Recent Advances in DiD Methods

A selective (and personal) perspective

Pedro H. C. Sant'Anna
Microsoft and Vanderbilt University

Brazilian Econometric Society (SBE) Meeting, December 2022

Popularity of Difference-in-Differences methods

Currie, Kleven and Zwiers (2020) at AEA P&P

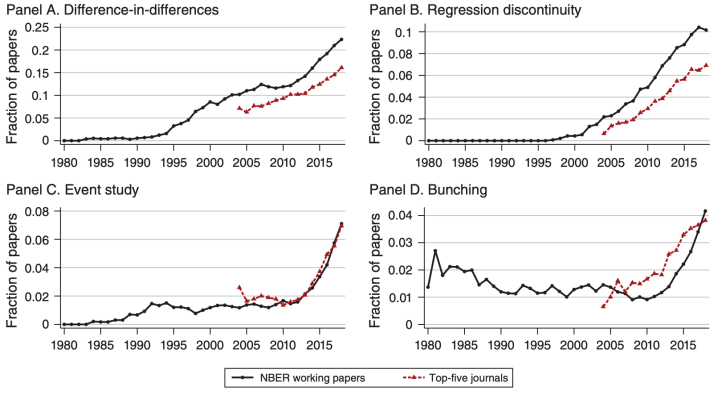


FIGURE 4. QUASI-EXPERIMENTAL METHODS

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 five-year moving averages.

Lookalike vs. Pre-Post vs. DiD

- Let's consider the case that **data do not come from an experiment or A/B test**.
- Since we are dealing with **observational data**, let's discuss some options

1. Rely on **lookalike** or **regression** or **double machine learning models.**

   Drawback: Rule out selection on unobservables.
   We need to have data on everything that affects treatment timing and outcome of interest (unconfoundedness assumption).

2. Rely on **Pre-Post analysis**

   Drawback: Does not account for potential trends in revenues.
   This is more reasonable if we study very short-run effects, but that is not usually the case.

The appeal of Difference-in-Differences

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.
- **DiD combines previous approaches to avoid their pitfalls**.
- **Advantage: Allow for selection on unobservables and for time-trends.**
  We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - **parallel trends assumption**.

Introduction

- The last few years have seen an explosion of econometrics on DiD, making it hard to keep up.

Introduction

- In Roth, Sant'Anna, Bilinski and Poe (2021), we attempt to synthesize the by-then recent literature and provide concrete recommendations for practitioners.
  - Canonical DiD setup.
  - Variation in treatment timing (problems and solutions).
  - Accessing and relaxing the parallel trends assumption (pre-tests, sensitivity analysis, incorporating covariates).
  - Inference with few clusters.

Introduction

- Since then, the literature kept evolving! Here is a sample of very recent topics:
  - DiD with continuous/multi-valued treatments.
    Callaway, Goodman-Bacon and Sant'Anna (2021)
  - When is DiD sensitive to functional form assumptions?
    Roth and Sant'Anna (2022)
  - What types of selection models are compatible with parallel trends?
    Ghanem, Sant'Anna and Wüthrich (2022)
  - How to incorporate Machine Learning into DiD?
    Chang (2020)
  - What if we have multiple treatments?
    de Chaisemartin and D'Haultfœuille (2022)

Structure of my Two Lectures

- I won't have time to cover everything I wish, so we will need to specialize.
- My main goals are to:

1. Expose everyone to the canonical DiD setup.
2. Discuss staggered treatment adoption setups

   2.1 Problems with Two-Way-Fixed Effects regressions
   Goodman-Bacon (2021), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021).

   2.2 Simple solutions to these problems
   Callaway and Sant'Anna (2021), Sun and Abraham (2021), Wooldridge (2021a), Borusyak, Jaravel and Spiess (2021)

3. Explain how we can embrace heterogeneity in staggered DiD setups and still identify useful parameters of interest

Let’s start with canonical DiD

Canonical DiD Setup

Canonical DiD Setup without Covariates

- Let's consider the canonical case:
  - 2 time periods: $t = 1$ (before treatment) and $t = 2$ (after treatment)
  - 2 groups: $G = 2$ (treated at period 2) and $G = \infty$ (untreated by period 2)
- $Y_t(g)$: Potential outcome at period t if units were exposed to treatment for the first time in period g.
- What causal parameter are we after?
- Main parameter of interest: Average Treatment Effect among Treated units

$$ATT \equiv \underbrace{\mathbb{E}\left[Y_{t=2}\left(2\right)|G=2\right]}_{\text{estimable from the data}} - \underbrace{\mathbb{E}\left[Y_{t=2}\left(\infty\right)|G=2\right]}_{\text{counterfactual component}}$$

Canonical DiD Setup without Covariates

Identification of the ATT is achieved via three main assumptions:

Assumption (SUTVA)

Observed outcomes at time t are realized as $Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$.

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ *for all groups in their pre-treatment periods, i.e., for all* $t < g$.

Assumption (Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right]$$

But how can these assumption help us?

Parallel Trends and the ATT

- We will start from the perspective that the *ATT* at time $t = 2$ is the target parameter.
- From the definition of the ATT and SUTVA, we have

$$\begin{aligned} ATT &\equiv \mathbb{E}\left[Y_{i,t=2}\left(2\right)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\ &= \underbrace{\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right]}_{\text{by SUTVA}} - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \end{aligned}$$

- Green object is estimable from data (under SUTVA).
- Red object still depends on potential outcomes, and our goal is to find ways to "impute" it.
- This is where PT and no-anticipation come into play!

Parallel Trends and the ATT

1) First, recall the PT assumption:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i=2\right]-\mathbb{E}\left[Y_{i,t=1}(\infty)|G_i=2\right]=\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i=\infty\right]-\mathbb{E}\left[Y_{i,t=1}(\infty)|G_i=\infty\right].$$

2) By simple manipulation, we can write it as

$$\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=2\right] = \mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i=2\right]+\left(\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=\infty\right]-\mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i=\infty\right]\right)$$

3) Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=2\right] = \underbrace{\mathbb{E}\left[Y_{i,t=1}\left(2\right)|G_i=2\right]}_{by\ No-Anticipation}+\left(\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=\infty\right]-\mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i=\infty\right]\right)$$

$$\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=2\right] = \underbrace{\mathbb{E}\left[Y_{i,t=1}|G_i=2\right]+\left(\mathbb{E}\left[Y_{i,t=2}|G_i=\infty\right]-\mathbb{E}\left[Y_{i,t=1}|G_i=\infty\right]\right)}_{by\ SUTVA}$$

