Difference-in-Differences Methods

A brief overview

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

Population Dynamics and Health Program Workshop
University of Michigan

CAUSAL SOLUTIONS

Popularity of Difference-in-Differences methods

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

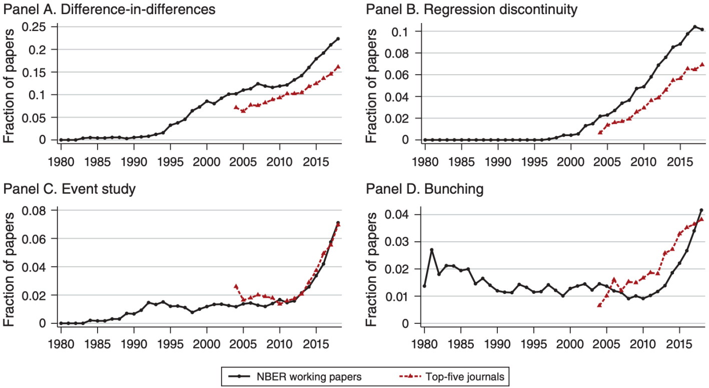


Figure 4. Quasi-experimental Methods

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 a randomized experiment**.

- Since we are dealing with **observational data**, let's discuss some options

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

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

Lookalike vs. Pre-Post vs. DiD

- Let's consider the case that **data do not come from a randomized experiment**.
- Since we are dealing with **observational data**, let's discuss some options

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

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

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

   Drawback: Does not account for potential trends in outcomes.
   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**.

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**.

We need to discuss why Parallel Trends is a plausible assumption in our application.

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 has 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 (2022a)
  - 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 4h Lecture

- 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.

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?

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

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

- 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

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

3) Now, exploiting No-Anticipation and SUTVA:

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

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

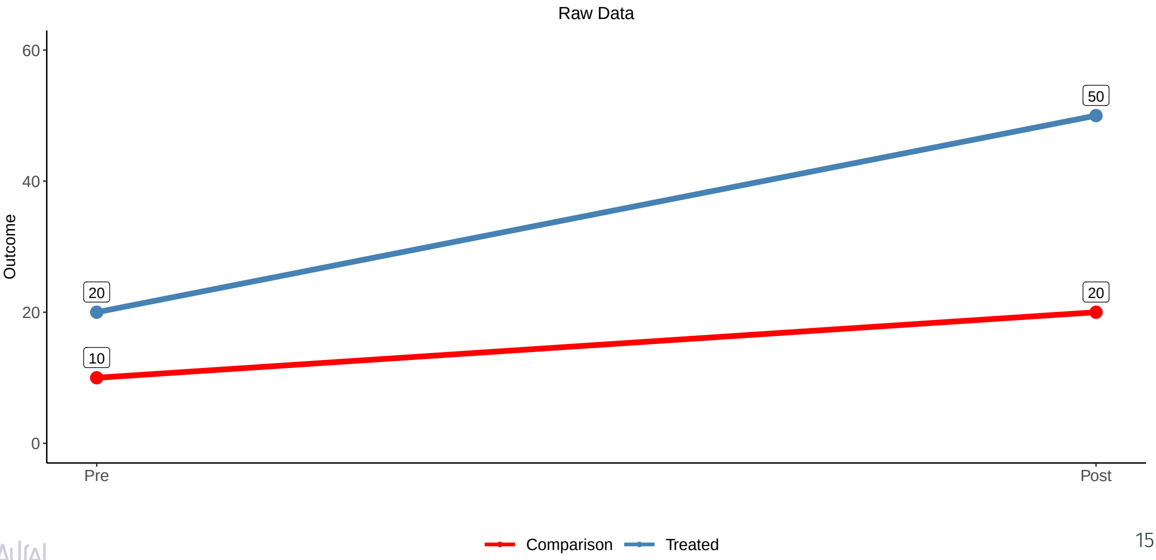
Parallel Trends and the ATT

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

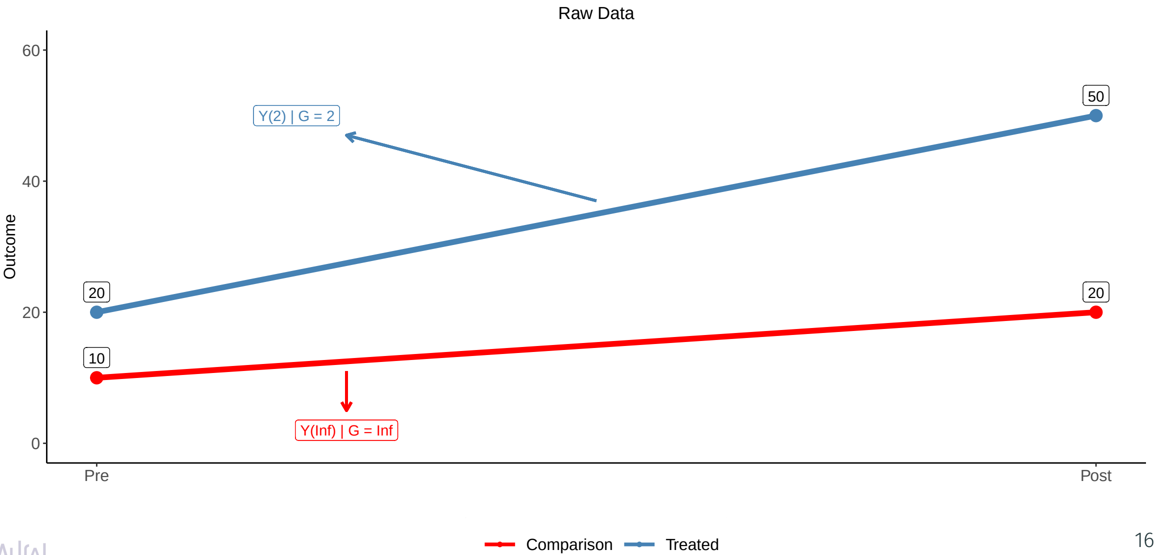
$$
\begin{aligned}
\textbf{ATT} &\equiv \mathbb{E}\left[Y_{i,t=2}\left(2\right)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i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

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

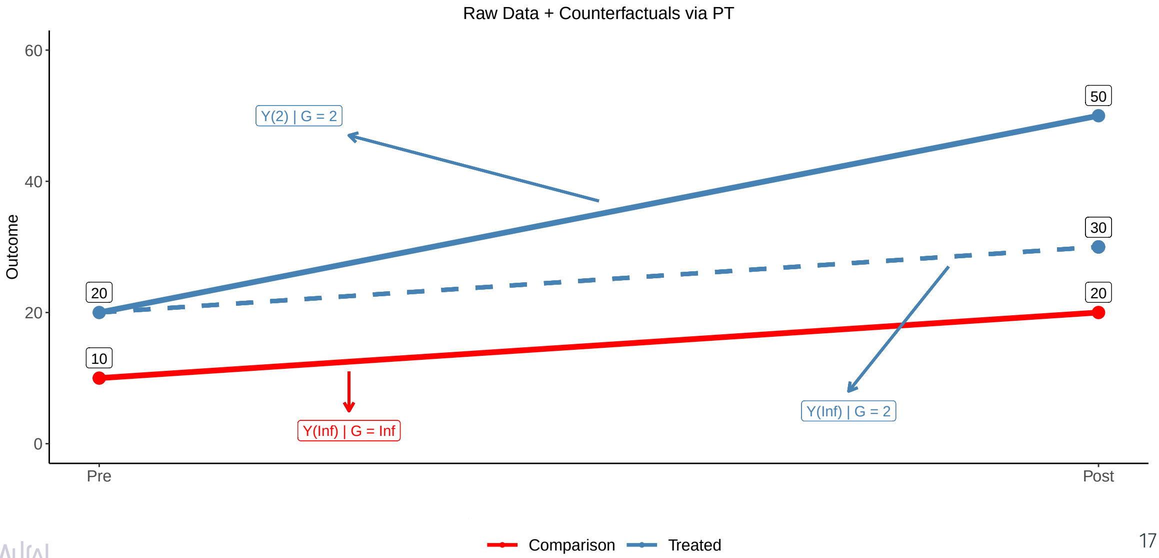
Parallel Trends via graphs



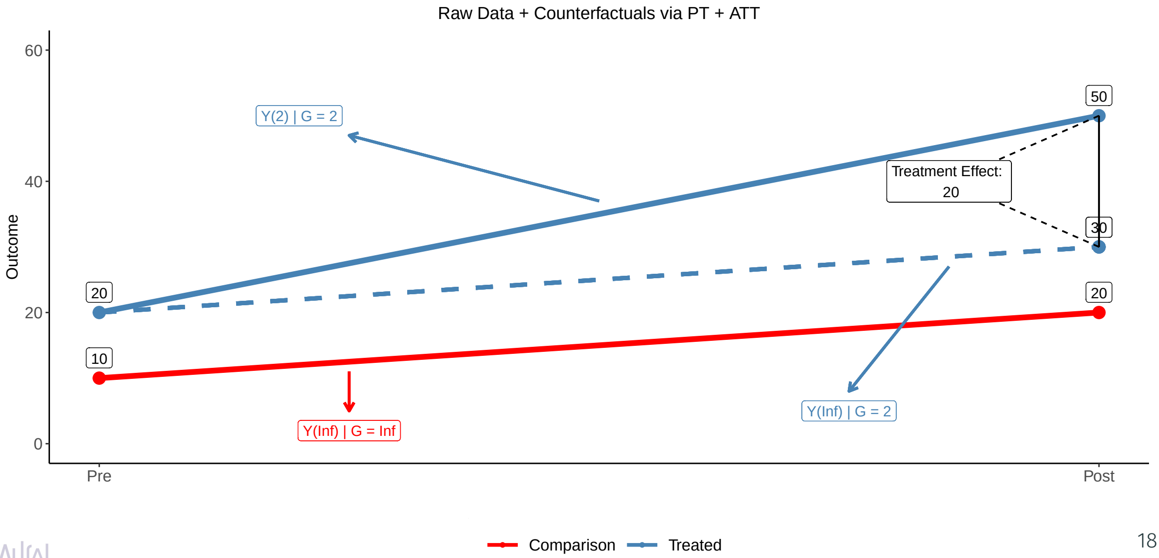
Parallel Trends via graphs



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:

