New Literature in Difference-in-Differences

A guide for practitioners

Erich Denk & Tim Simcoe

TPRI

11/10/2021

Motivation

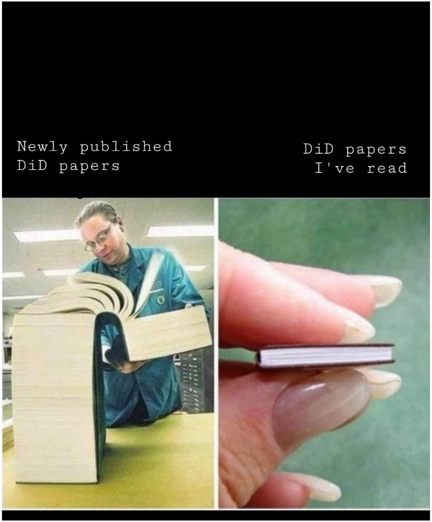


Figure 1: How many of us feel (Credit: Khoa Vu)

Outline

- Describe newly raised **potential** issues with standard DiD estimates
- Focus on 3 particularly relevant papers (there are many more!) and the problems of differential treatment timing
    - Goodman-Bacon (2019)
    - Callaway and Sant'Anna (2020)
    - Borusyak, Jaravel, and Speiss (2021)
- Discuss solutions and note software packages (if any) presented by these papers
- *Very brief* overview of a few other papers
- **Disclaimer: The canoncial 2x2 (two groups, two time periods) is perfectly fine!**

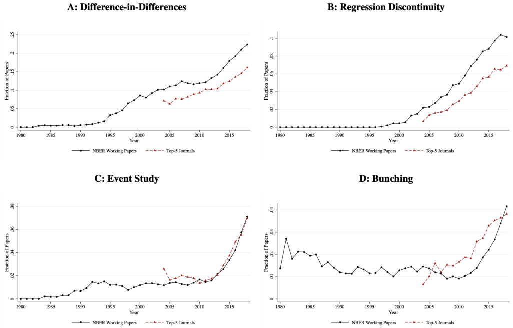
What is DiD?

- One of the most common research designs for evaluating non-random treatment applied to multiple groups
- Some units may never receive treatment or have not yet received treatment
- Observe each group before and after the treatment
- Examine the difference between the groups before, the differences after, and difference the differences
- New wave of literature raises concerns about canonical DiD regressions

Advantages of DiD

- Conceptually intuitive. John Snow and Cholera, Card & Krueger and the Minimum Wage
- Helpful in answering big and important questions. State policies, changes in corporate laws, technology diffusion etc.

Figure IV: Quasi-Experimental Methods



A: Difference-in-Differences

B: Regression Discontinuity

C: Event Study

D: Bunching

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.

Figure 2: Currie et al (2020) show 25% of NBER working papers use DiD

Brief Review of TWFE

In the 2x2 Case:

$$y_i = \alpha(POST) + \alpha(TREAT) + \beta^{DD}(POST * TREAT)$$

But with multiple time periods (and often multiple treatment periods):

$$y_{it} = \alpha_t + \alpha_i + \beta^{DD}(D_{it}) + \mu_{it}$$

- Two-Way Fixed Effects in panel data, is the most common DiD estimation
- Often, we just add covariates and assume *conditional* parallel trends

$$y_{it} = \alpha_t + \alpha_i + \beta^{DD}(D_{it}) + \theta \cdot X_{it} + \mu_{it}$$

Brief Review of TWFE (cont.)

- Identifying assumptions for TWFE
  - regressors are strictly exogenous conditional on the unobserved effect
  - Allows x_{it} to be arbitrarily related to unobserved covariates
  - Regressors will vary over time for at least some i and are not collinear
- Inference: Cluster standard errors by panel unit
  - Allows correlation in the μ_{it}'s for a given i
  - Need a reasonably large number of clusters