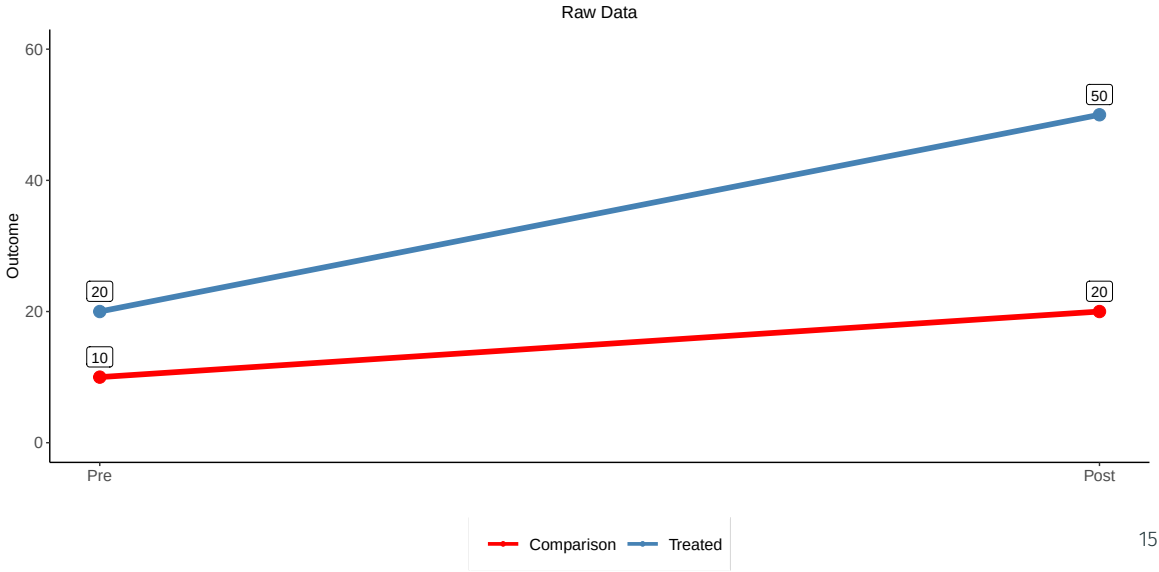
Parallel Trends and the ATT

- Combining these results together, we have that, under SUTVA + No-Anticipation + PT assumptions, it follows that

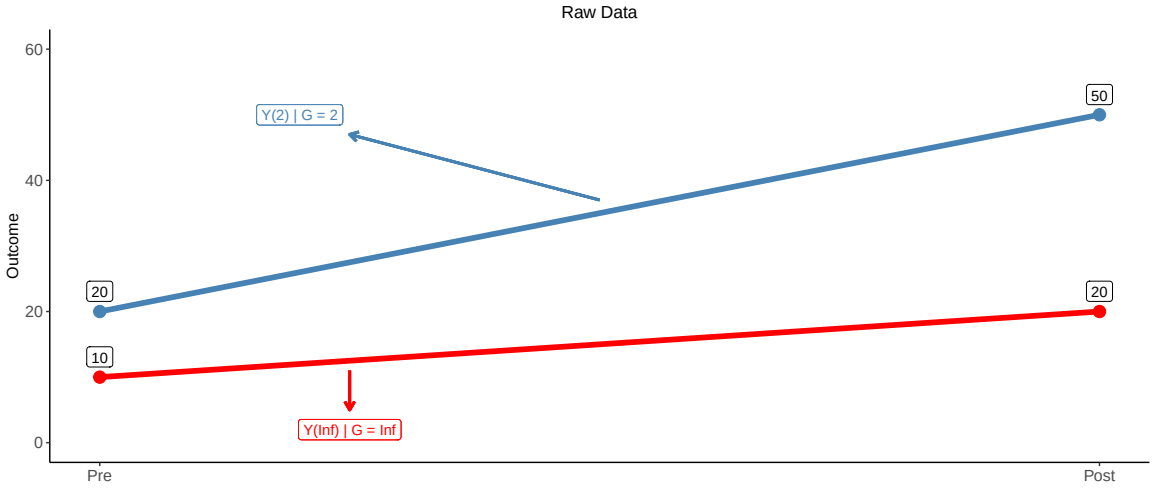
$$
\begin{aligned}
\textbf{ATT} &\equiv \mathbb{E}\left[Y_{i,t=2}\left(2\right)|G_i=2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i=2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i=2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i=2\right] - (\mathbb{E}\left[Y_{i,t=1}|G_i=2\right] + (\mathbb{E}\left[Y_{i,t=2}|G_i=\infty\right] - \mathbb{E}\left[Y_{i,t=1}G_i=\infty\right])) \\
&= (\mathbb{E}\left[Y_{i,t=2}|G_i=2\right] - \mathbb{E}\left[Y_{i,t=1}|G_i=2\right]) - (\mathbb{E}\left[Y_{i,t=2}|G_i=\infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i=\infty\right]) \\
&= \mathbb{E}\left[Y_{i,t=2} - Y_{i,t=1}|G_i=2\right] - \mathbb{E}\left[Y_{i,t=2} - Y_{i,t=1}|G_i=\infty\right]
\end{aligned}
$$

- This is “the birth” of the DiD estimand!

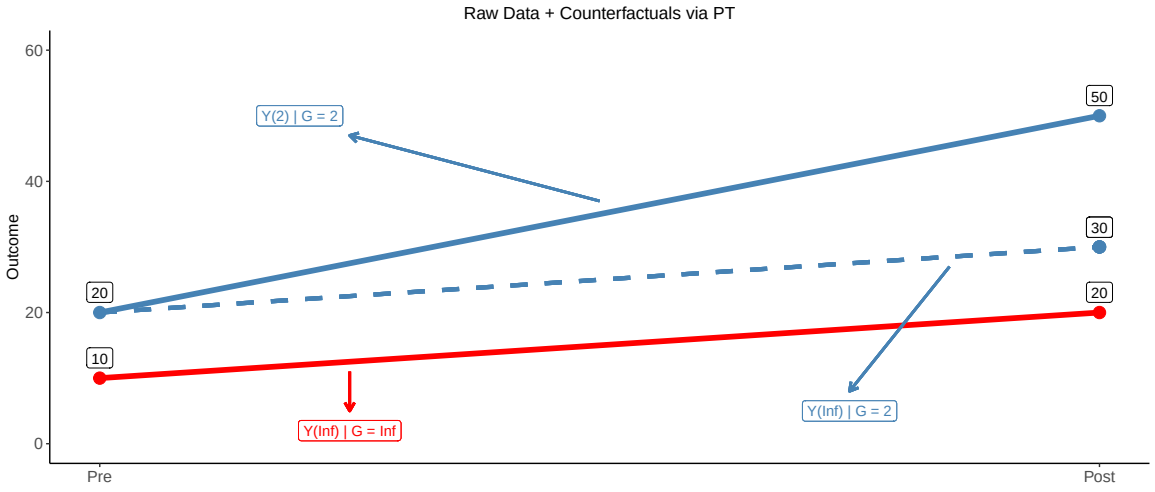
Parallel Trends via graphs



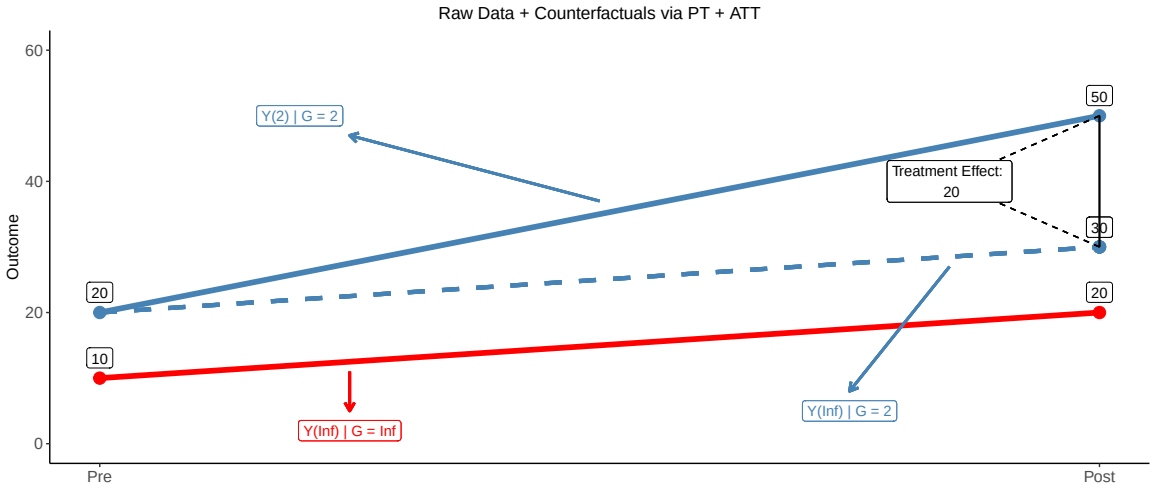
Parallel Trends via graphs



Parallel Trends via graphs



Parallel Trends via graphs



How do we estimate and make inference about the ATT?