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

- But how to get standard errors?

“Brute force” DiD estimator

- Canonical DiD Estimator for the ATT:

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

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

Let’s do an exercise in R

Parallel trends for different transformations of Y

Parallel Trends in levels

- **Parallel trends assumption (in levels):**

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

- If Y is the duration of claims measured in weeks, and treatment is an increase of cap (PT in levels)
  - PT would suggest that the average untreated claims' duration among workers who are affected by the increased cap would evolve the same as the average untreated claims' duration among workers who are not affected by the change in cap.
  - If the average change in untreated claims' duration among workers who are not affected by the change in cap is 2 weeks, these would serve as counterfactual changes for the average untreated claims' duration among workers who are affected by the change in cap
  - ATT would provide the average treatment effect (in weeks) among workers who are affected by the change in cap.

Parallel Trends in logs

- **Parallel trends assumption (in logs):**

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

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

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

- Under parallel trends (in logs), the ATT would take the format:

$$ATT = \mathbb{E}\left[\ln Y_{i,t=2}(2) - \ln Y_{i,t=2}(\infty) \mid G = 2\right] = \mathbb{E}\left[\ln \frac{Y_{i,t=2}(2)}{Y_{i,t=2}(\infty)} \,\middle|\, G = 2\right].$$

- ATT is measured in relative terms when you have PT in logs.

Parallel Trends in logs

- **Parallel trends assumption (in logs):**

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

- If Y is the duration of claims measured in weeks, and treatment is an increase of cap (PT in levels)
  - PT would suggest that the average log relative growth of untreated claims' duration among workers who are treated would be the same as the average log relative growth of untreated claims' duration among workers who are not treated
  - If the average log relative growth of untreated claims' duration among workers who are not treated is 0.10, these would serve as counterfactual changes for the average log relative growth of untreated claims' duration for workers that are treated.
  - ATT would provide the average treatment effect (in relative terms) among workers who are treated.

Which PT should we pick?

What if we take other transformations?

Roth and Sant’Anna (2022): When is PT sensitive to functional form?

- In Roth and Sant’Anna (2022b), we tackle these questions.
- We derive necessary and sufficient conditions for the DiD to be insensitive to functional form restrictions.
- We show that these conditions are falsifiable, and propose tests for them.
- Unfortunately, we won’t have time to go into details today.

Moving away from 2x2

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 without covariates

We have a DiD course with 30h of DiD covering way more details.

Next cohort starts on March 2023.

www.causal-solutions.com

Student price (also applicable for developing countries): $295

Promotional Price: $595

Discount Code Available until Feb 10, 2023: CSEARLY

$50 off from Student price and $100 off from Promotional price

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?

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!**
- Goodman-Bacon (2021) provides the most popular explanation.

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?

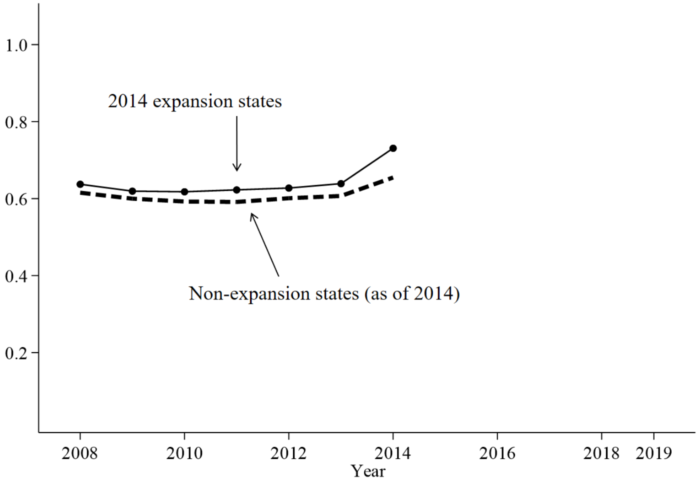
Example: Effect of ACA Medicaid Expansion on Health Insurance rate

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.

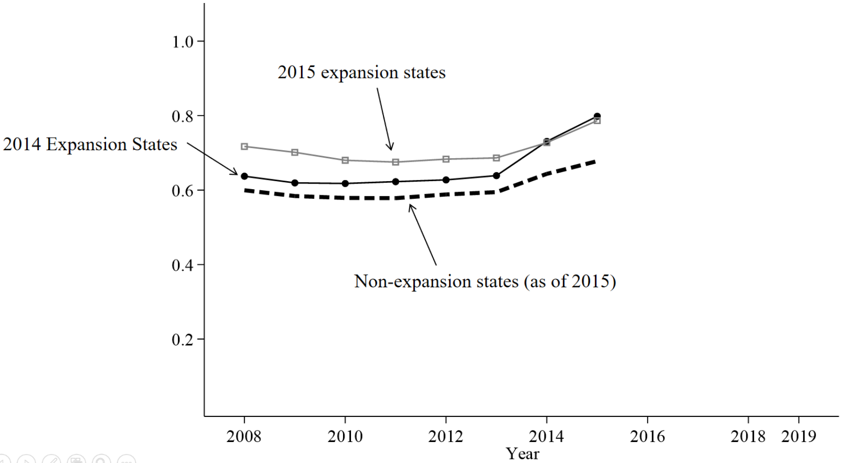
ACA Medicaid Expansion Circa 2014

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



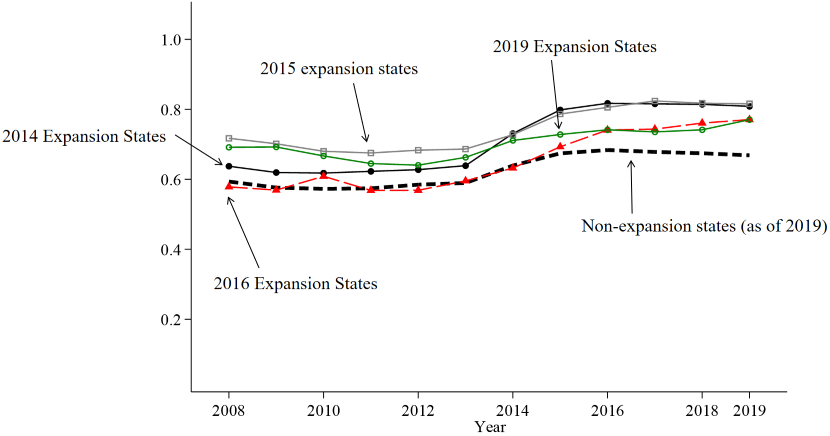
ACA Medicaid Expansion Circa 2015

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



ACA Medicaid Expansion Circa 2019

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:

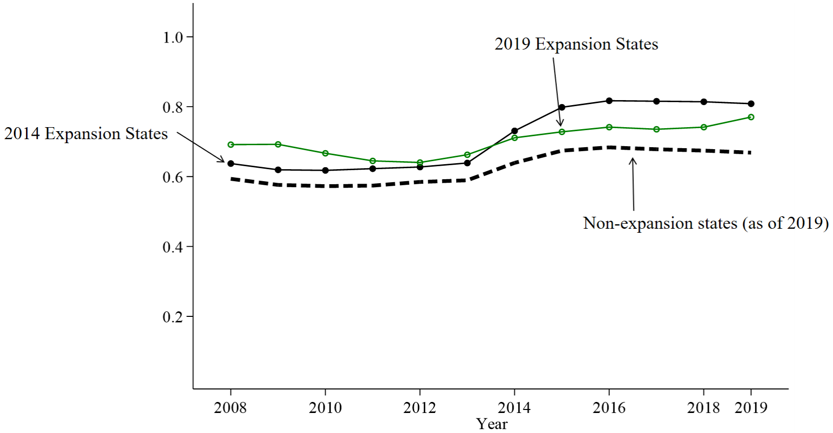
$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}$$

