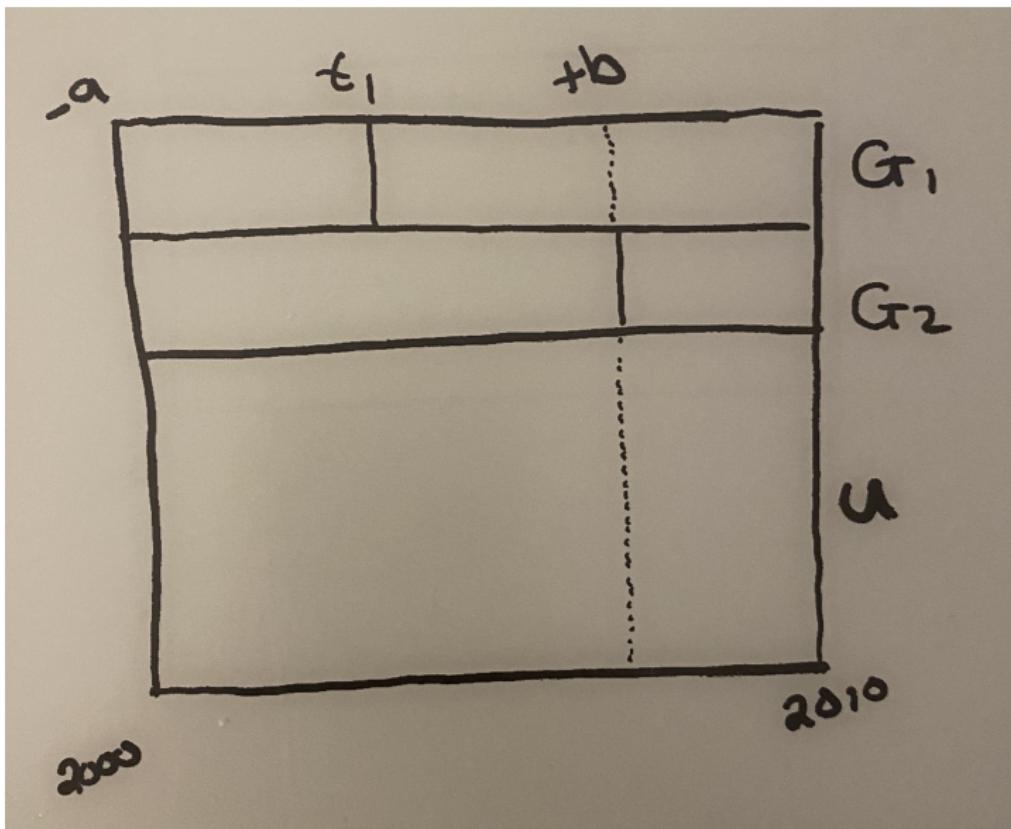
## Comments on the approach

- Hard part of this is probably just the careful balancing, saving datasets, then appending, but ultimately not difficult – just watch yourself.
- Because the data is now balanced *in event time*, there is *no differential timing*; it is a simple 2x2
- Recall that the reason TWFE is biased in DiD designs is (1) differential timing and (2) heterogeneity
- Stacked eliminates (1) making (2) irrelevant
- But recall the lessons we learned about including time-varying covariates from Sant'Anna and Zhao (2020) – stacked will suffer from those too

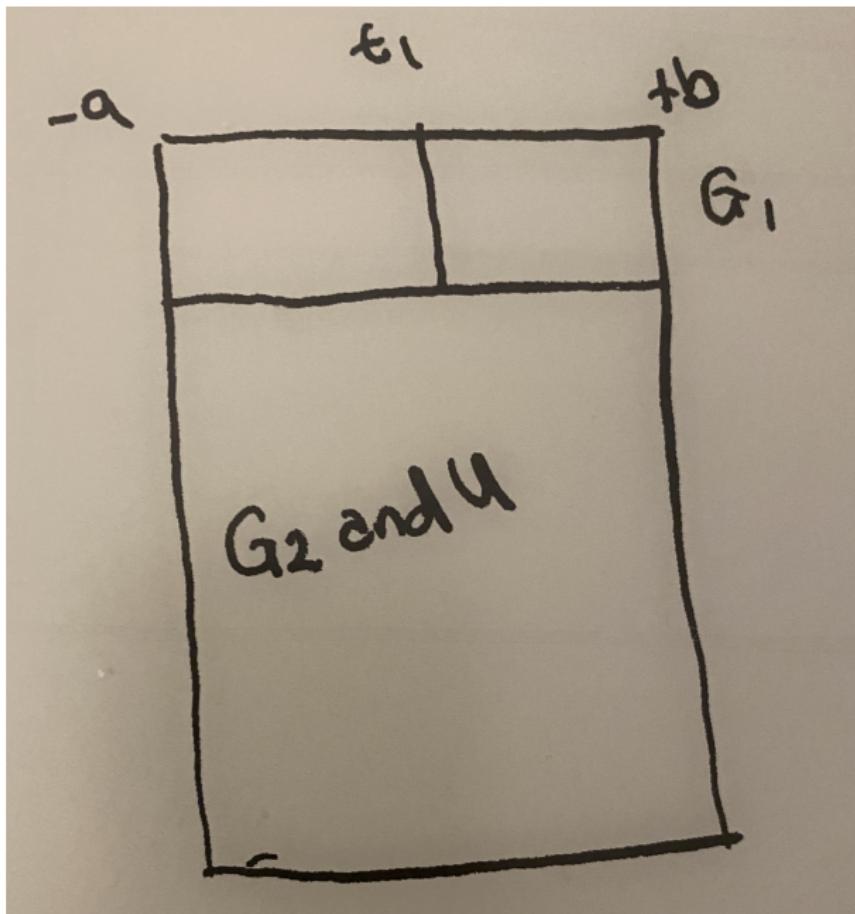
## Imbalanced in relative time with differential timing



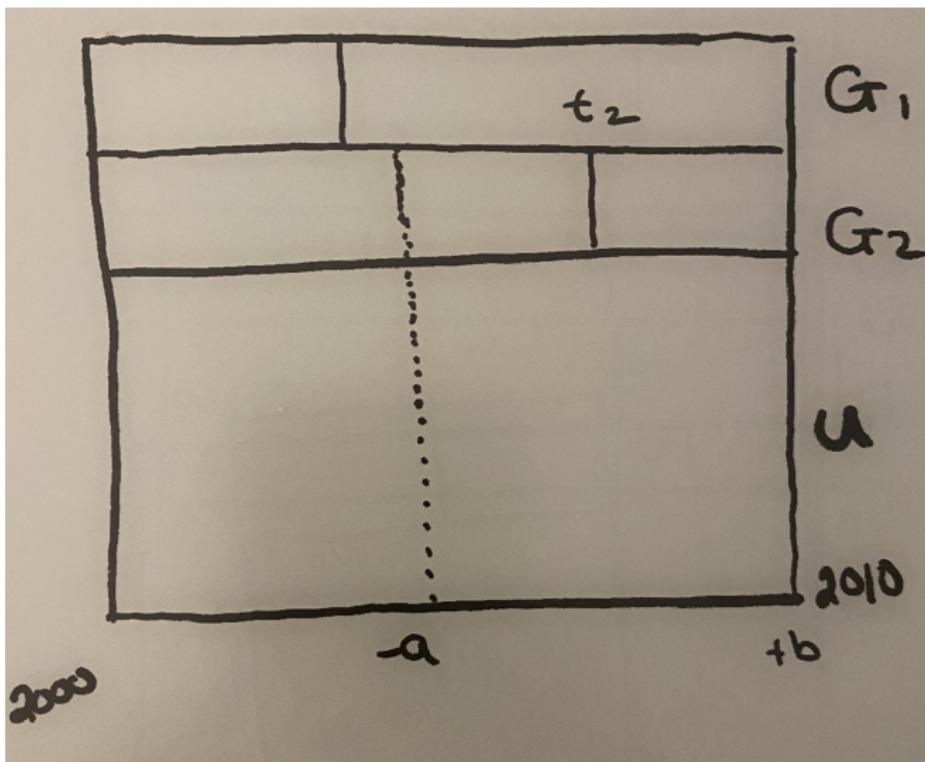
Creating  $G_1$  dataset: choose max pre (-a) and post (+b) periods



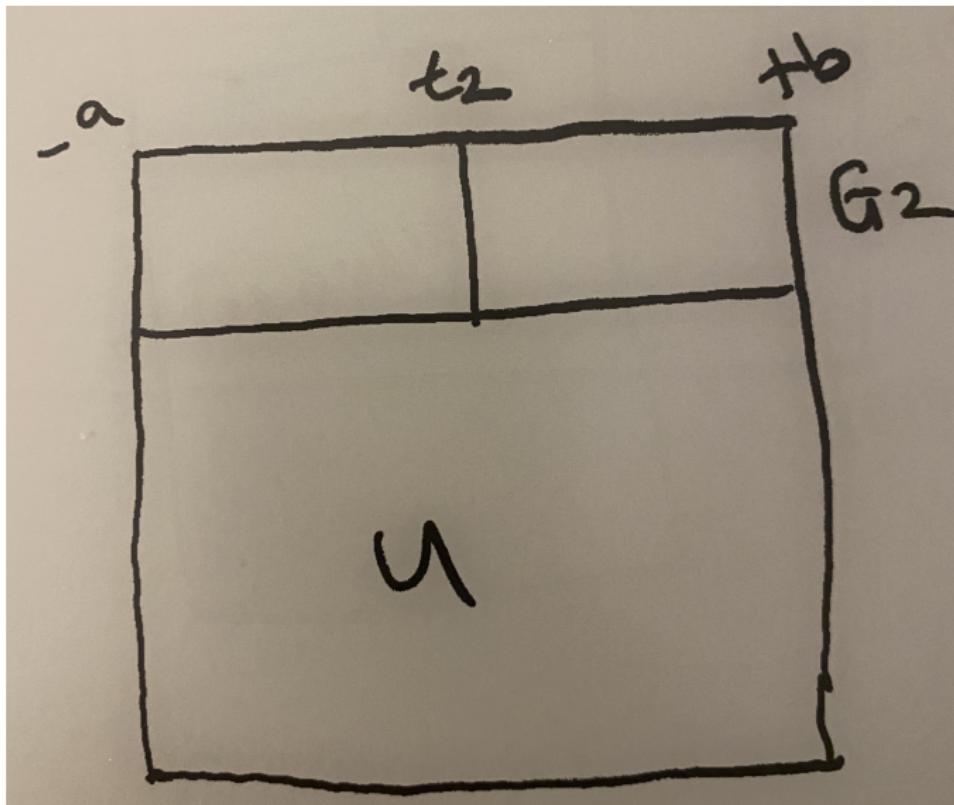
Creating  $G_1$  dataset: keep untreated units on  $[-a, +b]$