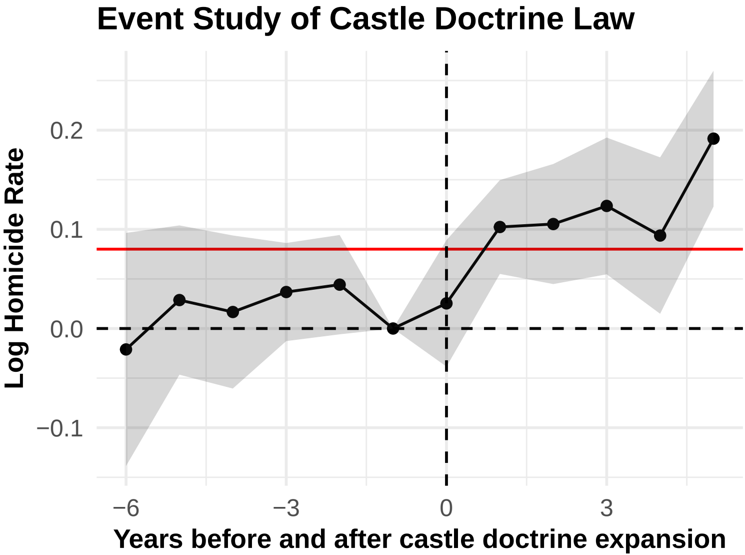
TWFE Conclusion

- Good reasons for TWFE specification
  - ATT under parallel trends
  - SEs easily obtained
  - Include multiple periods
  - Allow different treatment intensities
- But problems stemming from:
  - Varying time treatment/differential timing
  - Inclusion of covariates

Working example for presentation

- Cheng and Hoekstra (2012) on "Castle Doctrine"
- Research Q: *What is the effect of the passage of self-defense laws on homicides and violent crime?*
- Motivation: "Stand your ground" laws may make the expected cost of crime higher
- Findings: Opposite, an 8pp *increase* in the number of murders and non-negligent man-slaughters after passage of such laws

Castle Doctrine Event Study



Section 1

The Bacon Decomposition

Goodman-Bacon Key Facts

- Analyzes the TWFE estimator *if there is variation in when* treatment turns on.
- In essence, shows what β^{DD} is *algebraically*
- Some key facts
  - The DD estimator = weighted average of all the 2x2 DDs.
  - Every unit is necessarily part of the control group in *some* 2x2s
  - Weights come from the relative *size of the subgroup*
  - Estimates can change across specifications because the weights change, the 2x2 DD terms change, or both
  - Controls can introduce new and unintended identifying variation

All possible 2x2s in the three group case

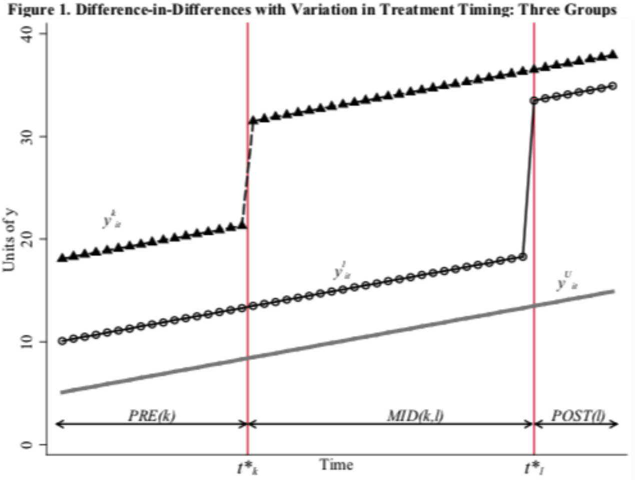
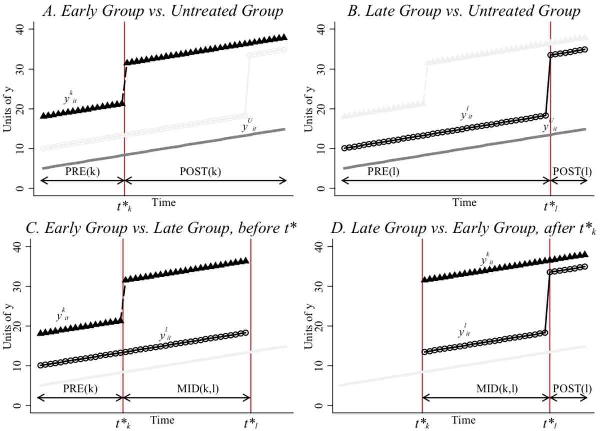


Figure 3: Goodman-Bacon Figure 1

The forbidden comparison

Panel D. Our late cohort is the treated and early treated cohort is control, but only after they are already treated.