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

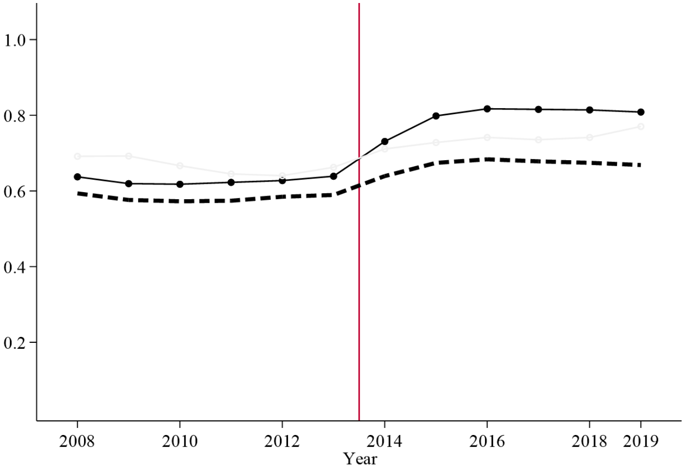
Three Groups Example

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



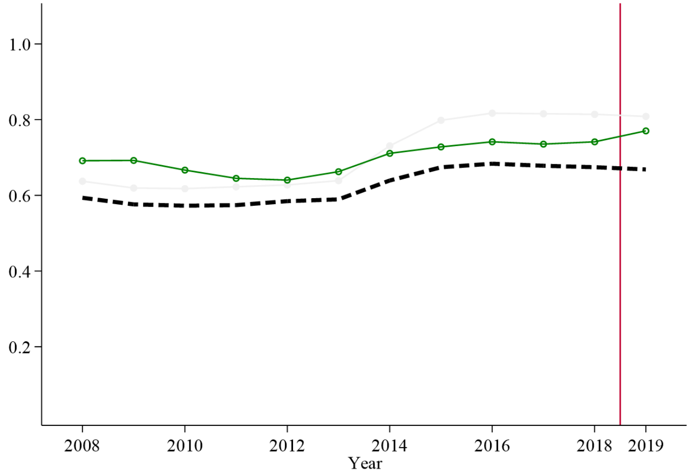
Treated in 2014 vs. Never-Treated

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



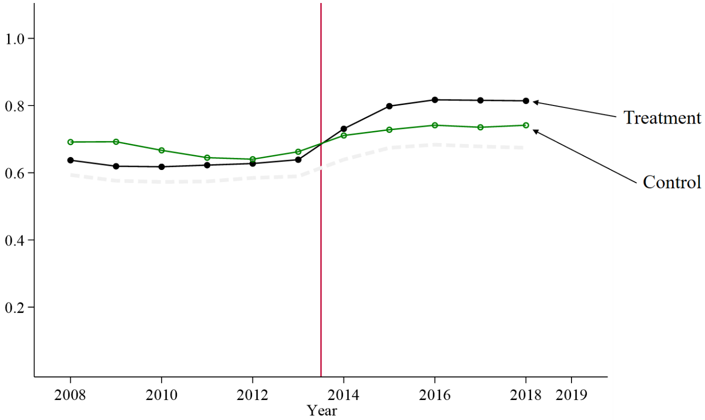
Treated in 2019 vs. Never-Treated

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



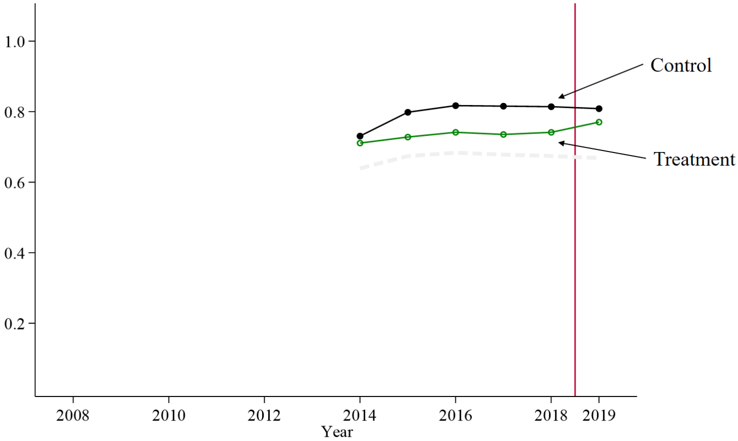
Treated in 2014 vs. Treated in 2019 ($t < 2019$)

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?

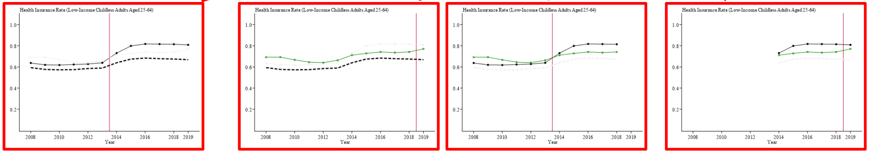
Bacon Decomposition

Bacon-Decomposition

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

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

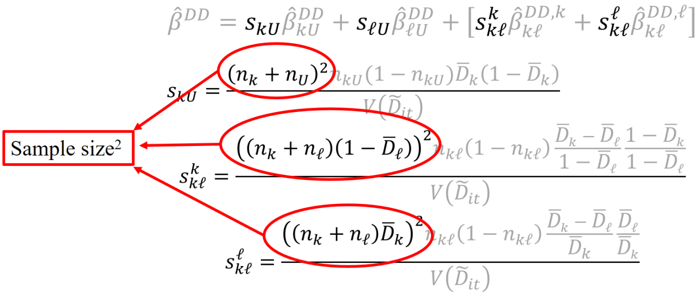
$$\hat{\beta}^{DD} = s_{kU}\boxed{\hat{\beta}^{DD}_{kU}} + s_{\ell U}\boxed{\hat{\beta}^{DD}_{\ell U}} + [s^{k}_{k\ell}\boxed{\hat{\beta}^{DD,k}_{k\ell}} + s^{\ell}_{k\ell}\boxed{\hat{\beta}^{DD,\ell}_{k\ell}}]$$



Bacon-Decomposition

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

Figure 10: Bacon-Decomposition: The weights



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

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

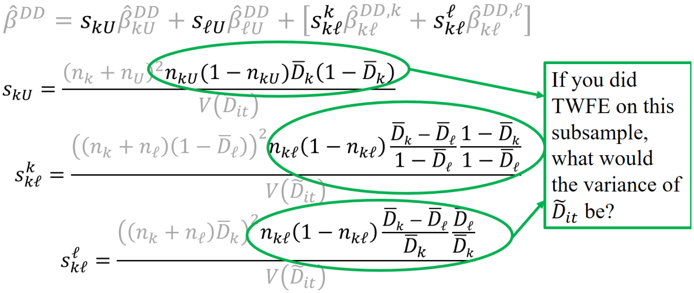
$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

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

Bacon-Decomposition

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

Figure 11: Bacon-Decomposition: The weights



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

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

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

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

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 $k = 1, \dots, K$ 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

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

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.

TWFE computes weighted-averages of 2x2 DiD's

- $\widehat{\beta} = 0.074$ in the empirical application.

- **OLS weights use sample size and variance**

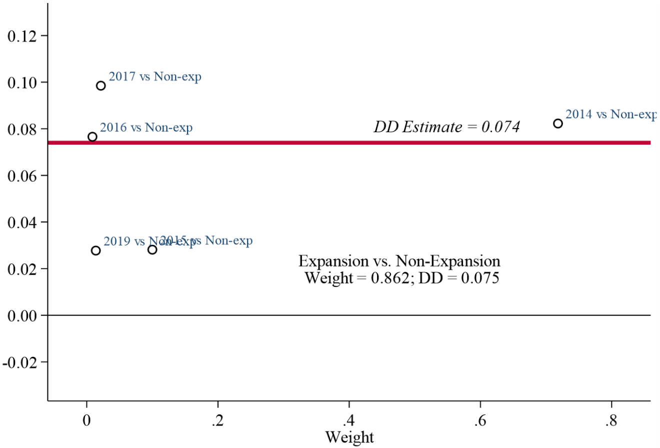
- Is that what you really want?

- **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 $\widehat{\beta}$?