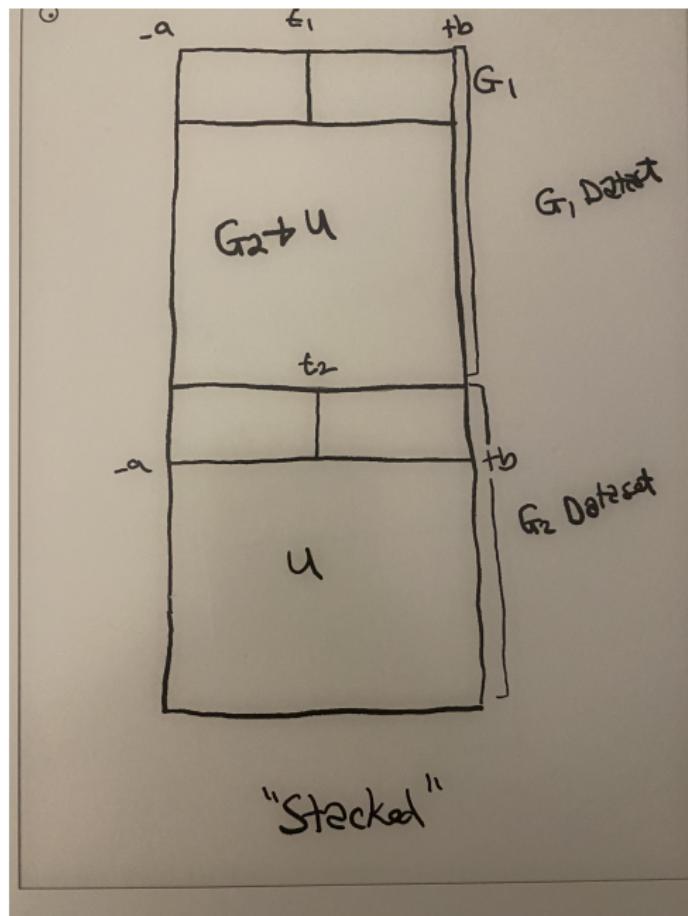
Creating  $G_2$  dataset: keep *only* untreated units on  $[-a,+b]$  intervals as controls



## Creating $G_2$ dataset: save the dataset



## Creating $G_2$ dataset: stack the datasets

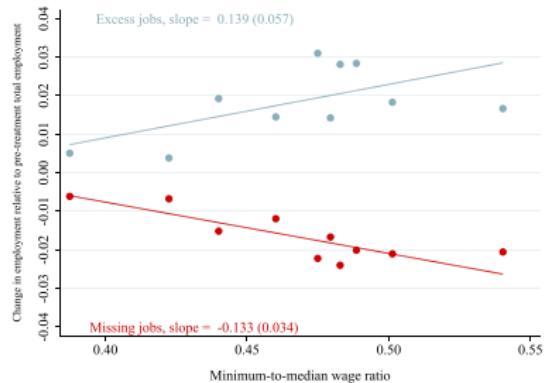


## Discussion of the stacked data

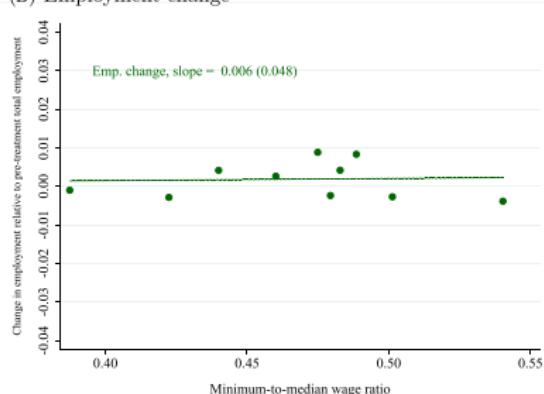
- Why doesn't  $G_1$  appear in  $G_2$ ?
- Notice that  $U$  appears in both  $G_1$  and  $G_2$  datasets.
- Unclear to me exactly what Cengiz, et al. (2019) run in their stacked regression, but we probably need to say it now that we want to have a control for “dataset-by-state” fixed effects some units appear more than once (e.g.,  $U$ ).

# Results

(A) Missing and excess jobs



(B) Employment change



# Results

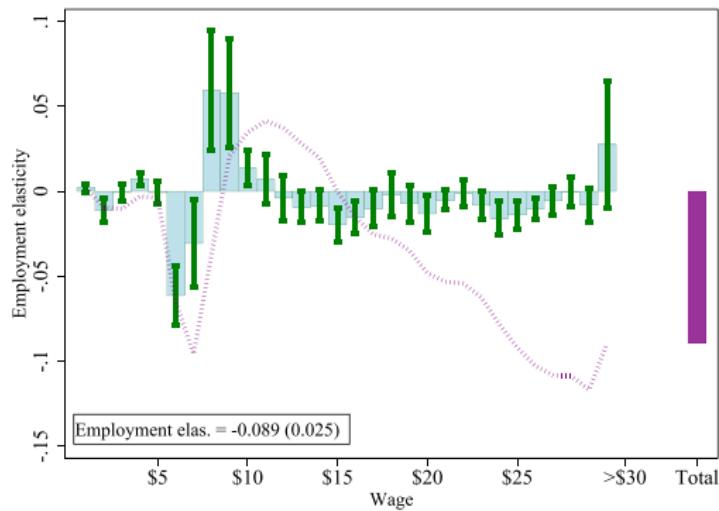
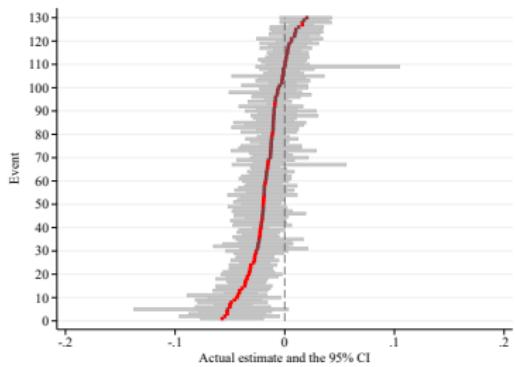


FIGURE VI

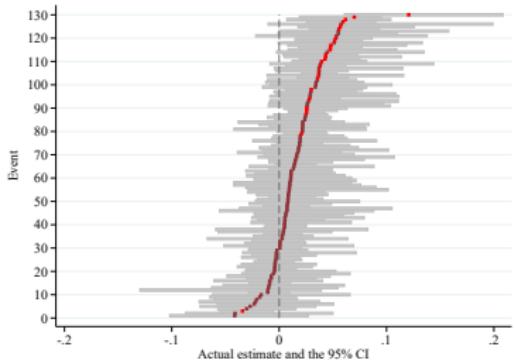
## **Upper part of wage distribution**

Logic is used to dismiss effects at upper part of distribution which is only place they find effects

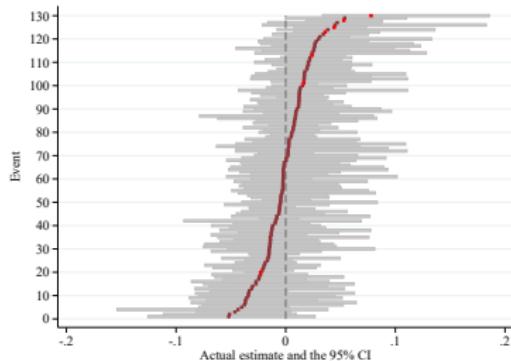
# Results



# Results



## Results



(c) Employment change ( $\Delta a + \Delta b$ )

## Unknown parameter

- Recall CS and SA start with parameter (e.g., group-time ATT) then build the estimator that finds it using aggregation
- Stacking goes in reverse: start with TWFE using a restructured dataset with “clean controls”
- So what is stacking identifying? VWATT? What are its assumptions? What are the weights?

## Identification

- Stacking is a TWFE estimation method, except that by balancing in relative event time, there is no longer any differential timing
- Thus identification requires a weighted parallel trends assumption, but how exactly?
- The parallel trends is *the same* as what we saw with our group-time representation
- Parallel is assumed to hold first *within* stacked dataset, not in the final regression model, which shapes the fixed effects we must employ to ensure it

## Gardner notation

$$Y_{cgpit} = \lambda_{cg} + \lambda_{cp} + \beta D_{cgph} + \varepsilon_{cgpit} \quad (6)$$

$c$  dataset;  $g$  group;  $p$  period;  $i$ th member of group  $g$ ;  $t$ th time period of period  $p$

$D_{cgp}$  is an indicator for whether group  $g$  is treated during period  $p$  of the group- $c$  dataset;  $\beta$  is the group-period ATT;  $\widehat{\beta}$  is estimate from TWFE

## Weighted ATT

$$\begin{aligned}\hat{\beta} &= \sum_{g=1}^G \sum_{p=g}^P w_{gp} \beta_{gp} \\ w_{gp} &= \frac{(1 - \pi_c) \pi_c \rho_c}{\bar{P} \sum_{c=1}^G (1 - \pi_c) \pi_c \rho_c}\end{aligned}$$

$\pi_c$  is the fraction of units treated in period  $p$  and  $\rho_c$  is the population share of observations for a given group/period

The stacked estimator weights each group's ATT by dataset-specific treatment variance and sample size which slightly overstates the true average bc  $\pi$  and  $\rho$  are both fractions.

Like Goodman-Bacon (2021), we see that the weights make  $\hat{\beta}$  slightly biased estimate of ATT.