“Brute force” DiD estimator

- Canonical DiD Estimator for the ATT:

$$\widehat{\theta}_n^{DiD} = (\bar{Y}_{g=2,t=2} - \bar{Y}_{g=2,t=1}) - (\bar{Y}_{g=\infty,t=2} - \bar{Y}_{g=\infty,t=1}) \,.$$

- But how to get standard errors?

- We can get the estimator’s asymptotic linear representation (influence function), but not many people like that.

‘TWFE” DiD estimator

- In practice, most of us would rely on the following TWFE regression specification:

$$Y_{i,t} = \alpha_0 + \gamma_0 1\{G_i = 2\} + \lambda_0 1\{T_i = 2\} + \underbrace{\beta_0^{twfe}}_{\equiv ATT}(1\{G_i = 2\} \cdot 1\{T_i = 2\}) + \varepsilon_{i,t},$$

  where we assume that $\mathbb{E}[\varepsilon_{i,t}|G_i, T_i] = 0$ *almost surely*.

- As long as number of treated and untreated “clusters” is large, we can use our favorite regression tools to estimate the ATT and make inferences about it.

Difference-in-Differences in Practice

- Many DiD empirical applications, however, deviate from the standard DiD setup:
  - Availability of covariates *X*;
  - More than two time periods;
  - Variation in treatment timing;
  - Non-binary treatments;
  - Parallel trends may not hold exactly.
  - Only a few treated and untreated clusters are available

Let’s focus on staggered treatment adoption

www.causal-solutions.com

Does TWFE “work” in setups with variation in treatment timing?

Recent Boom of New DiD Methods: TWFE Diagnostics

- What if we have staggered treatment adoption?
- It is tempting to use variations of the following TWFE specification:

$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}$$

  where $D_{i,t}$ is an indicator for unit i being treated by period t.

- Does β recover any interesting causal parameter of interest?
  - **Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), and Athey and Imbens (2021)** tackle this question.
- When TE are heterogeneous, β does not recover an easy-to-interpret parameter: **weighted average of ATT's, but some weights can be negative!**
- In my opinion, Goodman-Bacon (2021) explains this in the clearest way.

Traditional methods: TWFE regressions

- We know that, in the 2x2 case,

$$Y_{i,t} = \alpha_0 + \gamma_0 1\{G_i = 2\} + \lambda_0 1\{T_i = 2\} + \underbrace{\beta_0^{twfe}}_{\equiv ATT}(1\{G_i = 2\} \cdot 1\{T_i = 2\}) + \varepsilon_{i,t},$$

- It is tempting to "extrapolate" from this setup and use variations of the following TWFE specification to estimate causal effects:

$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}$$

  where dummies $D_{i,t} = 1\{t - G_i \geq 0\}$, where G_i indicates the period unit i is first treated (Group).

- $D_{i,t}$ is an indicator for unit i being treated by period t.

- For simplicity, let's assume that treatment is "irreversible": once a unit is treated, it is forever treated - aka **staggered design**

Does TWFE “work” in setups with variation in treatment timing?

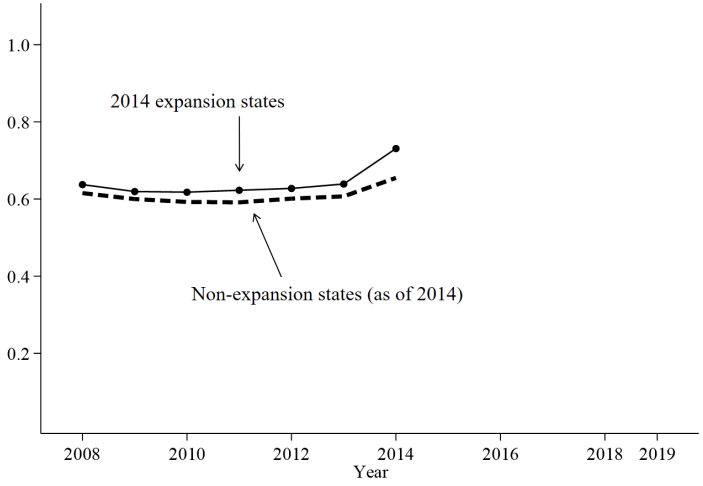
Example: Effect of ACA Medicaid Expansion on Health Insurance rate

Empirical Example: Medicaid Expansion

- To motivate our problem, let's look at a classical example: Medicaid Expansion
- We want to analyze its effect on health insurance rate among low-income, childless adults aged 25-64.

ACA Medicaid Expansion Circa 2014

Figure 1: Health Insurance Rate (low-income Childless Adults Aged 25-64)



ACA Medicaid Expansion Circa 2015

Figure 2: Health Insurance Rate (low-income Childless Adults Aged 25-64)

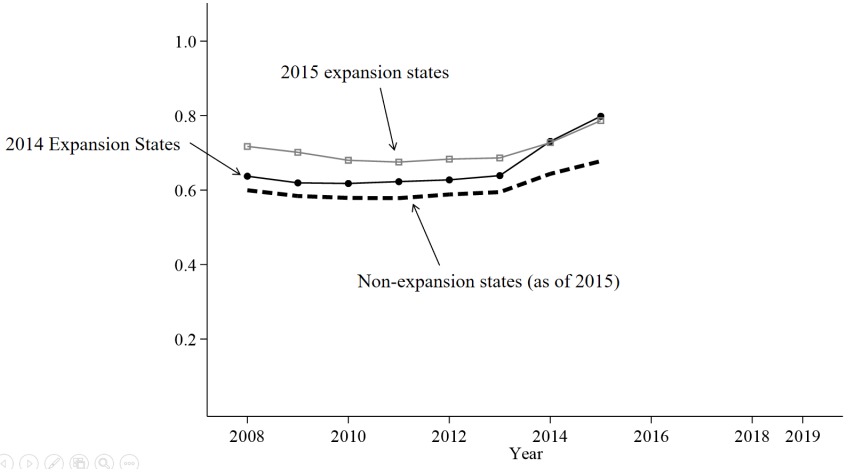
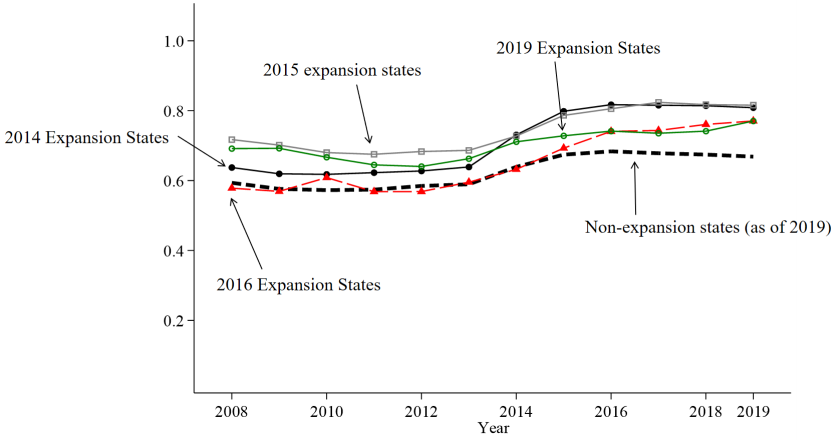