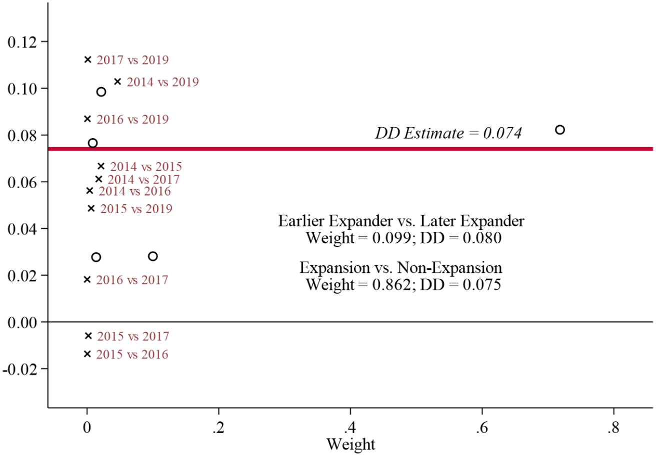
Bacon-Decomposition: Treated vs. Never-Treated

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

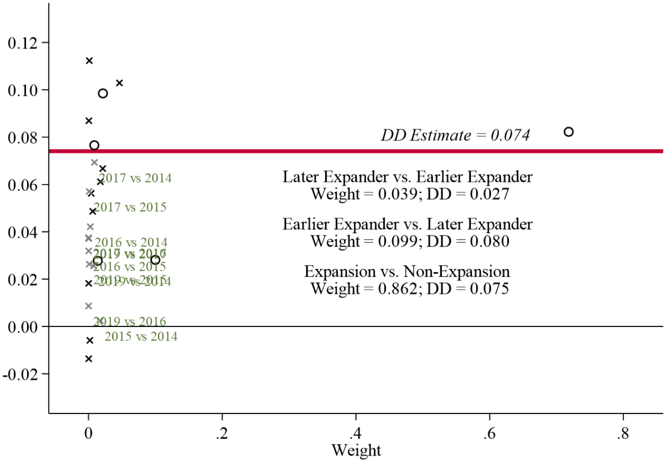


Figure 14: Bacon-Decomposition: The weights

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',

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

de Chaisemartin and D'Haultfœuille (2020) - Decomposition Results

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
- Intuition from Goodman-Bacon (2021): 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

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

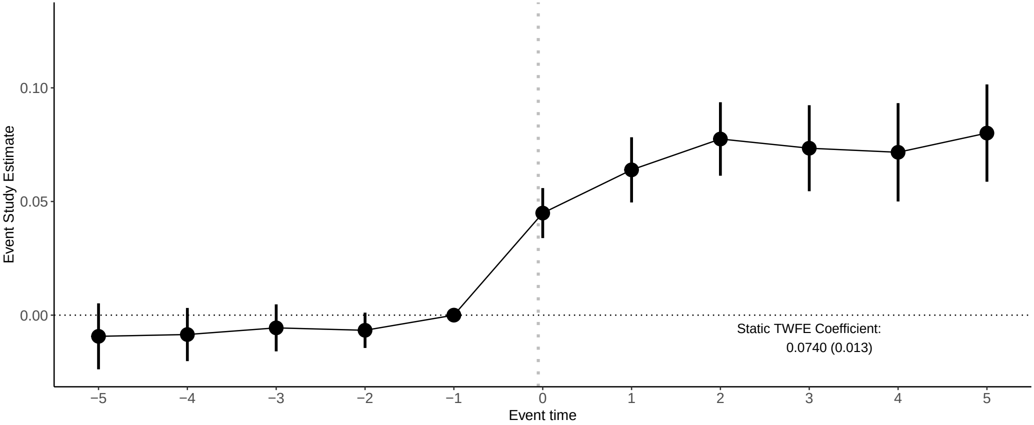
  with the event study dummies $D_{i,t}^{k} = 1\{t - G_i = k\}$, where G_i indicates the period unit i is first treated (Group).

- $D_{i,t}^{k}$ 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



Static TWFE Coefficient:
0.0740 (0.013)

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)

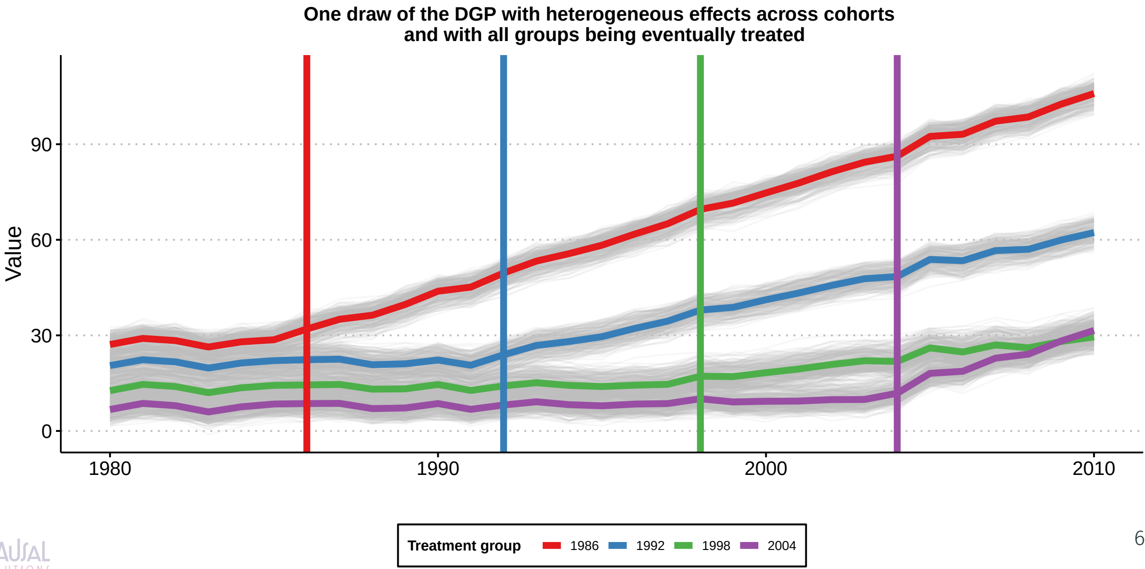
Problem with Event-Study via TWFE specifications: Sun and Abraham (2021)

- 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



One draw of the DGP with heterogeneous effects across cohorts
and with all groups being eventually treated

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.

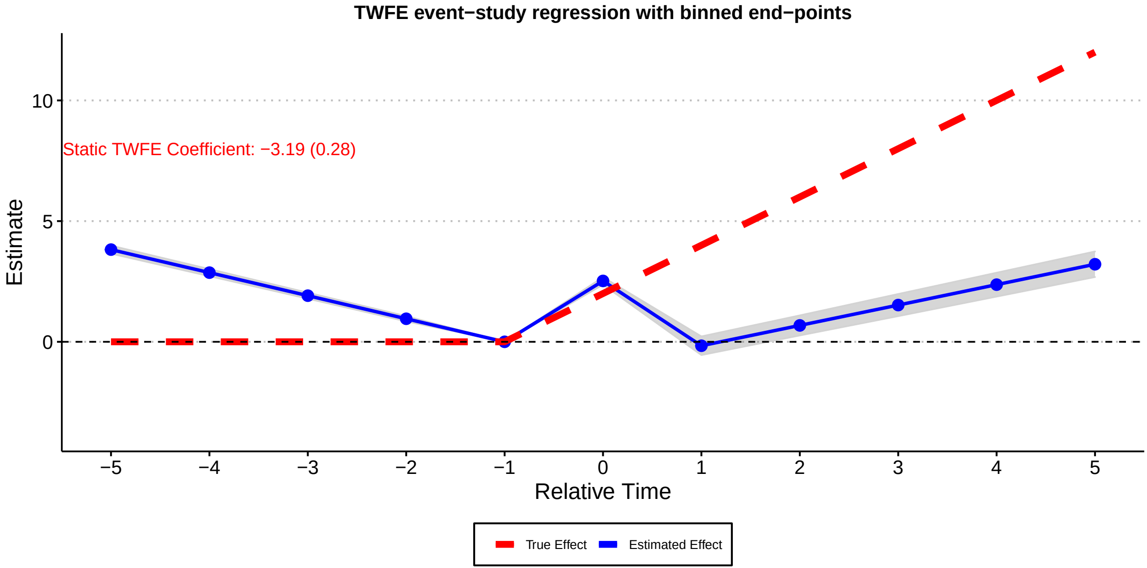
Traditional methods: TWFE event-study regression

- 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 ?

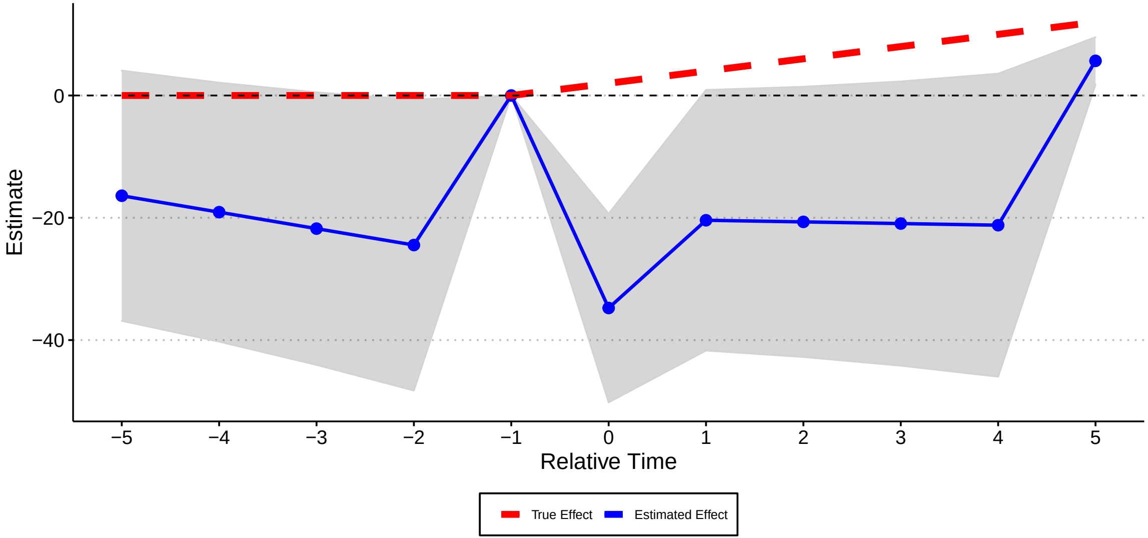
- Simulate data and repeat 1,000 times to compute bias and simulation standard deviations.