## Heterogeneity is nevertheless important

- When theory predicts that longrun effects of a policy may differ from shorrun effects, then exploring heterogeneity is required
- Effects may also differ for size of policy which creates problems for SUTVA bc of “no hidden variation in treatment”
- But heterogeneity analysis, we said, can introduce p-hacking even subconsciously or due to the “garden of forking paths” (Gelman and Loken 2013)

## P-hacking

*"Here's the thing: P-values of .05 aren't that hard to find if you sort the data differently or perform a huge number of analyses. In flipping coins, you'd think it would be rare to get 10 heads in a row. You might start to suspect the coin is weighted to favor heads and that the result is statistically significant.*

*But what if you just got 10 heads in a row by chance (it can happen) and then suddenly decided you were done flipping coins? If you kept going, you'd stop believing the coin is weighted.*

*Stopping an experiment when a p-value of .05 is achieved is an example of p-hacking. But there are other ways to do it – like collecting data on a large number of outcomes but only reporting the outcomes that achieve statistical significance. By running many analyses, you're bound to find something significant just by chance alone."*

## Alternatives to p-hacking heterogeneity

- ① **Preregistration of study designs:** Scientists publicly commit to an experiment's design before the data collection phase or the analysis phase. Much harder to 'cherry pick' results
- ② **Open data sharing:** Journals are increasingly requiring that researchers post their data online, or submit to a data repository
- ③ **Replication:** Some journals are publishing replications (e.g., Nature's ReScienceX)

Clemens and Strain (2021) use preregistration which is rare in labor or quasi-experimental work but is very common in development economics using RCTs

## Researchers must make choices

- Nick Huntington-Klein and several others published a 2021 article in *Economic Inquiry* in which two applied microeconomics papers were assigned to individual researchers with the task of replicating the results from raw data
- Results were concerning:

*"We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error."*
- Sometimes it isn't the garden of forking paths or p-hacking that is the problem – there is also uncertainty driven by the numerous reasonable choices that researchers must make in order to do any analysis in the first place

## Pre-registration and heterogeneity analysis

- Theoretical reasons to suspect effects of the minimum wage may differ depending on whether the bite is large or small
- Because their analysis would be conducting heterogeneity analysis, and heterogeneity analysis was one of the causes of the replication crisis, they would pre-register
- Pre-registered design as extensions of their earlier short-run analysis

## Variation in treatment

*"Consider an assessment of the causal effect of aspirin on headaches. For the potential outcome with both of us taking aspirin, we obviously need more than one aspirin tablet. Suppose, however, that one of the tablets is old and no longer contains a fully effective dose, whereas the other is new and at full strength. In that case, each of us may have three treatments available: no aspirin, the ineffective tablet, and the effective tablet. There are thus two forms of the active treatment, both nominally labeled "aspirin": aspirin+ and aspirin-." (Imbens and Rubin 2015)*

## SUTVA

Stable Unit Treatment Value Assumption requires that an individual receiving a specific treatment level cannot receive different forms of that treatment (called the “no hidden variations of treatments” by Imbens and Rubin 2015)

*“One strategy to make SUTVA more plausible relies on redefining the represented treatment levels to comprise a larger set of treatments, for example, Aspirin-, Aspirin+ and no-aspirin instead of only Aspirin and no-aspirin.” (Imbens and Rubin 2015)*

## Great Recession

- Great Recession saw a pause in minimum wage hikes followed by large increases in several states
- Increases in Cengiz, et al. (2019) were around 8 log points on average; minimum wage increases after the Great Recession range from 25 log points to as high as 60 log points (Clemens and Strain 2021)
- “DC, CA and NY had increased their minimum wages by 61, 50 and 53 percent respectively.”

## Methodology and Data

- Data: American Community Survey and Current Population Survey:
  - Years: 2011-2015 with pre-registration commitment to study through 2019
- Methodologies:
  - TWFE event study
  - Stacked regression (Cengiz, et al. 2019; Baker, et al. 2020)
  - Imputation estimator (Borusyak, Jaravel and Spiess 2021)
- Unique characteristics of treatment: large and small increases will be modeled separately

## Summary of findings

- Large increases in minimum wages reduced employment rates among individuals with low levels of experience and education by just over 2.5pp
- Relatively small minimum wage increases are variable and centered on zero much like what Cengiz, et al. (2019) found
- Medium-run effects are larger and more negative than short-run effects

## TWFE Event study specification

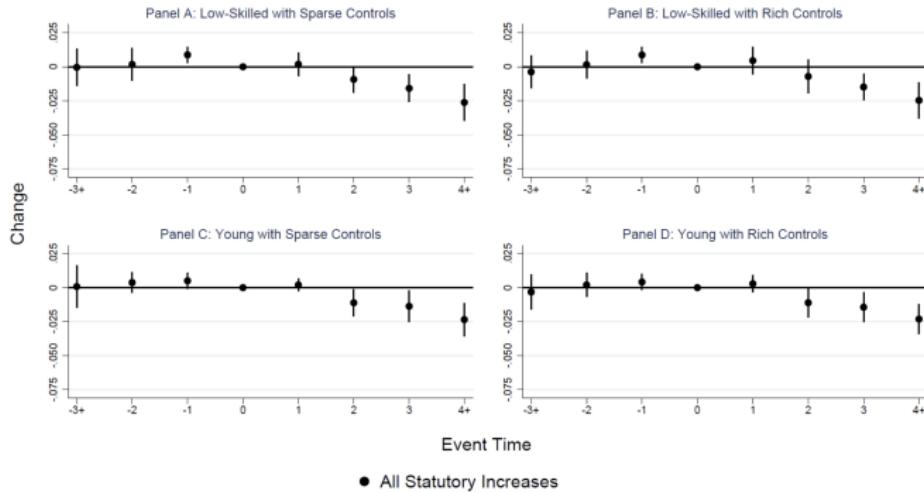
$$Y_{i,s,g(s),t} = \sum_{g(s) \neq 0} \beta_{g(s)} Policy_{g(s)} \times Post_t + \alpha_{1s} State_s + \alpha_{2t} Time_t + X_{i,s,t} \gamma + \varepsilon_{i,s,t}$$

$Y$  is binary for employment for person  $i$  in state  $s$  in policy category  $g(s)$  and time  $t$ . Samples are restricted to young (16-21yo) without high school, and young overall (16-25yo).  $X$  are the “rich controls” they label later and it includes median house price index, log aggregate personal income per capita, employment rates for different skill levels, and individual-level demographic controls.

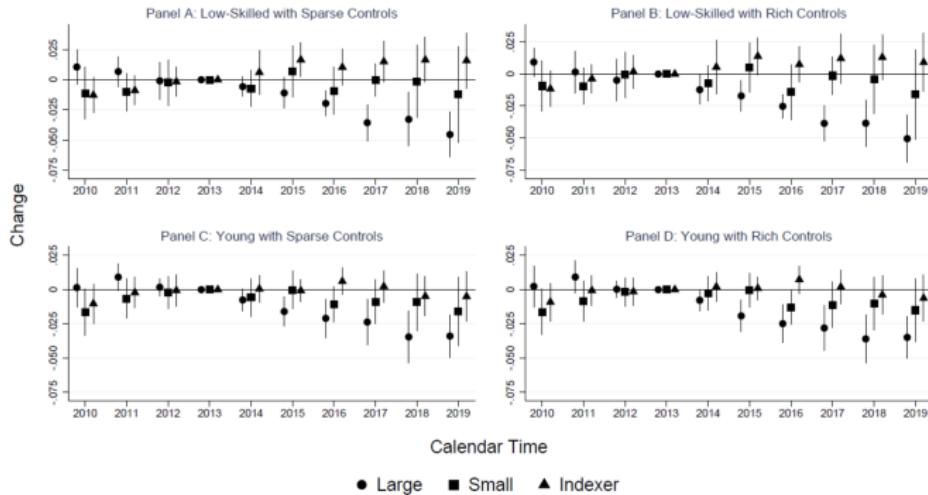
Coefficients of interest are the  $\beta$  terms and 2014 will be treated as the transition year. It measures the causal effect of state minimum wage policy changes on employment under standard, albeit nontrivial, assumptions such as homogenous treatment effects over time, no anticipation and parallel trends.

They will also estimate triple difference versions of this model with a within-state control group of untreated workers.

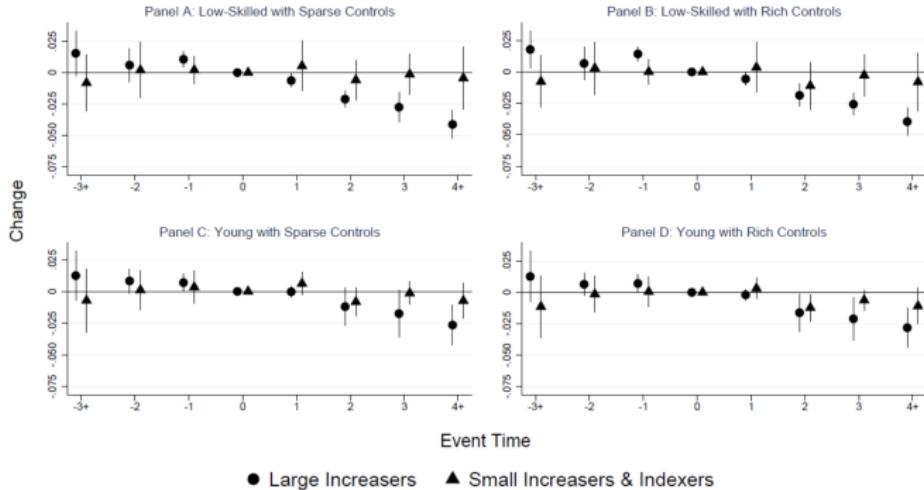
# TWFE Results



# TWFE Results



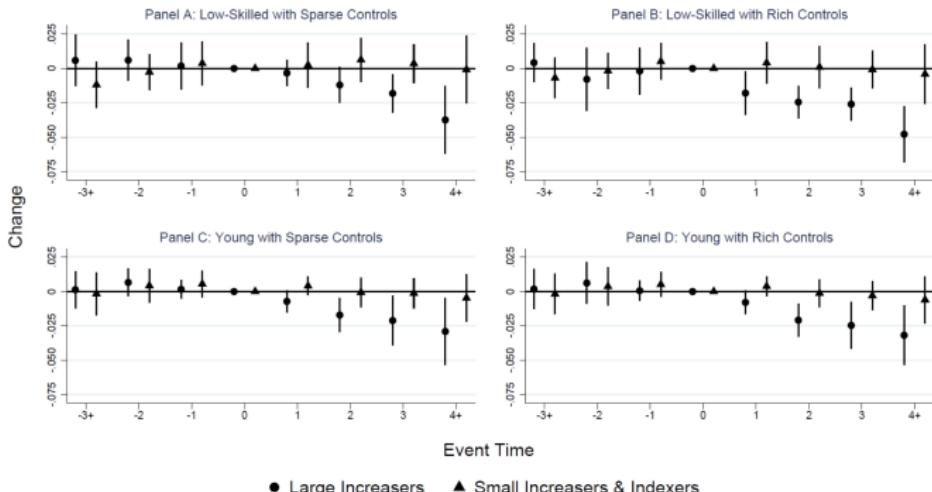
# TWFE Results



## Extension 1: Stacked regression

- Remember the big idea: create a new, much larger, dataset by appending the same dataset to itself over and over, not in balanced calendar time, but in balanced *event time*.
- Then estimate a simple TWFE model controlling for dataset-by-state fixed effects

# Stacked Regression Results



**Figure 10. Stacked Event Studies of Changes in Employment Following Large and Small Statutory Minimum Wage Increases:** This figure displays coefficients from the “stacked event study” estimator described by equation (5). Event Time is defined such that year “1” corresponds with the year during which a given state enacted its first minimum wage change due to legislation passed during our sample period. We compare estimates for large vs. small increases as defined in the main text. Panels A and B plot coefficients for low-skilled individuals defined as individuals ages 16–25 without a completed high school education. Panels C and D plot coefficients for young individuals defined as all individuals ages 16–21. The samples are from the ACS. Regressions with “sparse controls” include state and year fixed effects, as well as the log of annual *per capita* income and the annual average of the median house price index. Regressions with “rich controls” include all sparse controls plus the three-year lag of both the log of annual *per capita* income and the annual average of the median house price index, as well as a dummy variable for each education group and age. Error bars denote 95 percent confidence intervals around each estimated coefficient. Standard errors are clustered by state.