Figure 3: Health Insurance Rate (low-income Childless Adults Aged 25-64)



ACA Medicaid Expansion Circa 2019

- 23 states expanded circa 2014 - 4 did it earlier (ACA is effectively relabeled), we drop them.
- 3 states expanded circa 2015
- 2 states expanded circa 2016
- 1 states expanded circa 2017
- 2 states expanded circa 2019
- 16 states haven't expanded by 2019

OLS estimate of β

- Let $\widehat{\beta}$ be the OLS estimator of the following TWFE regression specification:

$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}$$

- What is $\widehat{\beta}$?

- Goodman-Bacon (2021) shows that we can answer this question following these three steps:
  1. Remove unit means

$$D_{i,t} - \overline{D}_i$$

  2. Remove time means of $(D_{i,t} - \overline{D}_i)$:

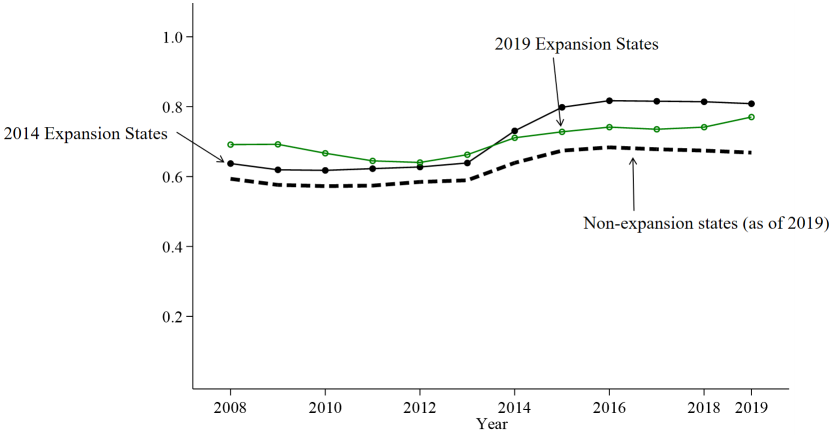
$$\widetilde{D}_{i,t} = (D_{i,t} - \overline{D}_i) - (\overline{D}_t - \overline{\overline{D}})$$

  3. Calculate univariate regression of $Y_{i,t}$ on $\widetilde{D}_{i,t}$:

$$\widehat{\beta} = \frac{(nT)^{-1} \sum_{i,t} Y_{i,t} \cdot \widetilde{D}_{i,t}}{(nT)^{-1} \sum_{i,t} \widetilde{D}_{i,t}^2}$$

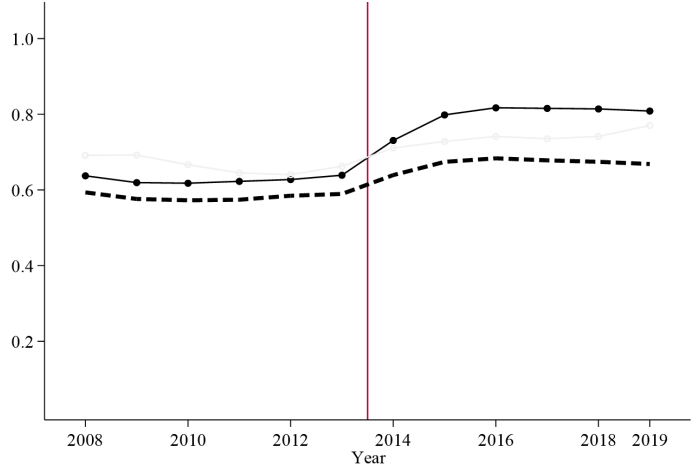
Three Groups Example

Figure 4: Health Insurance Rate (low-income Childless Adults Aged 25-64)



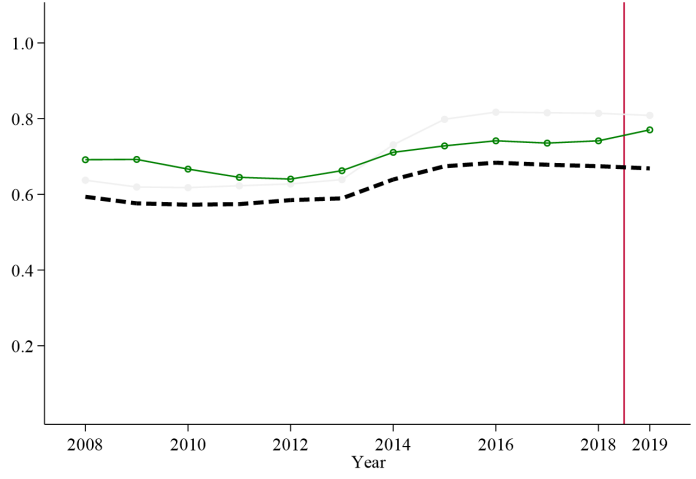
Treated in 2014 vs. Never-Treated

Figure 5: Health Insurance Rate (low-income Childless Adults Aged 25-64)



Treated in 2019 vs. Never-Treated

Figure 6: Health Insurance Rate (low-income Childless Adults Aged 25-64)

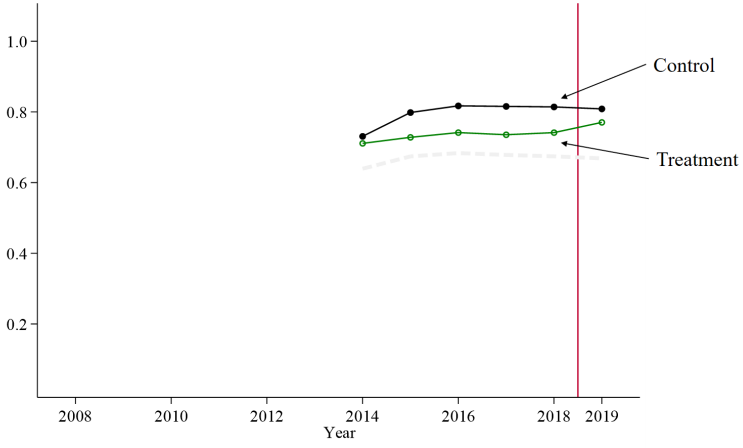


Treated in 2014 vs. Treated in 2019 ($t < 2019$)

Figure 7: Health Insurance Rate (low-income Childless Adults Aged 25-64)

Treated in 2019 vs. Treated in 2014 ($t \geq 2014$)

Figure 8: Health Insurance Rate (low-income Childless Adults Aged 25-64)



OLS estimate of β

- OLS is "variational hungry" and exploit all these 2x2 comparisons.
- But how does OLS aggregate them?
- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

$$\widehat{\beta} = s_{k,U} \cdot \widehat{\beta}_{k,U} + s_{\ell,U} \cdot \widehat{\beta}_{\ell,U} + \left[s_{k,\ell} \cdot \widehat{\beta}_{k,\ell} + s_{\ell,k} \cdot \widehat{\beta}_{\ell,k} \right]$$

- In our example:
  - $k = 2014$
  - $\ell = 2019$
  - $U =$ never-treated

Does TWFE “work” in setups with variation in treatment timing?