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

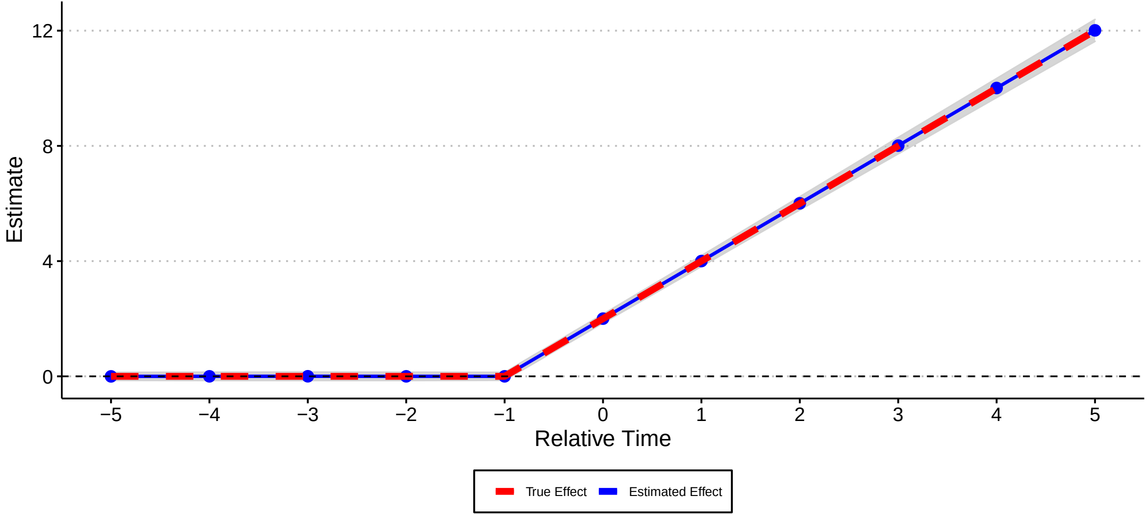
- What if we include all possible leads and lags in the TWFE event study specification, i.e., to set K and L to the maximum allowable in the data, making inclusion of $D_{i,t}^{<-K}$ and of $D_{i,t}^{>L}$ unnecessary ?

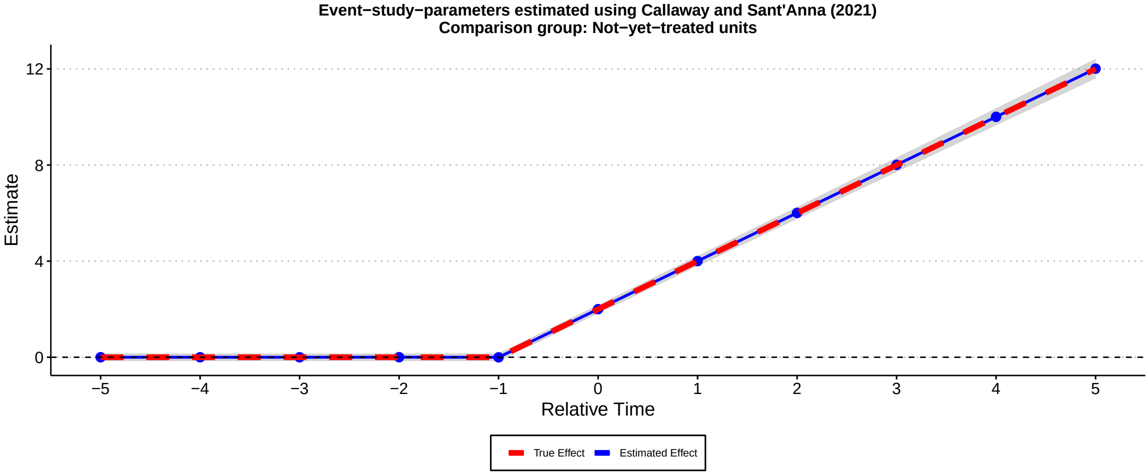
TWFE event−study regression with 'all' leads and lags



Is there hope?

Event−study−parameters estimated using Callaway and Sant'Anna (2021)
Comparison group: Last−treated−Cohort units




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!

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?

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, 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.

Main take-way message from all DiD proposed solutions

Clearly separate identification, aggregation, and estimation/inference steps!

In what follows, I will focus on Callaway and Sant'Anna (2021)'s approach.

Digging into Callaway and Sant’Anna (2021)

Callaway and Sant'Anna (2021) in practice

Can be implemented via the R package did.

Can be implemented via the Stata package csdid.

Can be implemented via the Python package differences.

Identification

Let’s talk about identification

Identification

Building block of the analysis

- If sample size was not a limitation (we have all the data in the world), what kind of question we would like to answer?

- In staggered setups, a parameter that is interesting and has clear economic interpretation is the $ATT(g,t)$

$$ATT\left(g,t\right) = \mathbb{E}\left[Y_t\left(g\right) - Y_t\left(\infty\right)\left|G_g = 1\right.\right], \ for\ t \geq g.$$

- Average Treatment Effect at time t of starting treatment at time g, among the units that indeed started treatment at time g.

Identifying Assumptions: No-Anticipation

- Given that we never observe $Y(\infty)$ in post-treatment periods among units that have been treated, we need to make assumptions to identify $ATT(g,t)$'s
- **No-Anticipation Assumption**: For all i,t and $t < g, g'$, $Y_{i,t}(g) = Y_{i,t}(g')$.
- Unit treatment effects are zero before treatment takes place.
- Exactly the same content as in the 2x2 case.

Parallel trend assumption based on a “never treated” group

Assumption (Parallel Trends based on a “never-treated”)

For each $t \in \{2, \ldots, T\}$, $g \in \mathcal{G}$ such that $t \geq g$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | C = 1]$$

Parallel Trends based on not-yet treated groups

Assumption (Parallel Trends based on “Not-Yet-Treated” Groups)

For each $(s, t) \in \{2, \dots, T\} \times \{2, \dots, T\}$, $g \in \mathcal{G}$ *such that* $t \geq g$, $s \geq t$

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | D_s = 0, G_g = 0].$$

ATT(g,t) Estimand: “never-treated” as comparison group

- Under no-anticipation and PT based on “never-treated”, we have

$$ATT_{unc}^{nev}(g,t) = \mathbb{E}[Y_t - Y_{g-1} | G_g = 1] - \mathbb{E}[Y_t - Y_{g-1} | C = 1].$$

- This looks very similar to the two periods, two-groups DiD result without covariates.

- The difference is now we take a “long difference”.

- Same intuition carries, though!

- This result appears in Callaway and Sant’Anna (2021) and Sun and Abraham (2021).

ATT(g,t) Estimand: not-yet treated as comparison group

- If one wants to use an the units that have not-yet been exposed to treatment by time t, we have a different estimand:

$$ATT_{unc}^{ny}(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|D_t = 0, G_g = 0].$$

- This looks similar to the two periods, two-groups DiD result without covariates, too.
- The difference is now we take a “long difference” , and that the comparison group changes over time.
- Same intuition carries, though!
- This result appears in Callaway and Sant’Anna (2021) and de Chaisemartin and D’Haultfœuille (2020), though de Chaisemartin and D’Haultfœuille (2020) focus exclusively in instantaneous treatment effects, i.e., the case with $g = t$.

Including covariates to “relax” PT

In practice, we can incorporate covariates into DiD to “relax” the PT assumption: it can allow for covariate-specific trends.

Callaway and Sant’Anna (2021) provide extensive treatment of how to do this reliably.

Can use regression, weighting, or doubly-robust methods

We do not have time to cover this in detail, today.

Aggregation

Second step: Aggregation

Summarizing ATT(g,t)

- $ATT(g,t)$ are very useful parameters that allow us to better understand treatment effect heterogeneity.

- We can also use these to summarize the treatment effects across groups, time since treatment, and calendar time.

- Practitioners routinely attempt to pursue this avenue:
    - Run a TWFE “static” regression and focus on the β associated with the treatment.
    - Run a TWFE event-study regression and focus on β associated with the treatment leads and lags.
    - Collapse data into a 2 x 2 Design (average pre and post-treatment periods).