## Briefly concluding remarks

- To understand the rest of their results, we have to first cover the imputation estimator that Borusyak, Jaravel and Speiss (2021) have created
- But to summarize what we found so far, large shocks are indeed credibly causing declines in employment, particular for the affected class of workers (young with or without a high school degree)
- More modest increase are null, though, which is what Cengiz, et al. (2019) also found

## Imputation

Some methods are more obviously imputations than others though.

I will consider something explicit imputation if it constructs counterfactual observations at the unit level, as opposed to implicit imputation (e.g., manual aggregation) which tends to directly estimate ATT measures such as CS.

We will discuss three explicit imputation methods: Borusyak, Jaravel and Spiess (2021) robust imputation estimator, Gardner (2021) two stage DiD and Athey, et al. (2021) matrix completion with nuclear norm

## Background

- First “new did” paper was Borusyak and Jaravel (2017) – a lot of what was simultaneously discovered elsewhere was in that paper
- We will discuss its successor – Borusyak, Jaravel and Spiess (2021)
- My interpretation: damning critique of OLS TWFE and a robust solution based on explicit imputation

## My Outline (versus their outline)

- ① Discussion of their interpretation of “basic” DiD assumptions
- ② Critique of TWFE OLS when strong assumptions don’t hold
- ③ Introduction of new assumptions
- ④ Robust efficient imputation estimator

## Broad view

- Under three standard DiD assumptions, TWFE OLS performs fine
- No anticipation creates some challenges for event studies that requires tweaks
- But one of them (treatment effect homogeneity) introduces major problems
- Remember: theirs was the first to bring attention to what happens when treatment effect heterogeneity occurs
- After detailed critique of TWFE OLS, they roll out a robust estimator
- BLUE like characteristics

## What are we after?

A key flavor of the new DiD papers is not merely to assume TWFE OLS recovers “reasonable” weighted averages of treatment effects, but to begin by explicitly naming the target parameter. Under what assumptions can we identify  $\tau_w$ ?

Estimation target:

$$\tau_w = \sum_{it \in \Omega_1} w_{it} \tau_{it} = w_1' \tau$$

Weights need not add up to one. Weights could be  $\frac{1}{N}$  for all  $it \in \Omega_1$ . We have a number of options.

## A1: Parallel trends

**Assumption 1: Parallel trends.** There exist non-stochastic  $\alpha_i$  and  $\beta_t$  such that:

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$$

with

$$E[\varepsilon_{it}] = 0$$

for all  $it \in \Omega$ . Can be extended (e.g., unit-specific trends). Only imposes restrictions on  $Y(0)$ , not treatment effects themselves. Notice how it is a TWFE assumption – it's actually the same data generating process as in baker.do.

## A2: No anticipation

- We saw this with SA, but I think it occurred slightly earlier with BJ (not sure)
- No anticipation effects means there are no treatment effects prior to the event date

$$Y_{it} = Y_{it}(0)$$

for all  $it \in \Omega_0$ .

- I think this is probably ruling out “Ashenfelter’s dip”
- It’s also an extension of SUTVA if I’m not mistaken because SUTVA requires that your outcome is a function of your current treatment status not your future treatment status

## A2: No anticipation (continued)

- Notice how as an assumption, it literally imposes  $\tau = 0$  for all pre-treatment periods.
- They argue that “some form of this assumption is necessary for DiD identification” because otherwise you don’t have a reference period
- Even before Goodman-Bacon (2021), Sun and Abraham (2020) and Borusyak and Jaravel (2017), I had seen a million applied papers and only seen references to PT, not NA
- It’s oftentimes treated as an implicit assumption that can be then tested using an event study, but they’ll discuss that as that confuses estimation with identification

### A3: Restricted causal effects

This is the one that places restrictions on what treatment effects can and cannot be (i.e., homogenous treatment effects). Notice the very detailed expression:

**Assumption 3 (Restricted causal effects):**  $B\tau_0$  for a known  $M \times N_1$  matrix  $B$  of full row rank.

If we can assume something like homogenous treatment effects, then TWFE actually is best because its ability to *correctly* extrapolate will increase efficiency. But it's when A3 is not tenable or not really ex ante justified by theory that we should be worried. There's an A3' that is a slight modification.

## Critique of Common Practice

- ① Under-identification in event studies
- ② Negative weighting
- ③ Spurious identification of long-run causal effects

## Critique: Underidentification problem

We saw some of this earlier with SA, but mind you, there was simultaneous discoveries and a chronology. This result was in BJ, for whatever that is worth to you.

**Lemma 1:** If there are no never-treated units, the path of [pre-treatment lead population regression coefficients] is not point identified in the fully dynamic OLS specification. In particular, adding a linear trend to this path  $\{\tau_h + k(h+1)\}$  for any  $k \in R$  fits the data equally well with the fixed effects coefficients appropriately modified.

In English, it means you're going to have a multicollinearity problem even worse than you thought when estimating the fully dynamic event study model (i.e., dropping only one lead for all base comparisons)

## Underidentification of lead coefficients

### Under-identification problem

Formally the problem arises because a linear time trend  $t$  and a linear term in the cohort  $E_i$  (subsumed by the unit FE) can perfectly reproduce a linear term in relative time  $K_{it} = t - E_i$ . Therefore a complete set of treatment leads and lags, which is equivalent to the FE of relative time, is collinear with the unit and period FEs.

Just one additional normalization is needed – drop  $\tau_{-a} = 0$  and  $\tau_{-1} = 0$ . This will break the multicollinearity. We saw this in SA also. So multiple people saw this at the same time.

## Under-identification and theoretical justifications

- Imposing any  $-a$  lead and  $-1$  lead to equal zero is somewhat ad hoc. Why those two and not some other two?
- Recall with SA – it mattered which ones you dropped because otherwise leads were contaminated
- This is again about NA – if you chose  $-a$  and  $-1$ , then you had some theoretical reason to assume NA held for them and not some other periods
- Researchers need an *a priori* reason to justify which leads they drop ideally
- I had a great one – Craigslist didn't announce or advertise or communicate intentions to enter markets before they did. NA was guaranteed
- You may need to scrutinize this.

## Negative weighting and violations of A3

Heterogeneous treatment effects creating problems *again*

- It's assumption 3 – homogeneity – that BJS (and really the first paper, Borusyak and Jaravel) showed was a problem for traditional event studies
- And we saw that earlier with Sun and Abraham
- What happens is that with heterogeneity, the weights on the treatment effects can become negative

## Negative weighting

Assume some simple static model with a single dummy for treatment. Then they lay out a second lemma

**Lemma 2:** If A1 and A2 hold, then the estimand of the static OLS specification satisfies  $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{OLS} \tau_{it}$  for some weights  $w_{it}^{OLS}$  that do not depend on the outcome realizations and add up to one  $\sum_{it \in \Omega_1} = 1$ .

The static OLS estimand cannot be interpreted as a “proper” weighted average, as some weights can be negative.

## Simple illustration

Table: TWFE dynamics

| $E(y_{it})$ | $i = A$                            | $i = B$                            |
|-------------|------------------------------------|------------------------------------|
| t=1         | $\alpha_A$                         | $\alpha_B$                         |
| t=2         | $\alpha_A + \beta_2 + \delta_{A2}$ | $\alpha_B + \beta_2$               |
| t=3         | $\alpha_A + \beta_3 + \delta_{A3}$ | $\alpha_B + \beta_3 + \delta_{B3}$ |
| Event date  | $E_i = 2$                          | $E_i = 3$                          |

$$\text{Static: } \delta = \delta_{A2} + \frac{1}{2}\delta_{B3} - \frac{1}{2}\delta_{A3}.$$

Notice the negative weight on the furthest lag. This is what you get when A3 is not satisfied..

## Short-run bias of TWFE

- TWFE OLS has a severe short-run bias
- the long-run causal effect, corresponding to the early treated unit A and the late period 3, enters with a negative weight (-1/2)
- The larger the effects in the long-run, the smaller the coefficient will be
- It's caused by "forbidden comparisons" (late to early treated) – we saw this with Goodman-Bacon (2021)
- Forbidden comparisons create downward bias on long-run effects with treatment effect heterogeneity, *but not with treatment effect homogeneity* – so it really is an A3 violation

## Spurious Long-Run Causal Effects

More A3 problems, this time finding long-run effects where there are none. Basically, you need to impose a lot of pre-trend restrictions to get estimates of long-run population regression coefficients. Even then you can't get them all.

OLS estimates are fully driven by unwarranted extrapolations of treatment effects across observations and may not be trusted unless strong ex ante justifications for A3 exist

**Lemma 4:** Suppose there are no never-treated units and let  $H = \max_i E_i - \min_i E_i$ . Then for any non-negative weights  $w_{it}$  defined over the set of observations with  $K_{it} \geq \bar{H}$  (that are not identically zero), the weighted sum of causal effects  $\sum_{it: K_{it} \geq \bar{H}} w_{it} T_{it}$  is not identified by A1 and A2.

## Modifications of general model

Modification of A1 to A1':

$$Y_{it}(0) = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}$$

Assumption 4 is introduced (homoskedastic residuals). This is key, because they will be building an “efficient estimator” with BLUE like OLS properties.