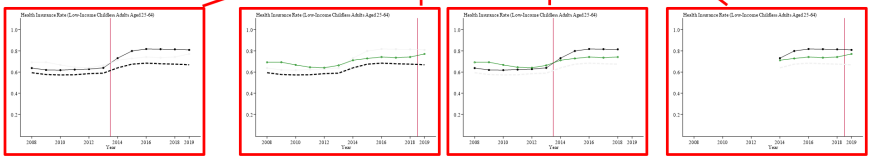
Bacon Decomposition

Bacon-Decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 9: Bacon-Decomposition: The 2x2 $\widehat{\beta}$

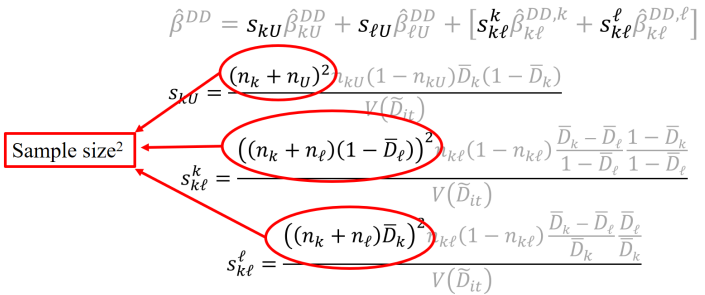
$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$



Bacon-Decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 10: Bacon-Decomposition: The weights



Bacon-Decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 11: Bacon-Decomposition: The weights

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU}(1 - n_{kU})\bar{D}_k(1 - \bar{D}_k)}{V(\tilde{D}_{it})}$$

$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\bar{D}_k\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

If you did TWFE on this subsample, what would the variance of $\tilde{D}_{it}$ be?

Bacon-Decomposition: General case

Theorem (Goodman-Bacon (2021) decomposition)

Assume that there are $k = 1, \ldots, K$ groups of treated units ordered by treatment time t_k^* and one "never-treated" group, U, which does not receive treatment in the data. The share of units in group k is n_k, and the share of periods that group k spends under treatment is $\overline{D}_k$. The regression estimate from a two-way fixed effects model is a weighted average all two-group DiD estimators:

$$\widehat{\beta} = \sum_{k \neq U} \left(s_{k,U} \cdot \widehat{\beta}_{k,U} \right) + \sum_{k \neq U} \sum_{\ell > k} \left(s_{k,\ell} \cdot \widehat{\beta}_{k,\ell} + s_{\ell,k} \cdot \widehat{\beta}_{\ell,k} \right),$$

where the weights are given by

$$s_{k,U} = \frac{(n_k + n_U)^2 \widehat{V}_{k,U}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)}, \quad s_{k,\ell} = \frac{\left((n_k + n_\ell)\left(1 - \overline{D}_\ell\right)\right)^2 \widehat{V}_{k,\ell}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)}, \quad s_{\ell,k} = \frac{\left((n_k + n_\ell)\,\overline{D}_k\right)^2 \widehat{V}_{\ell,k}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)},$$

such that $\sum_{k \neq U} s_{k,U} + \sum_{k \neq U} \sum_{\ell > k} \left(s_{k,\ell} + s_{\ell,k}\right) = 1$.

What does this mean to TWFE regressions?

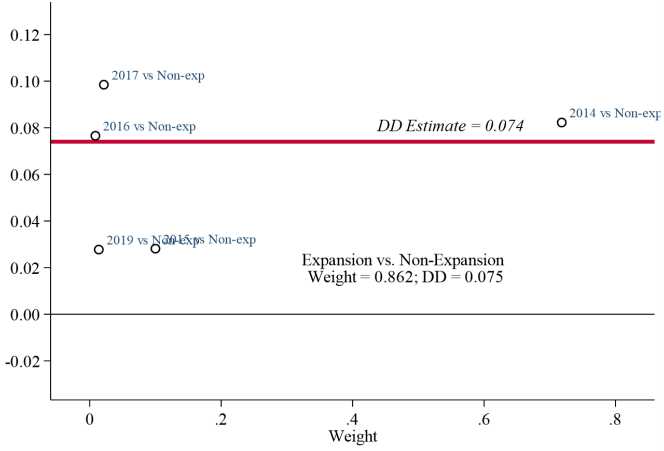
TWFE computes weighted-averages of 2x2 DiD's

- $\widehat{\beta} = 0.074$ in the empirical application.
- OLS weights use sample size and variance
- Is that what you really want?

- TWFE exploits all 2x2 DiD comparisons
  - Treated vs. "Never-treated"
  - Early-treated vs. Later-treated
  - Later-treated vs. Already-treated
- Are all these comparisons "reasonable" to attach a causal interpretation to $\widehat{\beta}$?

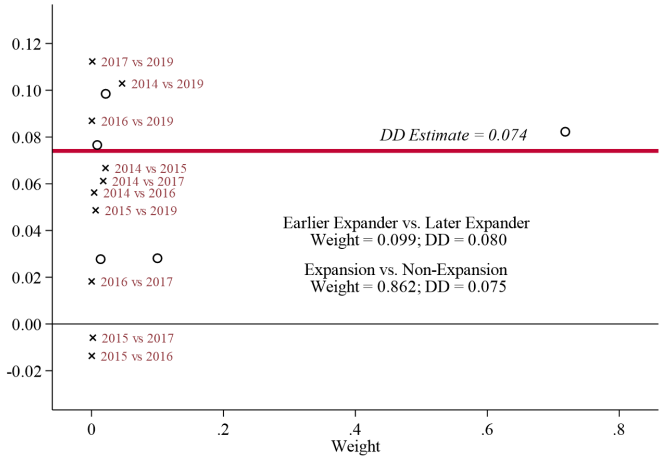
Bacon-Decomposition: Treated vs. Never-Treated

Figure 12: Bacon-Decomposition: The weights



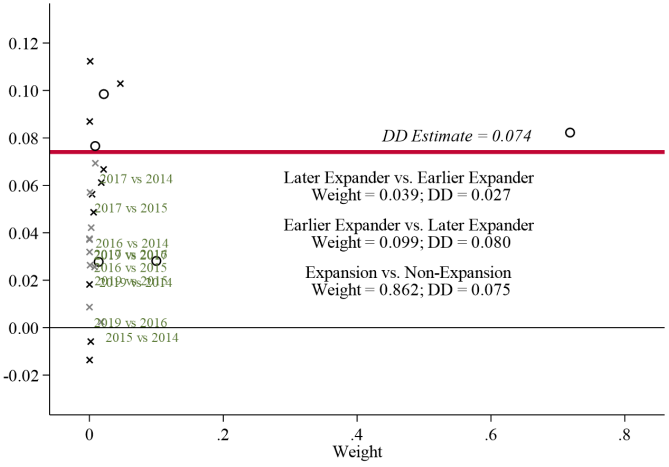
Bacon-Decomposition: Early-Treated vs. Later-treated

Figure 13: Bacon-Decomposition: The weights



Bacon-Decomposition: Later-treated vs. Early-Treated

Figure 14: Bacon-Decomposition: The weights



TWFE regressions, in general,

do not recover an easy-to-interpret

causal parameter of interest,

unless we rule out TE heterogeneity/dynamics

How do we know this?