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



A. Early Group vs. Untreated Group

B. Late Group vs. Untreated Group

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

*D. Late Group vs. Early Group, after t*ₖ*

The Weights

- The weights of the “Weighted 2x2” are determined by:
  1. Sample size (what share of units are in each treatment wave) - bigger groups get more weight
  2. Subgroup variance of treatment. OLS prefers groups where the Fixed Effect-adjusted treatment dummy varies more.
     - Bigger weights when treated and control times are equally sized
     - Early units have long post periods, short post periods. Opposite for late. Again, impacts variance
- These weights are just how OLS handles things all the time!

Goodman-Bacon Implications

- Differential trends in counter-factual outcomes *in a given timing group* can generate some bias proportional to the weight of the group
- Groups treated in the middle of the panel matter most
- Weights can be calculated and used in a balance test

Goodman-Bacon Implications (cont.)

- If treatment effects change monotonically over time or are “gradual” (not stepwise)
    - DD estimate is biased away from sign of true effect
    - This comparison group is “contaminated” and will cause us to underestimate the size of our treatment effect

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time

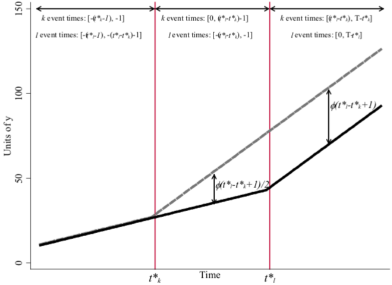


Figure 5: Goodman-Bacon Figure 3

Goodman-Bacon Implications (cont.)

- If treatment effects are constant
  - DD estimate is variance weighted ATE ($\neq$ ATT)
  - Again, middle panel groups get more weight than their sample size implies
- Using Callaway and Sant'Anna, Sun and Abraham, or a stacked setup will address much of the concern
- **In general, many of the other papers that are out there are focused getting rid of the "bad stuff" he's shown here**

Goodman-Bacon Statistical Packages

- Stata and R `bacondecomp`
  - calculates the decomposition
  - Scatter plot of 2x2s against their weights
  - Stores weights for future calculations
  - Good tool for seeing where the action is. Driven by one group, one type of comparison?
- Revisiting Cheng and Hoekstra...

	treated	untreated	estimate	weight	type
3	2006	2007	0.0420034	0.0045102	Earlier vs Later Treated
4	2008	2007	0.0437967	0.0090205	Later vs Earlier Treated
5	2010	2007	0.0065480	0.0022551	Later vs Earlier Treated
6	2009	2007	0.1495475	0.0060136	Later vs Earlier Treated
7	2007	99999	0.0592543	0.6103851	Treated vs Untreated
9	2006	99999	0.1450326	0.0503065	Treated vs Untreated

Equivalence of weighted sum and TWFE

If we multiply those estimates by the weights...

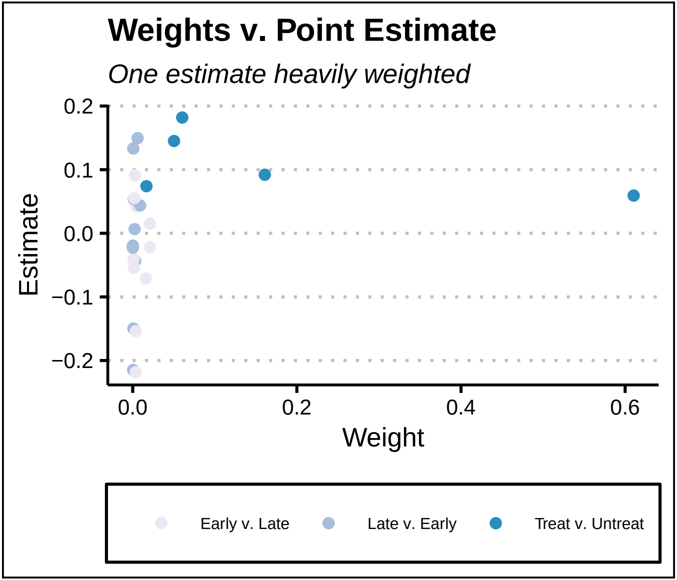
```
## [1] "Weighted sum of decomposition = 0.0694"
```

We get the same thing as estimating the TWFE

$$log(homicides) = \beta^{DD} + \alpha_i + \gamma_t$$

```
## [1] "Two-way FE estimate = 0.0694"
```

Which 2x2s are important?