Using A1' to A4, we get the “efficient estimator” which is for all linear unbiased estimates of  $\delta_W$ , the unique efficient estimator  $\widehat{\delta}_W^*$  can be obtained with 3 steps

## **Role of the untreated observations**

*"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others." – Imbens and Rubin (2015)*

*"The idea is to estimate the model of  $Y_{it}^0$  using the untreated observations and extrapolate it to impute  $Y_{it}^0$  for treated observations."*

## Steps

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

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

## Why is this working?

- Think back to that original statement of the PT assumption – you're modeling  $Y(0)_{it}$ .
- That is, without treatment – so the potential outcomes do not depend on any treatment effect
- Hence where we get treatment heterogeneity
- We obtain consistent estimates of the fixed effects which are then used to extrapolate to the counterfactual units for all  $Y(0)_{it \in \Omega_1}$
- I think this is a very cool trick personally, and as it is still OLS, it's computationally fast and flexible to unit-trends, triple diff, covariates and so forth (though remember what we said about covariates)

## Testing for parallel trends

- Perform pre-trend testing using untreated sample only
- This separation is preferable conceptually because it presents the conflation of using an identification assumption and validating it
- Traditional regression-based tests use the full sample, including the treated observations though
- Therefore it is not a test for A1 and A2; rather it is a joint test that is also sensitive to A3
- BJS test uses the untreated observations for which  $Y_{it}^0$  is ok under A2

## Test

- ① Choose an alternative model for  $Y_{it}^0$  richer than A1

$$Y_{it}^0 = A'_{it}\lambda_i + X'_{it}\beta + w'_{it}\delta + \tilde{\varepsilon}_{it}$$

- ② Estimate  $\delta$  with  $\hat{\delta}$  using OLS on untreated units only
- ③ Test  $\delta = 0$  using F-test or visually

# Comparisons to other estimators

Table 3: Efficiency and Bias of Alternative Estimators

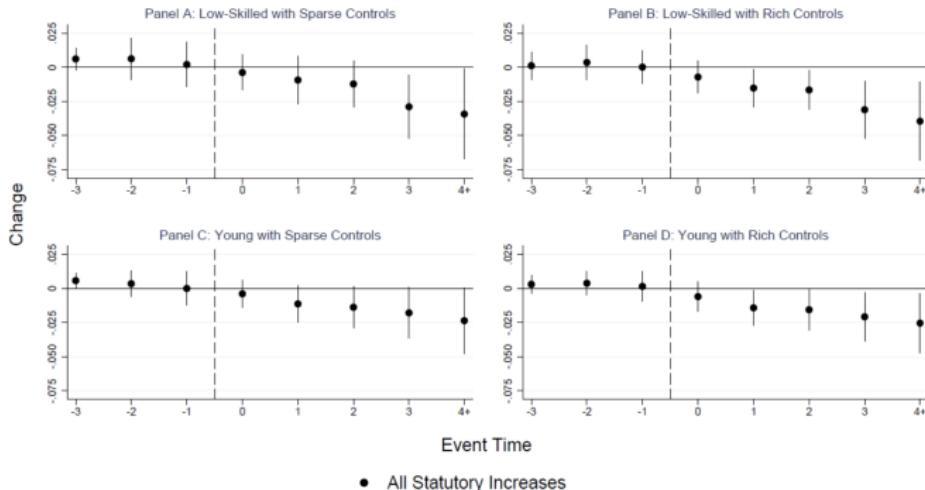
| Horizon | Estimator  | Baseline simulation |                 | More pre-periods | Heterosk. residuals | AR(1) residuals | Anticipation effects |
|---------|------------|---------------------|-----------------|------------------|---------------------|-----------------|----------------------|
|         |            | Variance<br>(1)     | Coverage<br>(2) |                  |                     |                 |                      |
| $h = 0$ | Imputation | 0.0099              | 0.942           | 0.0080           | 0.0347              | 0.0072          | -0.0569              |
|         | DCDH       | 0.0140              | 0.938           | 0.0140           | 0.0526              | 0.0070          | -0.0915              |
|         | SA         | 0.0115              | 0.938           | 0.0115           | 0.0404              | 0.0066          | -0.0753              |
| $h = 1$ | Imputation | 0.0145              | 0.936           | 0.0111           | 0.0532              | 0.0143          | -0.0719              |
|         | DCDH       | 0.0185              | 0.948           | 0.0185           | 0.0703              | 0.0151          | -0.0972              |
|         | SA         | 0.0177              | 0.948           | 0.0177           | 0.0643              | 0.0165          | -0.0812              |
| $h = 2$ | Imputation | 0.0222              | 0.956           | 0.0161           | 0.0813              | 0.0240          | -0.0886              |
|         | DCDH       | 0.0262              | 0.958           | 0.0262           | 0.0952              | 0.0257          | -0.1020              |
|         | SA         | 0.0317              | 0.950           | 0.0317           | 0.1108              | 0.0341          | -0.0850              |
| $h = 3$ | Imputation | 0.0366              | 0.928           | 0.0255           | 0.1379              | 0.0394          | -0.1101              |
|         | DCDH       | 0.0422              | 0.930           | 0.0422           | 0.1488              | 0.0446          | -0.1087              |
|         | SA         | 0.0479              | 0.952           | 0.0479           | 0.1659              | 0.0543          | -0.0932              |
| $h = 4$ | Imputation | 0.0800              | 0.942           | 0.0546           | 0.3197              | 0.0773          | -0.1487              |
|         | DCDH       | 0.0932              | 0.950           | 0.0932           | 0.3263              | 0.0903          | -0.1265              |
|         | SA         | 0.0932              | 0.954           | 0.0932           | 0.3263              | 0.0903          | -0.1265              |

Notes: See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

## Returning to the minimum wage

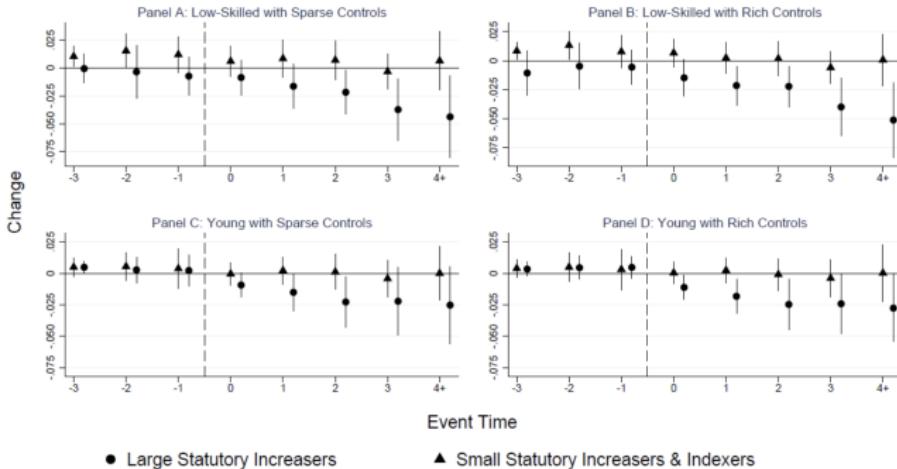
- Now we can return to the minimum wage study from earlier (Clemens and Strain 2021)
- Recall that stacked regression had found large negative effects on employment when minimum wage increases were large, but not when they were small
- The authors also implemented the BJS imputation estimator
- One comment abt the following graphics: BJS procedure does not have a “base” period in the same sense as the regression models do because it is not contrasting each period relative to some omitted group
- Rather it is imputing counterfactuals, and therefore we can calculate each period’s effect

# BJS Results



**Figure 11. Event Studies of Changes in Employment Following Statutory Minimum Wage Increases Using the BJS Imputation Estimator:** This figure displays coefficients obtained using the imputation estimator proposed by Borusyak, Jaravel and Spiess (2021) (BJS). For the BJS estimator, we code the first treatment year as the year in which a state's first statutory minimum wage increase took effect. Note that this appears graphically as "year 0" in the BJS figures, but corresponds with year 1 in the stacked event study figures. Panels A and B plot coefficients for low-skilled individuals defined as individuals ages 16–25 without a completed high school education. Panels C and D plot coefficients for young individuals defined as all individuals ages 16–21. The samples are from the ACS. Regressions with "sparse controls" include state and year fixed effects, as well as the log of annual average *per capita* income and the annual average state house price index used in our main regressions. Regressions with "rich controls" include all controls in the base controls regressions plus the three-year lag of log *per capita* income and the house price index, as well as a dummy variable for each education group and age. Error bars denote 95 percent confidence intervals around each estimated coefficient. Standard errors are clustered by state.

# BJS Results



**Figure 12. Event Studies of Changes in Employment Following Large and Small Statutory Minimum Wage Increases Using the BJS Imputation Estimator:** This figure displays coefficients obtained using the imputation estimator proposed by Borusyak, Jaravel and Spiess (2021) (BJS). For the BJS estimator, we code the first treatment year as the year in which a state's first statutory minimum wage increase took effect. Note that this appears graphically as "year 0" in the BJS figures but corresponds with year 1 in the stacked event study figures. We compare estimates for large vs. small increases as defined in the main text. Panels A and B plot coefficients for low-skilled individuals defined as individuals ages 16–25 without a completed high school education. Panels C and D plot coefficients for young individuals defined as all individuals ages 16–21. The samples are from the ACS. Regressions with "sparse controls" include state and year fixed effects, as well as the log of annual average *per capita* income and the annual average state house price index used in our main regressions. Regressions with "rich controls" include all controls in the base controls regressions plus the three-year lag of log *per capita* income and the house price index, as well as a dummy variable for each education group and age. Error bars denote 95 percent confidence intervals around each estimated coefficient. Standard errors are clustered by state.

## Comments abt the minimum wage study

- Elasticity of employment with respect to minimum wage is -0.124 and -0.082 for those without high school and the young, respectively
- Differences by size of minimum wage increase:
  - Large increases (around \$2.90): own-wage elasticity is -1.01 for 16-25yo with less than HS and -0.41 for 16 to 21yo (large effects)
  - Small increases (around \$1.90): own-wage elasticity is 0.46 (i.e., no employment effects)
  - Inflation-index increases (around \$0.90): own-wage elasticity is 0.16 (no effect) and -0.17 (no effect)

## Concluding remarks about the minimum wage study

Clemens and Strain (2021) illustrates three things:

- ① Sometimes theory may predict heterogeneous effects which requires researchers explore such theoretically motivated heterogeneity
- ② Since p-hacking is commonly associated with heterogeneity subsample analysis, we can partially protect against it through pre-registration
- ③ Robust DiD estimators should be used to double check for problems with TWFE when using DiD designs with differential timing

Reassuring that results are consistent across all models used. Do not count the minimum wage debate to be finished.

## 2SDiD

- I'd like to go back to a more traditional form of analysis by reviewing Gardner (2021)
- Like a few other papers, Gardner (2021) is both a diagnosis of the illness and a cure, and I'm putting his cure into an explicit imputation framework
- John Gardner is an assistant professor and applied econometrician at University of Mississippi – smart, cool, and former colleague of Brant Callaway of Callaway and Sant'Anna
- The cure will be nicely called two-stage difference-in-differences (2SDiD) – Nice name!