TWFE, Identifying Assumptions, and Causal Effects

- Goodman-Bacon (2021) decomposition is "mechanical" in the sense that it does not rely on any (causal) assumptions.
- To endow the decomposition with a causal interpretation, we need to make some assumptions - PT and no-anticipation, or restrict assignment mechanisms.
- It is also worth stressing that Goodman-Bacon (2021) decomposition is not "unique".
- If you choose a different "building block" than the "time-averaged" 2x2 DiD estimates, you get a different decomposition.
- Two alternative characterizations worth mentioning are those of Athey and Imbens (2021) and de Chaisemartin and D'Haultfœuille (2020).
- Let's zoom into de Chaisemartin and D'Haultfœuille (2020), as they impose additional assumptions to get causal effects interpretation

Does TWFE “work” in setups with variation in treatment timing?

de Chaisemartin and D’Haultfœuille (2020) Decomposition

de Chaisemartin and D'Haultfœuille (2020)

- de Chaisemartin and D'Haultfœuille (2020) consider a setup where treatment may turn on and off across time.
- For simplicity and easy-of-interpretation, we will focus on the staggered case (treatment is "irreversible").
- My notation will also impose a random sampling setup, which is different from what they do in their paper.
- However, it greatly simplifies the exposition.

de Chaisemartin and D'Haultfœuille (2020)

- Let us introduce the unit-specific treatment effect

$$\Delta_{i,t}^{g} = Y_{i,t}(g) - Y_{i,t}(\infty)$$

- Let $\epsilon_{i,t}$ be the error of the following TWFE specification:

$$D_{i,t} = \alpha_i + \alpha_t + \epsilon_{i,t}$$

- Consider the weights

$$w_{i,t} = \frac{\epsilon_{i,t}}{N_1^{-1} \sum_{i,t:D_{i,t}=1} \epsilon_{i,t}},$$

  where $N_1 = \sum_{i,t} D_{i,t}$

- **Strong unconditional PTA:** Assume that for every time period t and every group g, g',

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | G = g\right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | G = g'\right]$$

Theorem (de Chaisemartin and D'Haultfœuille (2020) decomposition)

Suppose SUTVA, No-anticipation and the Strong unconditional PT hold. Let β be TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}.$$

Then, it follows that

$$\beta = \mathbb{E}\left[\sum_{i,t:D_{i,t}=1} \frac{1}{N_1} w_{i,t} \cdot \Delta^g_{i,t}\right],$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N_1} = 1$, but $w_{i,t}$ can be negative.

- Weights are non-convex and can be negative
- Goodman-Bacon (2021) made it clear why: we are using already-treated units as comparison groups to "later treated" units; see also Borusyak and Jaravel (2017).

Do we have negative weights in our application?

- In our application, we do not have negative weights, though.
- This is expected, as most of the states got treated in 2014 and we have a relatively big “never-treated” group.
- Does this mean that TWFE “worked”?
- Weights being non-negative is a very minimal requirement.
- The fact that we do not really understand the weights attached to each ATT makes TWFE unattractive.

What happens when we consider a TWFE event-study specification?

Event-Study via TWFE specifications

Event-Study via TWFE specifications

- One of the main attractive features of observing multiple time periods is that we can attempt to "learn" about treatment effect dynamics.
- Status-quo in the literature is to consider variants of the TWFE event-study regression

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^k + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^k + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t}$$

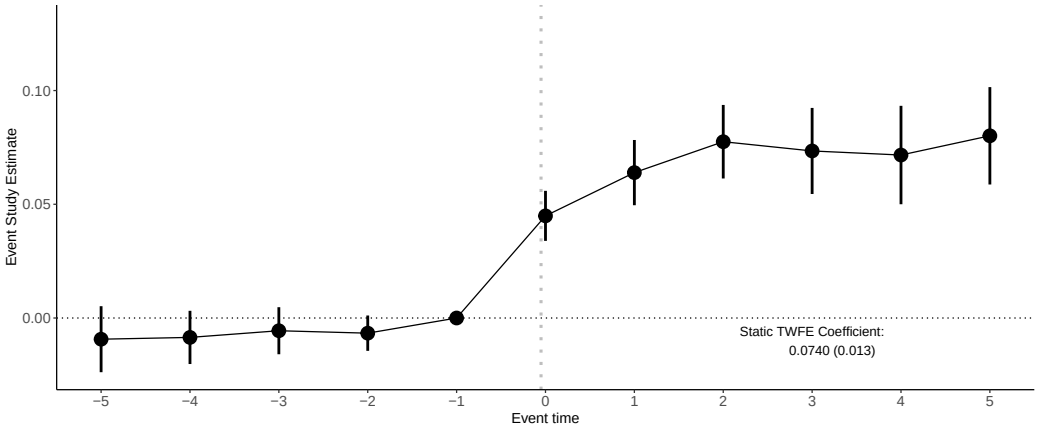
  with the event study dummies $D_{i,t}^k = 1\{t - G_i = k\}$, where G_i indicates the period unit i is first treated (Group).

- $D_{i,t}^k$ is an indicator for unit i being k periods away from initial treatment at time t.

Does this strategy “work”?

ACA Medicaid Expansion: TWFE Event-study specification

Figure 15: Health Insurance Rate (low-income Childless Adults Aged 25-64



Event-Study via TWFE specifications

- Can we (a priori) “trust” these results?
- What type of treatment effect parameter is being reported in this event-study?
- What kind of assumptions are we implicitly relying on?
- What kind of comparisons are being made “behind the scenes”?
- These are important questions!

Event-Study via TWFE specifications

Sun and Abraham (2021)

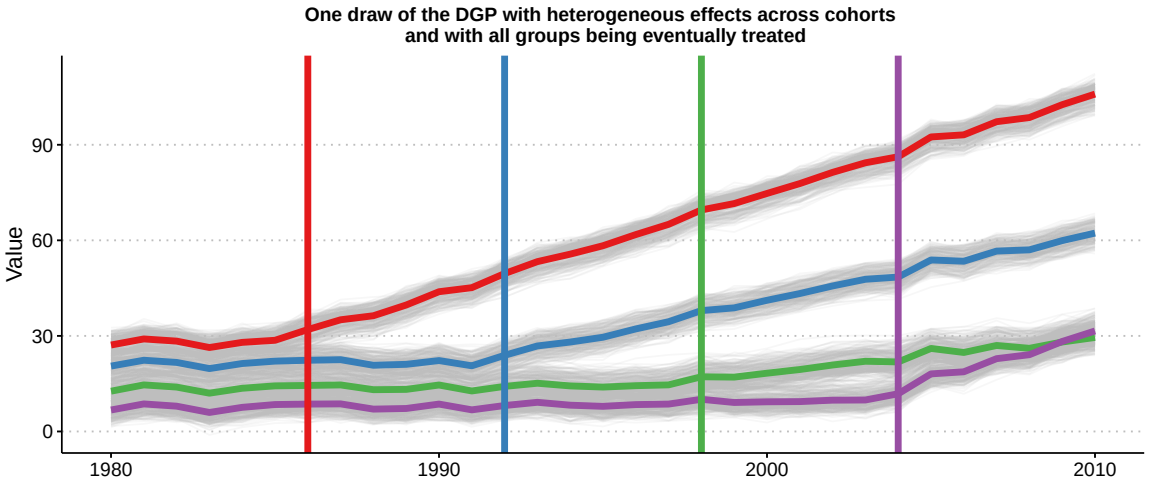
Problem with Event-Study via TWFE specifications: Sun and Abraham (2021)

- Sun and Abraham (2021) bring "bad" news, once again!
- Even when we impose the Strong unconditional parallel trends and the no-anticipation assumption, the OLS coefficients of the TWFE ES specification are, in general, very hard to interpret.
- Coefficient on a given lead or lag can be contaminated by effects from other periods
- Pre-trends can arise solely from treatment effects heterogeneity!
- Even under treatment effect homogeneity across cohorts (they all share same dynamics in event-time), the OLS coefficients can still be contaminated by treatment effects from the excluded periods.