Summarizing ATT(g,t)

- We propose taking weighted averages of the $ATT(g,t)$ of the form:

$$\sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \le t\} w_{gt} ATT(g,t)$$

- The two simplest ways of combining $ATT(g,t)$ across g and t are, assuming no-anticipation,

$$\theta_M^O := \frac{2}{T(T-1)} \sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \le t\} ATT(g,t) \tag{1}$$

and

$$\theta_W^O := \frac{1}{\kappa} \sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \le t\} ATT(g,t) P(G = g | C \ne 1) \tag{2}$$

- Problem: They "overweight" units that have been treated earlier

Summarizing ATT(g,t): Cohort-heterogeneity

- More empirically motivated aggregations do exist!
- Average effect of participating in the treatment that units in group *g* experienced:

$$\theta_S(g) = \frac{1}{T - g + 1} \sum_{t=2}^{T} \mathbf{1}\{g \leq t\} ATT(g, t)$$

Summarizing ATT(g,t): Calendar time heterogeneity

- Average effect of participating in the treatment in time period t for groups that have participated in the treatment by time period t

$$\theta_C(t) = \sum_{g=2}^{T} 1\{g \leq t\} ATT(g,t) P(G = g | G \leq t, C \neq 1)$$

Summarizing ATT(g,t): Event-study / dynamic treatment effects

- The effect of a policy intervention may depend on the length of exposure to it.

- Average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods

$$\theta_D(e) = \sum_{g=2}^{T} \mathbf{1}\{g + e \leq T\} ATT(g, g+e) P(G = g | G + e \leq T, C \neq 1)$$

- This is perhaps the most popular summary measure currently adopted by empiricists.

Summarizing ATT(g,t): Event-study

- When we compare $\theta_D(e)$ across two relative times e_1 and e_2, we have that

$$
\begin{aligned}
&\theta_D(e_2) - \theta_D(e_1) \\
&= \sum_{g=2}^{T} \mathbf{1}\{g + e_1 \leq T\} \underbrace{\left(ATT(g, g+e_2) - ATT(g, g+e_1)\right)}_{\text{dynamic effect for group } g} P(G = g | G + e_1 \leq T) \\
&\quad + \sum_{g=2}^{T} \mathbf{1}\{g + e_2 \leq T\} ATT(g, g+e_2) \underbrace{\left(P(G = g | G + e_2 \leq T) - P(G = g | G + e_1 \leq T)\right)}_{\text{differences in weights}} \\
&\quad - \sum_{g=2}^{T} \underbrace{\mathbf{1}\{T - e_2 \leq g \leq T - e_1\}}_{\text{different composition of groups}} ATT(g, g+e_2) P(G = g | G + e_2 \leq T)
\end{aligned}
$$

- **Balance sample in “event time” to avoid compositional changes that complicate comparisons across e.**

Third step: Estimation and Inference

Estimation and Inference

Estimation

- Identification results suggest a simple plug-in estimation procedure.
- Callaway and Sant'Anna (2021) provides high-level conditions for one to consider more general first-step estimators that allows for covariates and some flexible "data-adaptive" (machine learning) procedures.
  - Similar to Chen, Linton and Van Keilegom (2003) and Chen, Hong and Tarozzi (2008)

Inference

- Under relatively weak regularity conditions,

$$\sqrt{n}\left(\widehat{ATT}(g,t) - ATT(g,t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{gt}(\mathcal{W}_i) + o_p(1)$$

- From the above asymptotic linear representation and a CLT, we have

$$\sqrt{n}\left(\widehat{ATT}(g,t) - ATT(g,t)\right) \xrightarrow{d} N(0, \Sigma_{g,t})$$

  where $\Sigma_{gt} = \mathbb{E}[\psi_{gt}(\mathcal{W})\psi_{gt}(\mathcal{W})']$.

- Above result ignores the dependence across g and t, and “multiple-testing” problems.

Simultaneous Inference

- Let’s simplify and ignore anticipation issues for the moment.

Simultaneous Inference

- Let's simplify and ignore anticipation issues for the moment.
- Let $ATT_{g\leq t}$ and $\widehat{ATT}_{g\leq t}$ denote the vector of $ATT(g,t)$ and $\widehat{ATT}(g,t)$, respectively, for all $g=2,\ldots,T$ and $t=2,\ldots,T$ with $g\leq t$.

Simultaneous Inference

- Let's simplify and ignore anticipation issues for the moment.
- Let $ATT_{g \leq t}$ and $\widehat{ATT}_{g \leq t}$ denote the vector of $ATT(g, t)$ and $\widehat{ATT}(g, t)$, respectively, for all $g = 2, \ldots, T$ and $t = 2, \ldots, T$ with $g \leq t$.
- Analogously, let $\Psi_{g \leq t}$ denote the collection of ψ_{gt} across all periods t and groups g such that $g \leq t$.

Simultaneous Inference

- Let's simplify and ignore anticipation issues for the moment.
- Let $ATT_{g\leq t}$ and $\widehat{ATT}_{g\leq t}$ denote the vector of $ATT(g,t)$ and $\widehat{ATT}(g,t)$, respectively, for all $g=2,\ldots,T$ and $t=2,\ldots,T$ with $g\leq t$.
- Analogously, let $\Psi_{g\leq t}$ denote the collection of ψ_{gt} across all periods t and groups g such that $g\leq t$.
- Hence, we have
$$\sqrt{n}(\widehat{ATT}_{g\leq t}-ATT_{g\leq t})\xrightarrow{d}N(0,\Sigma)$$
where
$$\Sigma=\mathbb{E}[\Psi_{g\leq t}(\mathcal{W})\Psi_{g\leq t}(\mathcal{W})'].$$

Simultaneous confidence intervals

- How to construct simultaneous confidence intervals?

- We propose the use of a simple multiplier bootstrap procedure.

- Let $\widehat{\Psi}_{g\leq t}(\mathcal{W})$ denote the sample-analogue of $\Psi_{g\leq t}(\mathcal{W})$.

- Let $\{V_i\}_{i=1}^{n}$ be a sequence of *iid* random variables with zero mean, unit variance and bounded third moment, independent of the original sample $\{\mathcal{W}_i\}_{i=1}^{n}$

- $\widehat{ATT}^{*}_{g\leq t}$, a bootstrap draw of $\widehat{ATT}_{g\leq t}$, via

$$\widehat{ATT}^{*}_{g\leq t} = \widehat{ATT}_{g\leq t} + \mathbb{E}_n\left[V \cdot \widehat{\Psi}_{g\leq t}(\mathcal{W})\right]. \tag{3}$$

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^n$.

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^n$.
2. Compute $\widehat{ATT}^*_{g\leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^*(g,t)$, and form a bootstrap draw of its limiting distribution as

$$\hat{R}^*(g,t) = \sqrt{n}\left(\widehat{ATT}^*(g,t) - \widehat{ATT}(g,t)\right)$$

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{ATT}^{*}_{g \leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^{*}(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^{*}(g,t) = \sqrt{n}\left(\widehat{ATT}^{*}(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{ATT}^*_{g\leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^*(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^*(g,t) = \sqrt{n}\left(\widehat{ATT}^*(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.
4. Estimate $\Sigma^{1/2}(g,t)$ by
$$\widehat{\Sigma}^{1/2}(g,t) = \left(q_{0.75}(g,t) - q_{0.25}(g,t)\right) / \left(z_{0.75} - z_{0.25}\right)$$

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{ATT}^{*}_{g\leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^{*}(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^{*}(g,t) = \sqrt{n}\left(\widehat{ATT}^{*}(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.
4. Estimate $\Sigma^{1/2}(g,t)$ by
$$\widehat{\Sigma}^{1/2}(g,t) = \left(q_{0.75}(g,t) - q_{0.25}(g,t)\right) / \left(z_{0.75} - z_{0.25}\right)$$
5. For each bootstrap draw, compute $t-test^{*}_{g\leq t} = \max_{(g,t)}\left|\hat{R}^{*}(g,t)\right|\widehat{\Sigma}(g,t)^{-1/2}$.

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{ATT}^{*}_{g \leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^{*}(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^{*}(g,t) = \sqrt{n}\left(\widehat{ATT}^{*}(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.
4. Estimate $\Sigma^{1/2}(g,t)$ by
$$\widehat{\Sigma}^{1/2}(g,t) = \left(q_{0.75}(g,t) - q_{0.25}(g,t)\right) / \left(z_{0.75} - z_{0.25}\right)$$
5. For each bootstrap draw, compute $t-test^{*}_{g \leq t} = \max_{(g,t)} \left|\hat{R}^{*}(g,t)\right| \widehat{\Sigma}(g,t)^{-1/2}$.
6. Construct $\widehat{c}_{1-\alpha}$ as the empirical $(1-a)$-quantile of the B bootstrap draws of $t-test^{*}_{g \leq t}$.