Section 2

Ad Hoc Solutions: Callaway & Sant'Anna (2020), Borusyak, Jaravel, and Speiss (2021), and "Stacked DiD"

“Difference-in-Differences with Multiple Time Periods” Callaway and Sant’Anna (2020)

Considers the issues raised by Goodman-Bacon and attempts to find an ATT with setups of:

1. Multiple time periods ($T > 2$)
2. Variation in treatment timing
3. Parallel trends are conditional on observables / Availability of covariates

- Related to Abadie (2005)

Apply Callaway and Sant'Anna when...

1. Treatment effects are heterogeneous by time of adoption
2. Treatment effects change over time
3. Short-run effects are more pronounced (comparison groups shrink, so later estimates are less precise)

- Doesn't (yet) account for intensity of treatments (or multiple treatments, or switching of treatment status)

Advantages of Callaway and Sant’Anna

- *Pre-treatment covariates* can be included
- Aggregation schemes to summarize treatment effects
- Minimal parallel trend assumptions to identify the *ATT(g,t)*

How it works

- Sub-setting the data over and over to map back to the well-understood 2x2 case. Similar to Sun and Abraham (2020) and Chaisemartin and D'Haultfœuille (2020)

1. Identification of dis-aggregated parameters
2. Aggregation of these parameters
3. Estimation

- Each treatment cohort, g, will have its own ATT (except for the last treatment group in some cases)
- Choose parallel trends assumption: Which comparison group is appropriate for your application. Are never-treated very different from not-yet treated?

1. Use "never-treated" group for conditional parallel trends
2. Use "not-yet-treated" group for conditional parallel trends (usually larger)

Covariates

- Doubly robust:
    - As long as the model of the propensity score *or* model for outcome evolution are correctly specified, we will recover ATT. Two opportunities (like Sant'Anna and Zhao (2020)) to get the ATT right
- Careful using time-varying controls that might be impacted by the treatment

Aggregating Group-Time ATT

- Subsetting so much, we might lose precision
- Paper walks through the choices of weights, but we want to avoid the overweighting of units that are treated earlier

Options for the weights:

1. Event-Study/Dynamic treatment effect
2. Cohort heterogeneity (pre-time period measure)
3. Weight by calendar time

Assumptions

- Sampling is panel data
- Conditional Parallel Trends
- Irreversible Treatment
- Common support (via propensity score) a la Abadie (2005)
- Limited treatment anticipation (ATT is 0 pre-treatment)

Implications

- In essence, best to estimate narrow ATT per group-time. Can be useful if:
    1. Parallel trends only holds conditional on covariates
    2. Different comparison groups (never or not-yet treated)
    3. Units can anticipate treatment participation
- We might be (are very likely) interested in an aggregate estimate
    - IPW
    - Outcome regression
    - Doubly Robust
- Will arrive at the same answer as Sun and Abraham (2020) if you are not using covariates

R package `did`

- We see from implementing CS that we might have previously been *underestimating* our treatment effects.
- Can easily specify not-yet-treated and never-treated controls
- New Stata version by Rios-Avilia, Koren, Naqvi and Nichols

```
## [1] "Aggregate Group-Time ATE (Not Yet) = 0.1094"
```

```
## [1] "Aggregate Group-Time ATE (Never) = 0.1104"
```

Borusyak, Jaravel, and Spiess 2021

- Shows problems with TWFE due to heterogeneity of treatment timing, but offers some solutions

1. Separate Assumptions from Goals of estimator
2. Describes problems with common practice
3. Derive robust and efficient estimator from first principles
4. Large sample theory and inference using estimator
5. Approach to testing (separating from estimation)

Issues

- View: TWFE doesn't work because

1. Conflate the identifying assumptions of parallel trends and no anticipatory effects
2. Assumptions that restrict treatment effect heterogeneity
3. Specification of the estimand as a weighted average of treatment effects (Goodman-Bacon)