Event-Study via TWFE specifications

Stylized example using simulated data

Stylized example using simulated data



Stylized example using simulated data

- 1000 units ($i = 1, 2, \ldots, 1000$) from 40 states ($state = 1, 2, \ldots, 40$).
- Data from 1980 to 2010 (31 years).
- 4 different groups based on year that treatment starts: $g = 1986, 1992, 1998, 2004$.
- Randomly assign each state to a group.
- Outcome:

$$Y_{i,t} = \underbrace{(2010 - g)}_{\text{cohort-specific intercept}} + \underbrace{\alpha_i}_{N\left(\frac{state}{5},1\right)} + \underbrace{\alpha_t}_{\frac{(t-g)}{10}+N(0,1)} + \underbrace{\tau_{i,t}}_{\mu_g \cdot (t-g+1) \cdot 1\{t \geq g\}} + \underbrace{\varepsilon_{i,t}}_{N\left(0,\left(\frac{1}{2}\right)^2\right)}$$

- $\mu_{1986} = \mu_{2004} = 3,\ \mu_{1992} = 2,\ \mu_{1998} = 1$
- ATT for group g at the first treatment period is μ_g, at the second period since treatment is $2 \cdot \mu_g$, etc.

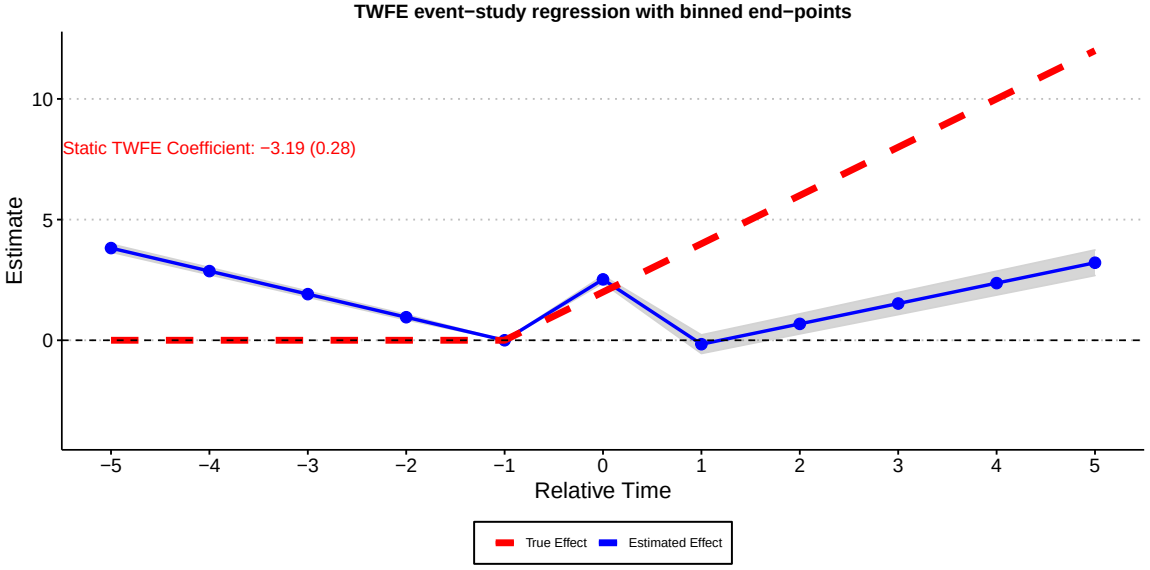
Traditional methods: TWFE event-study regression

- What if we tried to estimate the treatment effects using traditional TWFE event-study regressions,

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^k + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^k + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t},$$

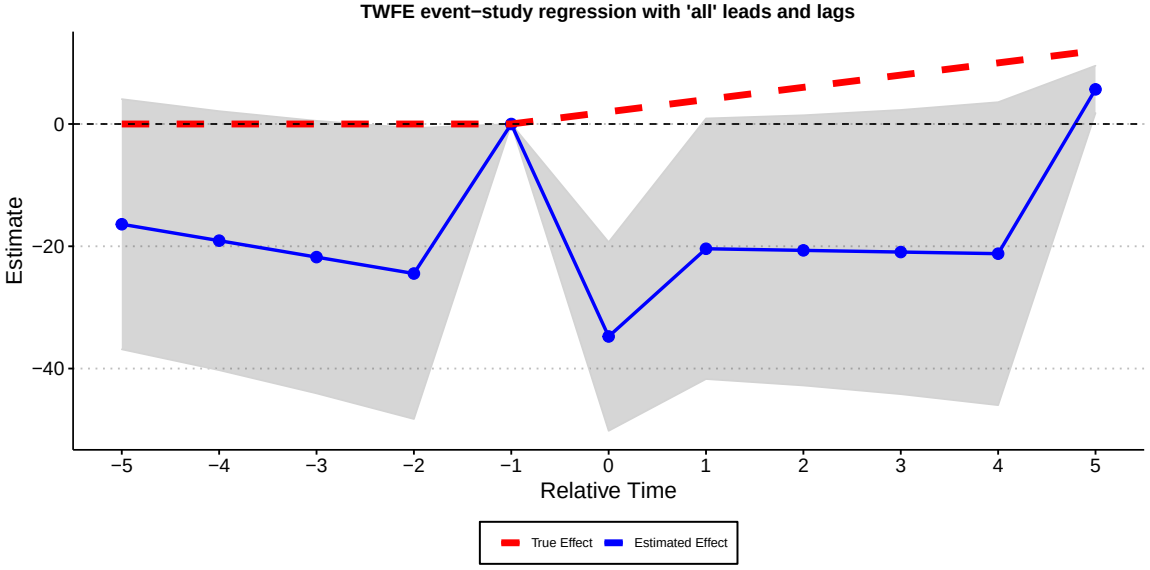
  with K and L to be equal to 5 ?

- Simulate data and repeat 1,000 times to compute bias and simulation standard deviations.