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{ATT}_{g\leq t}^{*}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^{*}(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^{*}(g,t) = \sqrt{n}\left(\widehat{ATT}^{*}(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.
4. Estimate $\Sigma^{1/2}(g,t)$ by
$$\widehat{\Sigma}^{1/2}(g,t) = (q_{0.75}(g,t) - q_{0.25}(g,t)) \,/\, (z_{0.75} - z_{0.25})$$
5. For each bootstrap draw, compute $t - test_{g\leq t}^{*} = \max_{(g,t)} \left|\hat{R}^{*}(g,t)\right| \widehat{\Sigma}(g,t)^{-1/2}$.
6. Construct $\widehat{c}_{1-\alpha}$ as the empirical $(1-a)$-quantile of the B bootstrap draws of $t - test_{g\leq t}^{*}$.
7. Construct the bootstrapped simultaneous confidence intervals for $ATT(g,t)$, $g \leq t$, as
$$\widehat{C}(g,t) = [\widehat{ATT}(g,t) \pm \widehat{c}_{1-\alpha} \cdot \widehat{\Sigma}(g,t)^{-1/2} \,/\, \sqrt{n}].$$

Simultaneous cluster-robust confidence intervals

- Sometimes one wishes to account for clustering.
- This is straightforward to implement with the multiplier bootstrap described above.
- Example: allow for clustering at the state level
  - draw a scalar U_s S times – where S is the number of states
  - set $V_i = U_s$ for all observations i in state s
- This procedure is justified provided that the number of clusters is “large”.

ACA Medicaid Expansion Example

Let’s go back to the ACA Medicaid Expansion Example

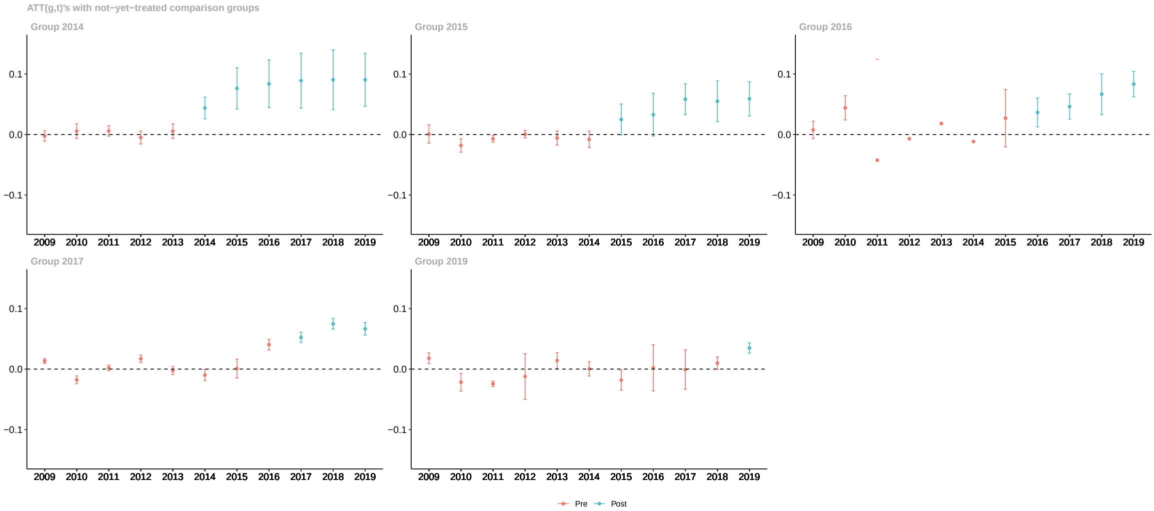
Let’s open R again

ACA Medicaid Expansion

- 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

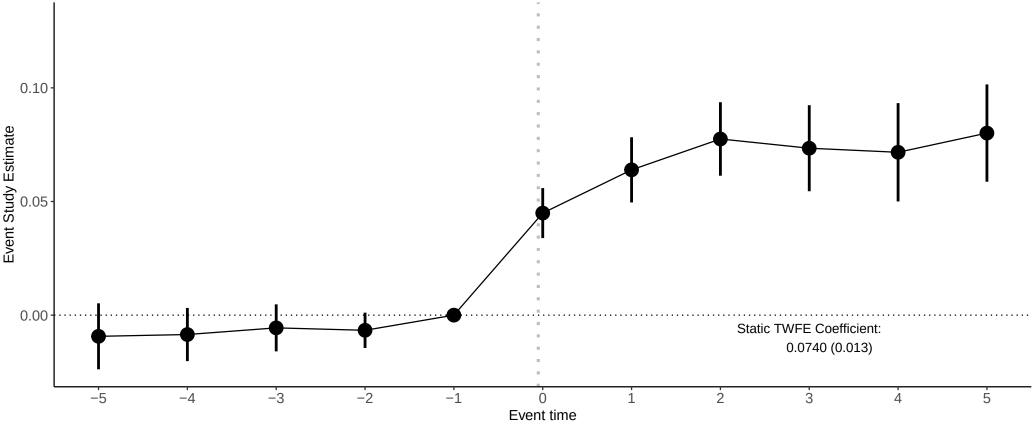
Challenge setup to make inference on ATT(g,t)'s per se

ACA Medicaid Expansion: Not-yet-treated as comparison group



ACA Medicaid Expansion: TWFE Event-study specification

Figure 16: Health Insurance Rate (low-income Childless Adults Aged 25-64)



ACA Medicaid Expansion: CS Event-study specification

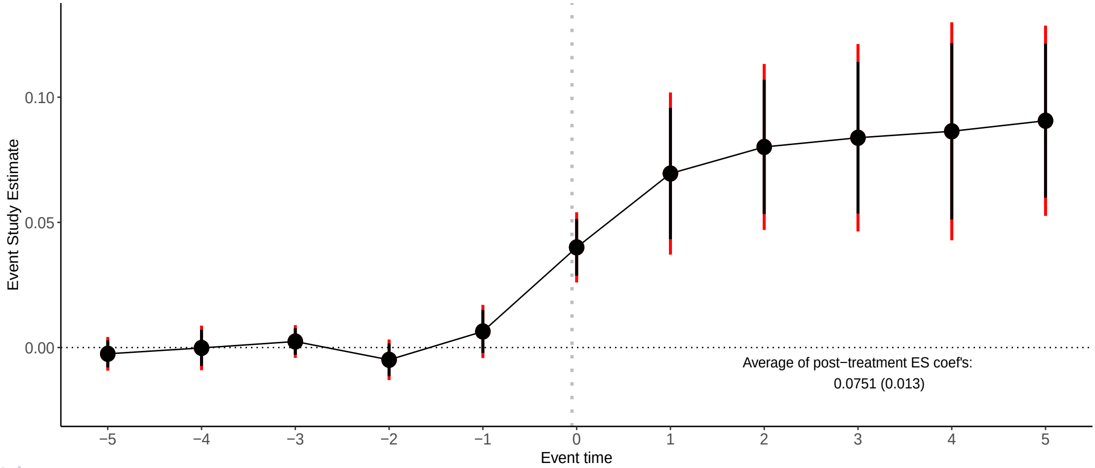
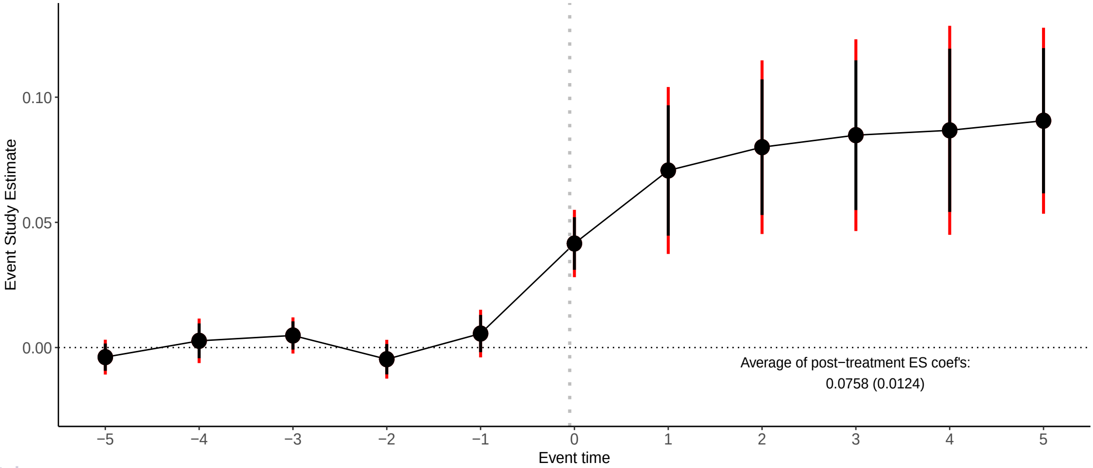


Figure 17: Results using “never-treated” as a comparison group

ACA Medicaid Expansion: CS Event-study specification

Figure 18: Results using “not-yet-treated” as comparison groups



Take-way messages

DiD procedures multiple time periods

- With multiple time periods and variation in treatment timing, TWFE does not respect our assumptions:
  - OLS is “variational hungry” and makes many comparisons of means
  - Some of these comparisons are bad: use already-treated units as a comparison group to “later-treated” groups
  - This can lead to “negative weighting” problems.
- Solution to the TWFE problem is simple
  - Separate the identification, aggregation and estimation/inference parts of the problem
- Use $ATT(g, t)$ as a building block so we can transparently see how things are constructed
- Many different aggregation schemes are possible: they deliver different parameters!
- Can allow for covariates via regressions adjustments, IPW and DR.