Issues, continued

- Not ruling out anticipation effects leads to a specification problem
  - “Fully dynamic” models (all leads and lags) are problematic
  - We estimate a model to validate pretrends, but are assuming no anticipation. . .
- Forbidden Comparison
- In no other method do we think it is OK to compare between different treatment cohorts when both have been treated
- similar to Goodman-Bacon

Response

- Response: Efficient estimator robust to treatment effect heterogeneity
  - Intuitive “imputation” form
  - Separate the assumption from the estimation

Implications/ Framework

1. Parallel Trends
2. No anticipation
3. Treatment-effect Model (optional)

Practice

Basically, an imputation estimator:

1. We know that our non-treated observations are

   $$Y^0 = \mu_i + \lambda_t$$

   and you can also use linear controls
2. Estimate: $\hat{\mu}_i$ and $\hat{\lambda}_t$ on all controls
3. Compute $\tau_{it} = Y_{1t} - \hat{Y}_{0t}$ to compute our weighted τ
4. Take averages of all the weighted τ_{it}s

Stata and R Packages : `did_imputation` and `event_plot`

Cengiz et al and Stacked DiD

- Cengiz et al (2019) Paper on minimum wage changes and low-wage jobs. Online Appendix D
- Create a set of unique datasets with each treated cohort separated out and “clean controls” for the corresponding time horizon
- Create “long” dataset, by appending (stacking) these newly created datasets together.

Stacked DiD

- Now we are estimating the following regression, where outcomes are regressed on treatment status and dataset-specific group and period fixed effects:

$$Y_{cgpit} = \lambda_{cg} + \lambda_{cp} + \beta D_{cgp} + \epsilon_{cgpit}$$

- where c is and indicator for dataset, g is an indicator for treatment cohort, i is the unit, and p is the time period.

Section 3

Other DiD papers/techniques

Worth looking into...

- Sun and Abraham (2021)
- Chaisemartin and D'Haultfœuille (2019)
- Gardner (2021) 2-Stage Difference in Differences (also nice explanation of Stacked DiD)
- Athey (2021) Imputation via matrix completion. Try to directly estimate the counterfactual.

Section 4

Concluding Remarks

What's next?

- New literature seems a worthwhile investment of time for empirical researcher
- Only scratched surface on these papers, meant to be almost a syllabus
- When do these papers apply to your setting?
  - Don't mention Goodman-Bacon if all treated units are treated at the same time!
  - Depends on problems you face and the assumptions you want to make

Section 5

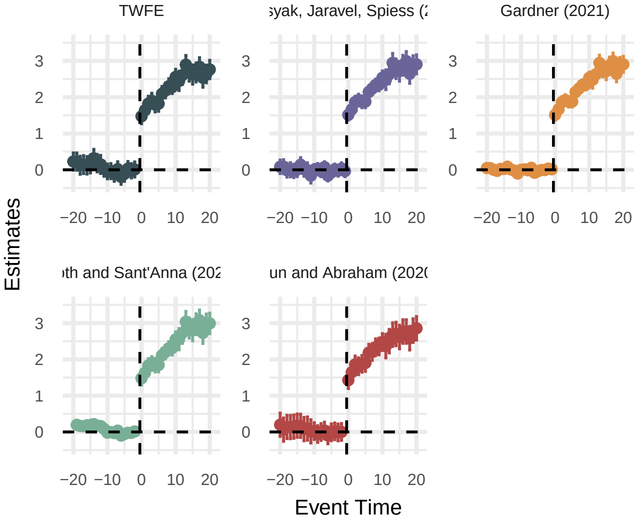
Appendix

More Resources

- Wooldridge 2-day DiD seminar December 14-15
- Scott Cunningham's "CodeChella" recordings
- Wooldridge on "saving" TWFE
- Taylor Wright's DiD Reading Group (with video)

All-in-One Package

- In both Stata and R `did2s` with function `event_study` shows you all the estimators in one plot!


TWFE
syak, Jaravel, Spiess (2
Gardner (2021)
th and Sant'Anna (202
un and Abraham (2020
Estimates
Event Time