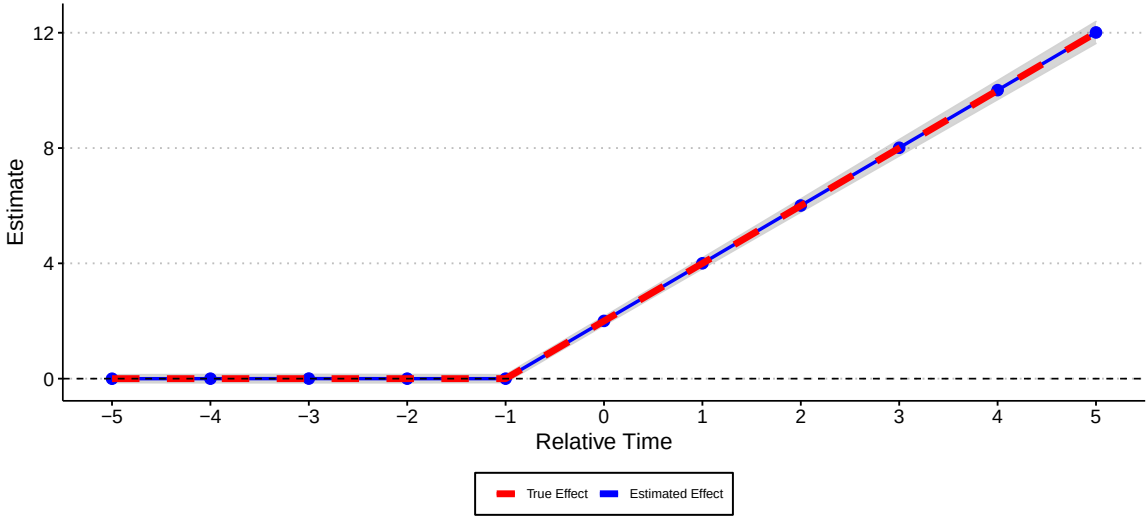

TWFE event−study regression with binned end−points
Static TWFE Coefficient: −3.19 (0.28)
Estimate
10
5
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

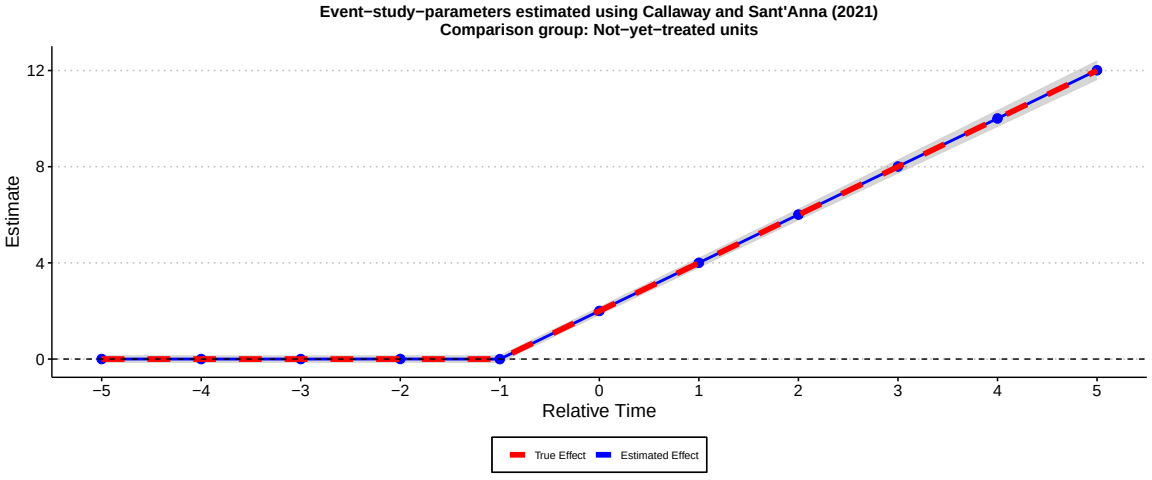
Traditional methods: TWFE event-study regression

- What if we include all possible leads and lags in the TWFE event study specification, i.e., to set K and L to the maximum allowable in the data, making inclusion of $D_{i,t}^{<-K}$ and of $D_{i,t}^{>L}$ unnecessary ?


TWFE event−study regression with 'all' leads and lags
Estimate
0
−20
−40
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

Is there hope?


Event−study−parameters estimated using Callaway and Sant'Anna (2021)
Comparison group: Last−treated−Cohort units
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect


Event−study−parameters estimated using Callaway and Sant'Anna (2021)
Comparison group: Not−yet−treated units
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

Recent Boom of New DiD Methods: Solutions to the TWFE problems

- The problems associated with using standard TWFE specifications are evident.
- OLS is variational hungry but causal inference is variational cautious!
- How to solve the TWFE problem in DiD setups?
- Ensure that you only make the comparisons you want to
- **Callaway and Sant'Anna (2021)** propose a guided and transparent way to do this!
  - Allow for covariates, different comparison groups, panel and repeated cross-sections.
  - **Separate the analysis into identification, aggregation, and estimation/inference.**

Addressing the TWFE problems

Recent Boom of New DiD Methods: Solutions to the TWFE problems

- Callaway and Sant'Anna (2021) is not the only game in town:
  - **Sun and Abraham (2021)**: Proposed estimator coincides with CS when there are no covariates and use the never-treated/last-treated cohort as a comparison group. However, this paper has many other results about the pitfalls of TWFE that are not in CS.
  - **Gardner (2021)**, **Borusyak et al. (2021)** and **Wooldridge (2021b)**: Propose "imputation"/regression based methods to recover cohort-time ATT's . These three papers do not nest nor is nested by CS, but identification assumptions are sometimes stronger. Benefit: more precise estimates when these assumptions are correct.
  - **Wooldridge (2021a)**: Propose estimators that are suitable for nonlinear models. It relies on alternative types of parallel trends assumptions, e.g. 'ratio-in-ratios" if exponential model. If use canonical link functions, standard errors can be easily estimated.

Recent Boom of New DiD Methods: Solutions to the TWFE problems (cont.)

- Callaway and Sant'Anna (2021) is not the only game in town:
  - **de Chaisemartin and D'Haultfœuille (2020, 2021)**: Estimator coincides with CS when there are no covariates, uses not-yet-treated units as comparison group, and treatment is staggered. However, these two papers allow for treatment turning on-off, which is not allowed in CS. de Chaisemartin and D'Haultfœuille (2020), though, relies on stronger assumptions and rules out dynamic treatment effects.

    When covariates are available, these papers do not nest nor are nested by CS. However, they seem to implicitly impose homogeneity assumptions wrt to X (e.g., ATT does not vary according to age).
  - **Roth and Sant'Anna (2021)**: When treatment timing is as-good-as-random, we can do much better than DiD in terms of efficiency. However, it requires more than PT. Does not nest nor is nested by CS.

References

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2021, (Forthcoming).

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2017.

_ , _ , and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," 2021.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

_ , Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna, "Difference-in-Differences with a Continuous Treatment," *arXiv:2107.02637*, 2021.

Chang, Neng-Chieh, "Double/debiased machine learning for difference-in-differences models," *The Econometrics Journal*, 2020, *23* (2), 177––191.

Currie, Janet, Henrik Kleven, and Esmée Zwiers, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, May 2020, *110*, 42–48.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

_ and _, "Difference-in-Differences Estimators of Intertemporal Treatment Effects," 2021.

_ and _, "Two-way Fixed Effects Regressions with Several Treatments," *arXiv:2012.10077*, 2022.

Gardner, John, "Two-Stage Difference-in-Differences," Technical Report, Working Paper 2021.

Ghanem, Dalia, Pedro H. C. Sant'Anna, and Kaspar Wüthrich, "Selection and parallel trends," *arXiv:2203.09001[econ]*, 2022.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2).

Roth, Jonathan and Pedro H. C. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2022, *Forthcoming*.

_ **and Pedro H.C. Sant'Anna**, "Efficient Estimation for Staggered Rollout Designs," *Working Paper*, 2021.

_ **, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe**, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *arXiv:2201.01194*, 2021.

Sun, Liyan and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2).

Wooldridge, Jeffrey M., "Nonlinear Difference-in-Differences with Panel Data," *Working Paper*, 2021.

Wooldridge, Jeffrey M, "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *Working Paper*, 2021, pp. 1–89.