## Highlights

- Why does TWFE fail under differential timing? Violates strict exogeneity under heterogeneity
- The logic of the failure suggests an obvious, but previously unknown, solution which is the 2SDiD
- I'll explain 2SDiD, focus on the parallel trends implications, and show we can get a consistent and unbiased estimate of group and relative time fixed effects
- If you can get consistent and unbiased estimates of group and relative time fixed effects, then you can delete them and run normal analysis
- We'll work through some code

## Background

- By now, we all agree that TWFE just doesn't handle heterogeneity under differential timing very well
- We've seen in the Goodman-Bacon decomposition why – it's caused by TWFE implicitly calculating late to early 2x2s, which are a source of bias
- But some of you are coming straight from a panel econometrics course that maybe didn't use potential outcomes notation
- Isn't strict exogeneity enough for consistent estimates? What then does strict exogeneity have to do with heterogeneity and differential timing?
- Everything

## More background

*"It seems natural that TWFE should identify the ATT" – Gardner (2021)*

It just seems like TWFE with a DiD will estimate the ATT with weights that we'll find intuitive. Was this just a conjecture and was never true? Why isn't this working?

## High level discussion

- TWFE identifies the ATT when the heterogeneous effects are distributed equally across all groups and periods, but since that is a knife-edge situation, it is likely that TWFE will not in our applications meet this special scenario
- In the two group case, that is what happens though which is why TWFE worked fine there
- Metaphorically, the two group case that we always used to pin our intuition of what DiD was doing was the exception not the rule
- Goodman-Bacon (2021) shows the problem is caused by late-to-early comparisons; Gardner (2021) will show that the problem is misspecification
- Think of these as different perspectives on the same problem

## Model misspecification

*"Misspecified DiD regression models project heterogenous treatment effects onto group and period fixed effects rather than the treatment status itself"*

Spoiler: This analysis of the problem suggests solution – why don't we remove those?

## 2SDiD

“What’s the name of that kid from Mexico?” – Ted Lasso

“Dani Rojas” – Nate the Great

“Great name” – Ted Lasso

- Two stage DiD is a great name because of its connection to that classic IV model 2SLS
- If you can link it to 2SLS in your mind, it may help you because it’ll show you that Gardner’s model is a two stage model
- First stage – estimate the group and relative time fixed effects using only the  $D = 0$  observations
- Second stage – using predicted values based off those fixed effect coefficients, run your model off the transformed outcome
- Get the standard errors right just like 2SLS by taking the first stage into account

## More high level

- The second step recovers the average difference in outcomes between treated and untreated units after removing group and period fixed effects
- What I like about Gardner's method is its pleasant familiarity, its speed
- But note, it's not going to allow you to do the kind of heterogeneity analysis that CS allows for
- Some of the differences will be due to slightly different PT assumptions, and some will because 2SDID will be using all of the data for analysis, not just the baseline for calculating the DID estimates

## Notation

$i$ : panel units

$t$ : calendar time – think of real dates

$g \in \{0, 1, \dots, G\}$  – groups

$p \in \{0, 1, \dots, P\}$  – relative time or “periods”

Periods are successive. Group 0 – never treated. Group 1 – treated in period 1, 2, and on. Group 2 – treated in period 2, etc.

## Parameters

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

It's a group-time ATT but expressed in a more traditional econometric notation that you could easily find in Wooldridge or some such

## Modeling basics

Under parallel trends, mean outcomes will satisfy the following equation

$$E\left[Y_{gpit}|g, p, D_{gp}\right] = \lambda_g + \gamma_p + \beta_{gp}D_{gp}$$

In two-group, group and period effects are eliminated with dummies because TWFE uses dummies to demean across multiple dimensions. Then TWFE identifies ATT. But this does not hold when average effects vary across group and period. There are many ways to express a treatment effect's across group and time, but Gardner presented it as a weighted average of the coefficients for only that group-period situation:

$$E\left(\beta_{gp}|D_{gp} = 1\right) = E\left(Y_{gpit}^1 - Y_{gpit}^0|D_{gp} = 1\right)$$

## Strict exogeneity violation

Rewriting the above we get:

$$\begin{aligned} E\left[Y_{gpit}|g, p, D_{gp}\right] &= \lambda_g + \gamma_p + E\left[\beta_{gp}|D_{gp} = 1\right]D_{gp} \\ &\quad \left[\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)\right]D_{gp} \end{aligned}$$

The problem is there's this weird new error term and it isn't mean zero under heterogenous treatment effects spread across group and period. Unlike the two group case, the coefficient on  $D_{gp}$  from TWFE doesn't identify the average  $E(\beta_{gp}|D_{gp} = 1)$

So let's see Gardner's solution, but note – his solution was suggested by the problem itself. Gardner is thoughtful and observant.

## DiD regression estimand

- So if TWFE isn't recovering  $E(\beta_{gp}|D_{gp} = 1)$ , then what is it recovering?
- He shows that under PT, the coefficient on  $D_{gp}$  is:

$$\beta^* = \sum_{g=1}^G \sum_{p=g}^P w_{gp} \beta_{gp}$$

- So then – what are the weights  $w_{gp}$ ?
- Groan – It's a huge mess, and I hate even showing it to you because I find the weights almost impossible to decipher, but maybe you'll have a better go at it than me

## Weights

$$w_{gp} = \frac{\left\{ [1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} = 1)] \right\} P(g, p)}{\sum_{g=1}^G \sum_{p=g}^P \left\{ [1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} = 1)] \right\} P(g, p)}$$

Terms:

- $P(D_{gp} = 1|p)$ : share of units treated in period  $p$
- $P(D_{gp} = 1|g)$ : share of periods in which  $g$  is treated
- $P(D_{gp} = 1)$ : share of unit  $\times$  time treated
- $P(g, p)$ : population share of observation corresponding to group  $g$  and period  $p$

I thought about changing all those probabilities into means, but honestly, it really didn't help me at all. But Gardner notes that this is from theorem 1 of deChaisemartin and D'Haultfoeuiller (2020) and his Appendix A

## Estimation

$$Y_{gpit} = \lambda_g + \gamma_p + \beta D_{gp} + \varepsilon_{gpit}$$

This specification assumes a conditional expectation function that is linear in group, period and treatment status. But when the model is misspecified, it will attribute some of the heterogeneity impacts of the treatment to group and period fixed effects. The longer the treatment, the greater  $\bar{D}$  is, the more that group's treatment effects will be absorbed by group fixed effects. When misspecified, TWFE doesn't recover  $E[\beta|D = 1]$ .

## Statistical issues

- Common support: “as long as there are untreated and treated observations for each group and period,  $\lambda_g$  and  $\gamma_p$  are identified from the subpopulation of untreated groups and periods.”
- Identification: “the overall group  $\times$  period ATT is identified from a comparison of mean outcomes between treated and untreated groups after removing group and period effects.”

## Estimation: First stage

First stage:

$$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

using only  $D_{gp} = 0$ , retaining the fixed effects. Collect the  $\widehat{\lambda}_g$  and  $\widehat{\gamma}_p$ .

## Estimation: Second stage

Second stage:

$$\begin{aligned}\hat{y}_{gpit} &= y_{gpit} - \widehat{\lambda}_g - \widehat{\gamma}_p \\ \widehat{y}_{gpit} &= \alpha + \beta D_{gp} + \psi_{gpit}\end{aligned}$$

Why does this work? Parallel trends assumption implies:

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

But because

$$E\left\{ [\beta_{gp} - E(\beta_{gp}|D_{gp} = 1)]D_{gp}|D_{gp} \right\} = 0$$

## Estimand

Then this procedure will identify  $E(\beta_{gp}|D_{gp} = 1)$ . Consistency and unbiasedness proofs.

This is  $E(\beta_{gp}|D_{gp} = 1) = \sum^G \sum^P \beta_{gp} P(g, p|D_{gp} = 1)$ . It will tend to put more weight, by definition, on groups earlier into their treatment. But this isn't the same as the negative weighting that BJS say occurs oof the long lags. It just means there are more of them.

Event studies are:

$$y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^P \beta_r D_{rgp} + \varepsilon_{gpit}$$

Just change the second stage with the transformed outcome.

## Inference

- Standard errors are wrong on the second stage because the dependent variable uses estimates obtained from the first stage.
- The asymptotic distribution of the second stage can be obtained by interpreting the two-stage procedure as a joint GMM

|                                 |                                     |
|---------------------------------|-------------------------------------|
| Basics                          |                                     |
| Covariates                      |                                     |
| Weighted Group-Time ATT         | Imputation based robust estimator   |
| Stacking                        | 2SDiD                               |
| Imputation DiD                  | Matrix completion with nuclear norm |
| Alternative estimators          |                                     |
| Basic suggestions going forward |                                     |

## Big idea

*“The main part of the article is about the statistical problem of imputing the missing values of  $Y$ . Once these are imputed, we can estimate the causal effect of interest,  $\delta$ .”*

*“To estimate average causal effect of the treatment on the treated units, we impute the missing potential control outcomes” – Athey, et al. (2021)*

## Overview

- Athey, et al. (2021) unites two literatures – unconfoundedness and synthetic control
- Combines computer science with statistics to create the matrix completion with nuclear norm (MCNN) estimator
- Nuclear norm regularization is used for the imputation

## What is matrix completion

- Completing a matrix means guessing at the correct values that are missing
- Hence the “completion” is just another name for “filling in” the matrix
- In causal inference, if the matrix is a matrix of potential outcomes (e.g.,  $Y^0$ ), then missingness is caused by treatment assignment

Here's a matrix of potential outcomes,  $Y^0$ , representing units at time  $t$  that had not been treated.

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & Y_{it}^0 \end{pmatrix}$$

Now imagine a treatment assignment, SUTVA, that flips treatment from 0 to 1 in the last period  $t$ :

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

Ask yourself: why are there question marks in the last column?

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Matrix completion seeks to do the following:

Matrix completion with nuclear norm will impute the last column using regularized regression:

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & \widehat{Y_{1t}^0} \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & \widehat{Y_{2t}^0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & \widehat{Y_{it}^0} \end{pmatrix}$$

And once you have those, you can calculate individual level treatment effects that can be used to aggregate to the ATT

## History of matrix completion

- Open competition by Netflix in 2006 – winner would get \$1m if they could improve predictive model by ten points on RMSE
- Invited a ton of competition – from MIT teams to regular everyday joes working out of their home office
- Everyone was given a database which was then tested by Netflix on a holdout dataset
- Quick progress was made followed by very slow advances
- Winner was announced in 2009

## Netflix prize

- Gigantic sparsely populated matrix (100m users ranking 100k movies)
- I like Silver Linings Playbook and Lars and the Real Girl and you like Silver Linings Playbook
- Probably you'll also like Lars and the Real Girl
- So we are using correlations in the columns to "complete" missing values
- When you think about it, while it seems predictive (and it is), isn't it really a causal design?
- "If I watch Lars and the Real Girl, will I like it?"