If you want learn more about DiD, check our course!

CAUSAL
SOLUTIONS

www.causal-solutions.com

CAUSAL SOLUTIONS

Student price (also applicable for developing countries): $295

Promotional Price: $595

Discount Code Available until Feb 10, 2023: CSEARLY

$50 off from Student price and $100 off from Promotional price

For any questions/comments, you can contact me via

✉pedrosantanna@causal-solutions.com

✉pedro.h.santanna@vanderbilt.edu

✉psantanna@microsoft.com

🐦@pedrohcgs

Bonus Topic: Violations of Parallel Trends

Recent Boom of New DiD Methods: Violations of PT

- What if treatment Parallel Trends Assumption is violated?

Recent Boom of New DiD Methods: Violations of PT

- What if treatment Parallel Trends Assumption is violated?

- **Rambachan and Roth (2022):** Shows how you can use pre-trends to bound ATT's when PT are violated.

- Build on Manski and Pepper (2015) but provide new and practically relevant uniformly valid inference procedures. New rationale for violations of PT, too!

- Can be easily combined with Callaway and Sant'Anna (2021) - *https://github.com/pedrohcgs/CS_RR*.

- **This is my favorite paper of this "batch" of new DiD papers.**

Why do I like this paper so much?

- Currently common practice on pre-test has limitations with important practical consequences.

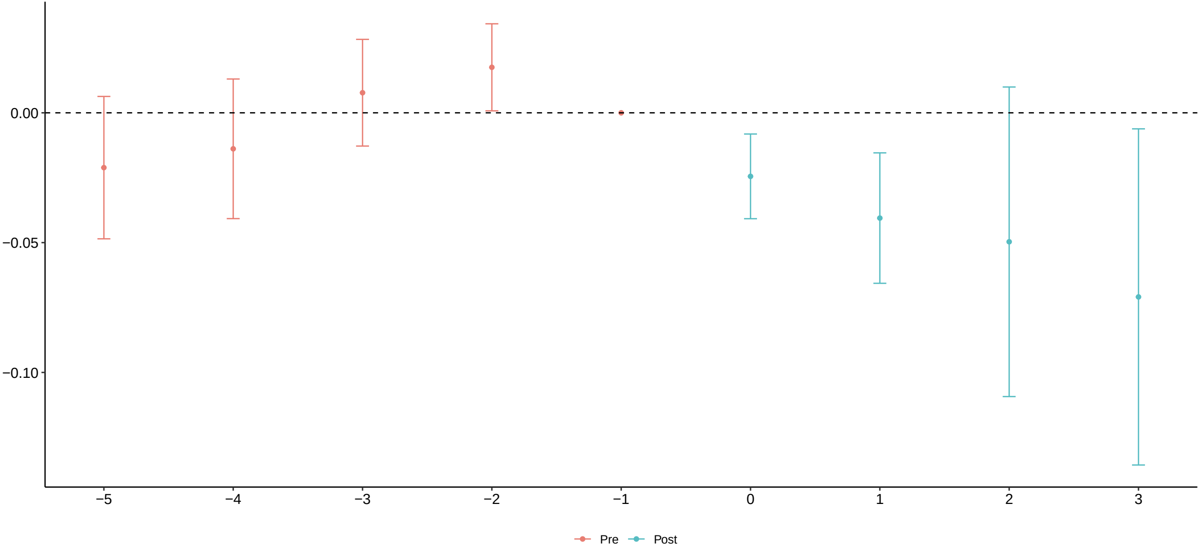
- However, as a good econometrician, instead of sitting in our Ivory Tower, we need to seek several practical, easy-to-use tools that can alleviate some of these problems.

- This is what Rambachan and Roth (2022) do!
- In my view, the sensitivity analysis procedures in Rambachan and Roth (2022) are fundamental to improving the reliability and transparency of DiD procedures.

- Let's briefly show this using the did and HonestDiD R packages, which implements Callaway and Sant'Anna (2021) and Rambachan and Roth (2022), respectively.

Combining Callaway and Sant’Anna (2021) and Rambachan and Roth (2021)

```
# Install the packages (I used the Github versions)
devtools::install_github("bcallaway11/did");
devtools::install_github("asheshrambachan/HonestDiD")

#Load the packages
library(did); library(HonestDiD); library(dplyr); library(here)

# Load data used by Callaway and Sant'Anna (2021)
min_wage <- readRDS((here("data",'min_wage_CS.rds')))

#--------------------------------------------------------------------------------------------
# Formula for covariates
xformla <- ~ region + (medinc + pop ) + I(pop^2) + I(medinc^2)  + white + hs  + pov
#--------------------------------------------------------------------------------------------

# Estimate ATT(g,t)'s using DR DiD with never-treated as comparison group
CS_never_cond <- did::att_gt(yname="lemp", tname="year", idname="countyreal", gname="first.treat",
                             xformla = xformla, control_group="nevertreated", data = min_wage,
                             panel = TRUE, base_period="universal", bstrap = TRUE, cband = TRUE)

# compute event-study aggregation
CS_es_never_cond <- aggte(CS_never_cond, type = "dynamic", min_e = -5, max_e = 5)
ggdid(CS_es_never_cond,
      title = "Event-study aggregation \n DiD based on conditional PTA and using never-treated as comparison group ")
```

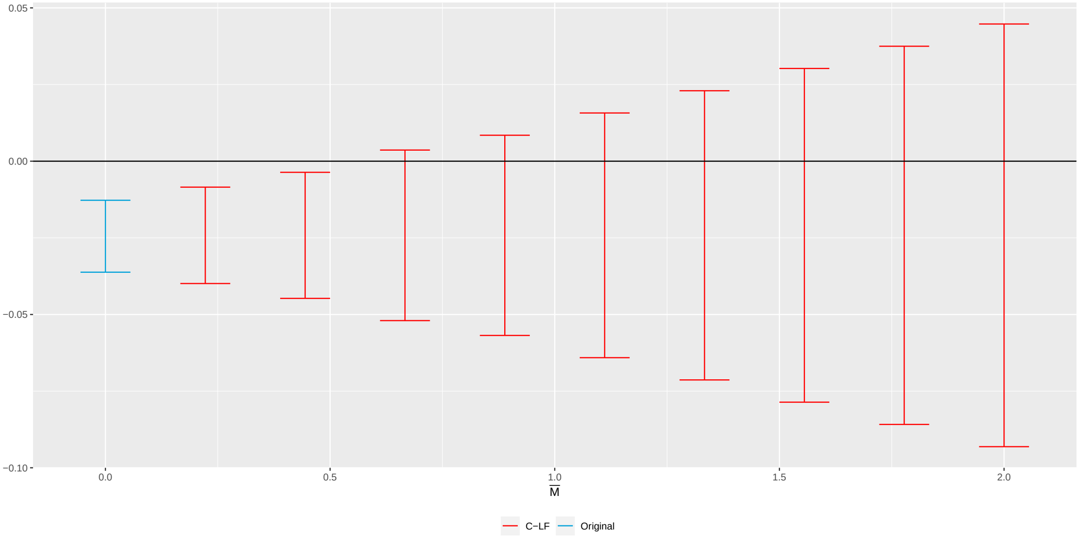


Event−study aggregation
DiD based on conditional PTA and using never−treated as comparison group

Rambachan and Roth (2022) after Callaway and Sant'Anna (2021)

```
# Brant and I have written a wrapper for HonestDiD that allows one to use aggte did outputs as inputs

# Here we apply the wrapper, and use the ``relative magnitude'' type of sensitivity analysis

# Doing it for instantaneous treatment effect, e = 0
hd_cs_rm_never <- honest_did(es = CS_es_never_cond,
                             e = 0,
                             type="relative_magnitude")

# Plot results
cs_HDiD_relmag <- createSensitivityPlot_relativeMagnitudes(hd_cs_rm_never$robust_ci,
                                         hd_cs_rm_never$orig_ci)
cs_HDiD_relmag
```

Sensitivity Analysis based on “relative magnitude” restrictions



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.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, apr 2008, *36* (2), 808–843.

— , **Oliver Linton, and Ingrid Van Keilegom**, “Estimation of semiparametric models when the criterion function is not smooth,” *Econometrica*, 2003, *71* (5), 1591–1608.

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).

Manski, Charles F and John V Pepper, "How Do Right-To-Carry Laws Affect Crime Rates? Coping With Ambiguity Using Bounded-Variation Assumptions," Working Paper 21701, National Bureau of Economic Research November 2015. Series: Working Paper Series.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 2022, *Forthcoming.*

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

— and — , "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.