## Types of imputation

- I didn't always think of causal inference in terms of imputation because often the method was just taking existing values and manipulating them, rather than filling in missing values
- But the fundamental problem of causal inference states that causal inference is a missing data problem, so it makes sense you'd be imputing
- I tend to think therefore in terms of implicit and explicit imputation methods
- Borusyak, et al. (2021) and Athey, et al. (2021) both seem more like "explicit" imputation methods
- Callaway and Sant'Anna (2020) on the other hand is an implicit method, as is did methods more generally

## Two literatures

- Lots of moving parts in this interesting paper, so my goal here is purely explainer and mostly high level at that.
- I want you to be competent and conversant in it so we also have some R code
- There's two literatures they want you to have in your mind:
  - ① Unconfoundedness –  $(Y^0, Y^1) \perp\!\!\!\perp D|X$  – sometimes explicitly imputes (nearest neighbor), sometimes more implicit (inverse probability weighting)
  - ② Synthetic control – literally calculating a counterfactual as a weighted average over all donor pool units
- Their MCNN method will show that both are “nested” within the general framework they've developed making them actually special cases

## Differences

- Conceptually different in the way they exploit patterns for causal inference
- Unconfoundedness assumes that patterns over time are stable *across units*
- Synth assumes patterns across units are stable *over time*
- Regularization (particularly the nuclear norm) nests them both

## The Gist

- Factor models and interactive effects model the observed outcome as the sum of a linear function of covariates and a unobserved component that is a low rank matrix plus noise
- Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components with the rank itself estimated

## Three contributions

- ① Formal results for non-random missingness when block structure allows for correlation over time. Nuclear norm is important here
- ② Shows unconfoundedness and synth are in fact matrix completion methods
  - they all have the same objective function based on the Frobenius norm for the difference between the latent matrix and the observed matrix
  - Each approach imposes different sets of restrictions on the factors in the matrix factorization
  - MCNN by contrast doesn't impose any restrictions – just regularization to characterize the estimator
- ③ Applies the method to two datasets, but I'm going to skip that bc I find that stuff tedious once I've muscled my way through a paper like this

## Block structure

- Lots of jargon in this article – unconfoundedness, vertical and horizontal regression, fat and thin matrices.
- Unfortunately, you need to learn it all so let me try and organize it
- We define the matrix first in terms of its block structure which is describing where and when the missingness is occurring in the matrix

## Unconfoundedness

- Much of the unconfoundedness literature estimates an ATE under unconfoundedness
- But it tends to focus only on a simple setup where the missingness is the last period
- Think about LaLonde (1986) – NSW treats the workers, and then you don't observe  $Y^0$  for the treated group in the *last period*
- This is the “single-treated-period block structure” because only one *period* is missing

## Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

## Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

Notice, this is the synthetic control design because a single unit (unit  $i$ ) is missing  $Y^0$  for the 3rd and  $t$ th periods.

## Staggered adoption

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & ? & ? & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

So all of these so-called designs can be expressed in terms of missingness in the block structure, and our job therefore is to find an estimator that is general enough to manage all of them. Their MCNN will be that.

## Thin and Fat matrices

- We also have to consider the relative number of panel units  $N$  and time periods  $T$  because this also shapes which regression style will be used for imputation
- Thin matrices are basically where  $N \gg T$ , but fat matrices are ones where  $T \gg N$
- Approximately square ones are where  $T$  is approximately equal to  $N$
- My conditionally accepted JHR (fingers crossed) had around 400 panel units and 180 months, so it was  $N \gg T$  which is a thin matrix

## Vertical and horizontal regression

- Two special combinations of missing data patterns and matrix shape need special attention because they are the focus of large but separate literatures
- Unconfoundedness has that single-treated period block structure with a thin matrix ( $N \gg T$ ).
- You use a large number of units and impute missing potential outcomes in the last period using controls with similar lagged outcomes
- This is the horizontal regression – imagine just running OLS on the lags and taking predicted values
- The horizontal regression holds under unconfoundedness

## **Vertical regression**

Doudchenko and Imbens (2016) and Pinto and Furman (2019) show that Abadie, Diamond and Hainmueller (2011) can be interpreted as regressing the outcomes for the treated prior to treatment on the outcomes for controls in the same period

## Fixed effects and factor models

- Both horizontal and vertical regressions exploit other patterns
- An alternative to each of them though is to consider an approach that allows for the exploitation of both stable patterns over time and stable patterns across units
- This is where their matrix completion with nearest neighbor model comes in – it does that very thing

## Matrix completion with nuclear norm

- Model the  $N \times T$  matrix of complete outcomes data matrix  $Y$  as:

$$Y = L^* + e$$

where  $E[e|L^*] = 0$

- The error term can be thought of as measurement error if you need a frame to think about it
- So you have this complete matrix,  $L^*$ , and zero mean conditional independence holds

## Assumption 1

Apart from the unconfoundedness assumption, we have this weird assumption!

### Assumption 1

$e$  is independent of  $L^*$  and the elements of  $e$  are  $\sigma$ -sub-Gaussian and independent of each other

Lots of matrix forms can be defined this way. But let's not get lost in the weeds – we are still just trying to estimate  $L^*$ ! That's the main storyline, not the side quest, to use Red Dead Redemption words I understand

## All imputations are wrong but some are useful

- You can impute something a million different ways.
- $1 + 1 + 1 + 1 = 4$  is an imputation of the fifth unknown element and frankly just looking at it, seems wrong.
- You could minimize the sum of squared differences but if the objective function doesn't depend on  $L^*$ , the estimator would just spit back  $Y$  and  $\delta = 0$ .
- They add a penalty term  $||\lambda||$  to the objective function, but even then, not all of them do well.
- Turns out, it actually matters whether you regularize the fixed effects or not (just like it matters whether you regularize the constant in LASSO apparently – I decided to take their word for it)

## Estimator

$$L^* = \widehat{L} + \widehat{\Gamma} \mathbf{1}_T^T + I_N \widehat{\Delta}^T$$

where the objective function is:

$$= \arg \min_{L, \Gamma, \Delta} \left\{ \frac{1}{O} \| P_0(Y - L - \Gamma \mathbf{1}_T^T - \mathbf{1}_N \Delta^T) \|_F^2 + \Lambda \| L \| \right\}$$

## Fixed effects and regularization

- The penalty will likely be the nuclear norm but notice that the fixed effects are outside the penalty term. You could subsume them into  $L$ , they say, but they recommend you not doing this.
- Fraction of observations is relatively high and so the fixed effects can actually be estimated separately (apparently that is one difference between MCNN and the rest of the MC literature)
- The penalty will be chosen using cross-validation

## Other norms

- One thing I thought was interesting was that the nuclear norm allowed for the construction of a low rank  $L^*$  matrix, but other norms actually would have weird properties
- I remember once me asking Imbens (like I had even a clue what I was talking about), "Why not use elastic net? Why are you using the nuclear norm?" He said elastic net would spit out all zeroes. I remember thinking "Why did I think I would understand what he told me?"
- One advantage of NN is its fast and convex optimization programs will do it, whereas some others won't because of the large  $N$  or  $T$  issues
- There's almost like a cross walk, too, between this and Borusyak, et al. (2021) but I don't quite see it except they both leverage imputation

## Conclusion

- Let's just review the R code. It uses gsynth
- We'll look at Cheng and Hoekstra (2013) again, because frankly it's a dataset I know
- Ultimately, this is just another model though that can be used for differential timing but at the moment, no one knows how it performs in simulations alongside Borusyak, et al. (2021), Callaway and Sant'Anna (2020) or any of the others
- So I can't really answer questions about when to use it and not to – it comes down to these very narrow assumptions
- You choose the estimator based on the problem you're studying and the assumptions – you must justify it, no one else can, but you do so by appealing to assumptions

## Sharp DiD

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

## Fuzziness

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

## Fuzzy estimators

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

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

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

## Two proposed estimators

Propose two other estimators

- ① Time corrected Wald ratio,  $Wald_{TC}$  – relies on PT within subgroup of units sharing the same treatment at the first date
- ② Changes in changes extension,  $Wald_{CiC}$  – extension of Athey and Imbens (2006) “changes in changes” paper. Generalizes CiC to fuzzy. CiC is invariant to outcome scaling but puts restrictions on the full distribution of potential outcomes instead of the mean

## Personal takeaway

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

## Most basic notation

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

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

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

Sharp:  $D = G \times T$ ; Fuzzy:  $D \neq G \times T$

## Cases under consideration

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

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

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

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

## Cases under consideration

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

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

$\text{Wald}_{DiD}$  identifies the LATE of switchers under PT and stable ATE assumption and LATE of treatment and control group switchers are the same

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

## Fuzzy design assumptions

A1: Dominating growth of treated units in the treatment group

The treatment group is the one experiencing the larger increase in its treatment rate.

This rules out the case where the two groups experience the same evolution of their treatment rates. Let  $R_{gt} \sim R|G = g, T = t$ ; Assumption 1 implies the following conditions:

$$E(D_{11}) > E(D_{10})$$

$$E(D_{11}) - E(D_{10}) > E(D_{01}) - E(D_{00})$$

## Fuzzy design assumptions

A2: Stable percent of treated units in the control group

$0 < E(D_{01}) = E(D_{00}) < 1$  means there is stable percent of treatment units in the control group.

This is a special case where number of treatment units in control group is fixed.

## Fuzzy design assumptions

### A3: Treatment participation equation

In the treatment group, no one switches from treatment to control.  
Formally this is

$$D = 1 \text{ if } V \geq v_{gt} \text{ with every } V \perp\!\!\!\perp T|G$$

Where  $V$  is the propensity to get treatment,  $v_{gt}$  is a threshold specific to each group/time

## A little more notation

- We say a unit is treated as  $D(t) = 1\{V \geq v_{gt}\}$
- Switchers are units who go from control to treatment between 0 and 1  $S = \{D(0) < D(1), G = 1\}$
- LATE is for switchers:  $\Delta = E(Y_{11}(1) - Y_{11}(0)|S)$
- LQTE is also for switchers:  $\tau_q = f_{y_{11}(1)|S(q)}^{-1} - F_{y_{11}(0)|S(q)}^{-1}$

## Switcher LATE/LQTE

Why only switchers?

- Sometimes only ones affected are switchers; a policy occurs but only eligibility for some. Switchers end up treated
- Identifying more than the LATE places more restrictions and this already has like 8 assumptions

## First estimator: $\text{Wald}_{DiD}$

Commonly used strategy in these fuzzy designs is to normalize the DiD on the outcome by the DiD on the treatment status itself (because remember, in the fuzzy design, units are *becoming* treated as well as *being in treatment groups*)

$$\text{Wald}_{DiD} = \frac{\text{DiD}_Y}{\text{DiD}_D}$$

## Wald-DiD

Let  $S' = \{D(0) \neq D(1), G = 0\}$  be control group switchers. Then we define relevant parameters as:

$$\begin{aligned}\Delta' &= E(Y_{01}(1) - Y_{01}(0)|S') \\ \alpha &= \frac{[P(D_{11} = 1) - P(D_{10} = 1)]}{DiD_D}\end{aligned}$$

## Assumptions

### A4: Parallel trends

Standard assumption. Not worth repeating for the millionth time.

## Assumptions

### A5: Stable treatment effect over time

In both groups, the average effect of going from 0 to  $d$  units of treatment among units with  $D(0) = d$  is stable over time. This is the same as assuming that among these units, the mean of  $Y(d)$  and  $Y(0)$  follow the same evolution over time

$$E \left[ Y(d) - Y(0) | G, T = 1, D(0) = d \right] - \\ E \left[ Y(d) - Y(0) | G, T = 0, D(0) = d \right] = 0$$

for units in the switching population

## Assumptions

A6: Homogenous treatment effect over time

Switchers have the same LATE in both groups. This isn't necessary in sharp DiD, just fuzzy

## Wald DiD theorems

There's a reason we just listed six assumptions. We need them for this traditional scaled DiD method for fuzzy designs called the Wald DiD. We'll go in order.

### Theorem 1: Wald DiD

If A1, A3-A5 hold, then Wald DiD equals

$$\alpha\Delta + (1 - \alpha)\Delta'$$

but if A2 or A6, then Wald DiD equals  $\Delta$

## Interpretation of theorem 1: case 1

Case 1: when treatment grows in the control group, then  $\alpha > 1$ .  
Then if we assume A1, A3-A5, a lot of things cancel out under A1, A3-A5, but the Wald DiD becomes a weighted *difference* of the LATEs of treatment and control group switchers in period 1.

Since it is a difference in LATEs, then even two positive LATEs can flip sign if the first is less than the second.

But if you assume A6, you just get the LATE.

## Interpreting theorem 1: case 2

Case 2: When treatment diminishes in controls, then  $\alpha < 1$ .

Then under A1, A3-A5, Wald DiD will equal a weighted average of LATEs of treatment and control group switchers in period 1.

This quantity will not reverse signs, but won't equal the LATE without A6.

### Interpreting theorem 1: case 3

Case 3: Treatment rate is stable in control, then  $\alpha = 1$  and Wald DiD will equal LATE under A1, A3-A5.

This requires that the ATE among units treated at  $T=0$  remain stable over time – necessary condition.

Under A1, A3-A4, Wald DiD is equal to LATE plus a bias term involving several LATEs, and unless they cancel out exactly, Wald DiD will be different from the LATE

## Alternative estimators

- Wald TC – Time Corrected Wald DiD
- Wald CiC – Changes in changes generalization to fuzzy design

Now we review alternative assumptions under which Wald TC or Wald CiC identify the LATE of switchers in the fuzzy. First let's look at Wald TC which won't depend on A4-A5.

## Alternative assumptions for the Wald TC

### A4': Conditional parallel trends

This requires  $Y(0)$  mean average follow the same trends as all the other groups.

## Wald TC estimator

Wald TC equals

$$\frac{E(Y_{11}) - E(Y_{10} + \delta_{D_{10}})}{E(D_{11} - E(D_{10})}$$

where

$$\delta_d = E[Y_{d_01}] - E[Y_{d_00}]$$

which is the change in mean outcome between periods 0 and 1 for controls and treatment status  $d$  (not groups T and C – individual units  $d$ ).

## Theorem 2

Theorem 2 and the Wald TC

If A1-A3 and A4', then Wald TC equals  $\Delta$

Note that: Wald TC equals

$$\frac{E(Y|G=1, T=1) - E(Y + (1-D)\delta_0 + D\delta_1|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

This is almost the Wald DiD ratio except for that second term with the  $Y + (1 - D)\delta_0 + D\delta_1$  instead of just  $Y$ .

This arises because time can independently affect the outcome.

When treatment is stable for a group  $G$ , then  $\delta_0 = 0$ .

## Comment on Theorem 2

Wald TC equals

$$\frac{E(Y|G=1, T=1) - E(Y + (1-D)\delta_0 + D\delta_1|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

The numerator of Wald TC compares the mean outcome in the treatment group in the post period 1 to the counterfactual mean we would have had if switchers had remained untreated.

Then normalized by the change in switching, we get the LATE for switchers

## **Wald CiC**

Here we have continuous outcomes and an estimator for quantiles of the LATE called LQTE. New assumption is complicated but is needed for the Wald CiC

## Assumptions for changes in changes Wald ratio

### A7: Monotonicity and time invariance of unobservables

Potential outcomes are strictly increasing functions of some scalar unobserved heterogeneity term whose distribution is stationary over time. Also imposes the distribution of that unobserved heterogeneity be stationary within subgroups of units sharing the same treatment status at baseline.

## Data restrictions

### A8: Data restrictions

First,  $Y$  must have the same support in each of the eight  $D \times G \times T$  cells (common support). Second, the distribution of  $Y$  be continuous with positive density in each of the eight cells.

This will allow us to bound treatment effects (Athey and Imbens 2006). Now the ugliest estimator ever.

## Wald CiC estimator

Let  $Q(y) = F_{Y_{01}}^{-1} \cdot F_{Y_{00}}(Y)$  be the quantile-quantile transform of  $Y$  from period 0 to 1 in the control group. Also let:

$$F_{CiC,d(Y)} = \frac{P(D_{11} = d)F_{Y_{d11}} - P(D_{10} = d)F_{Y_{d10}}}{P(D_{11} = d) - P(D_{10} = d)}$$

And our Wald CiC estimator is:

$$W_{CiC} = \frac{E(Y_{11}) - E(Q_{D10}(Y_{10}))}{E(D_{11}) - E(D_{10})}$$

### Theorem 3: Wald CiC

Theorem 3: Wald CiC

Under A1-A3 and A7-A8, then  $W_{CiC}$  is the LATE and equivalently we get the LQTE

$$W_{CiC} = \frac{E(Y|G=1, T=1) - E((1-D)Q_0(Y) + DQ_1(Y)|G=1, T=0)}{E(D|G=1, T=1) - E(D|G=1, T=0)}$$

### **Comment on theorem 3**

Almost the standard Wald DiD except for that  $(1 - D)Q_0(Y) + DQ_1(Y)$  instead of  $Y$  in the second term of the numerator. So again, we are simply making adjustments for the fuzziness but under different set of assumptions. This term accounts for the fact that time directly affects the outcome, but in a CiC setup.

## **Which to use**

It's about choosing your poison. Do you want A4' or A7?

When T and C have different outcome distributions conditional on D in the first period, then scaling of the outcome may have large effect on the Wald-TC. Whereas Wald-CiC isn't sensitive to the scaling of Y.

But when the two groups have similar outcome distributions conditional on D in the first period, Wald-TC may be preferable as A4' only restricts the mean of the potential outcomes, whereas Wald-CiC restricts the entire distribution

## Extensions to non-binary, ordered treatment

### Theorem 6

Under continuous treatments, the estimators we've been considering are equal to the average causal response parameter that Angrist and Imbens (1995) discuss. This parameter is a weighted average over all values of  $d$  of the effect of increasing treatment from  $d - 1$  to  $d$  for any switchers where treatment status goes from strictly below to strictly above  $d$  over time.

Theorem 6 extends to a continuous treatment. Under theorem 6, each of the estimators is identifying a weighted average of the derivative of potential outcomes with respect to changing  $d$

## Stata code

Only code I know of at the moment is the fuzzydid the authors published in Stata Journal. But it allows you to specify which estimator. Here's sample code for Wald DiD:

```
fuzzydid lngonf g_decr post1 inverse_fee, did breps  
(1000) cluster(county1)
```

where *g<sub>decr</sub>* is the treatment group dummy, *post1* is the post period dummy, and *inversefee* is our continuous treatment variable. We specify the Wald DiD by noting *did* after the comma.

## Concluding remarks

- Paper is hard but worth it. It's possible your controls are getting treated for unrelated reasons, but this is testable
- The Wald DiD is a conventional approach but suffers bias without a layering in of assumptions
- Alternative estimators for when control group stabilization isn't possible or you don't want to impose treatment effect homogeneity are available
- `fuzzydid` can handle continuous treatments as well as dummies.

← Thread



Analisa Packham  
@analisapackham

...

referees who keep suggesting Calloway and Sant'Anna (2019) and Goodman-Bacon (2019) when the treatment happens to everyone in the same year:

plz stop it

10:58 AM · Jul 15, 2021 · Twitter Web App

23 Retweets 16 Quote Tweets 540 Likes



- Differential timing with heterogeneity – Bacon, Callaway and Sant'anna, etc.
- Covariates – Abadie, Sant'Anna and Zhao
- Fuzzy - de Chaisemartin and D'Haultfoeuille

## Concluding remarks on DD

- You're probably going to write a paper using DiD at least once in your life, but probably more
- Even if you don't, you're going to read a lot of papers using DiD, referee them, or advise students using them
- It's in your best interest to make the fixed cost investment in the new econometrics of DiD because the old methods are mostly harmful
- Good news is we are at the conclusion of this wave of papers, software is now widely available, solutions tend to have common features, and overall presentations (static and dynamic) aren't all that different

## Concluding remarks

- Simple 2x2 has its own problems when estimated using TWFE  
*if you include covariates*
- Stronger assumptions needed to include covariates, and bias can be large
- Don't control for covariates that could be affected by the outcome (e.g., COLLIDER BIAS!! DAG!! BOOGIEMAN!)
- Why pay more for the same car?

## Concluding remarks

- Main problem in differential timing is heterogeneity and the use of already-treated units as controls
- Honestly, I'll just put my neck out there – if you have any reason to believe homogenous treatment effects hold from theory, fine. Use TWFE
- But with differential timing and not a priori theory, you *cannot* use TWFE. It is biased, and it does not obey a “no sign flip” property, weights can be negative, etc etc.
- CS has additional benefits like examining heterogeneous responses by timing – this is part of the value of defining target parameters as weighted averages

## Concluding remarks

- Causal claims depends on valid assumptions, high quality and appropriate data, and appropriate estimators
- Use this opportunity to remember how much fun econometrics is
- Don't sweat whether you learned everything in this seminar – check out my substack “Causal Inference: the Remix” for simple explainers, go back to the papers, talk to the authors (they are all very smart, but also extremely kind people)
- Have fun! Remember that applied work is exciting, so don't sweat it. Don't forget how great it is to learn something new
- Don't forget that season 2 of Ted Lasso came out yesterday. It felt the same, but different (metaphor)