Recent Advances in DiD Methods

A selective (and personal) perspective

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

Guest Lecture at University of Washington, January 2022

Popularity of Difference-in-Differences methods

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

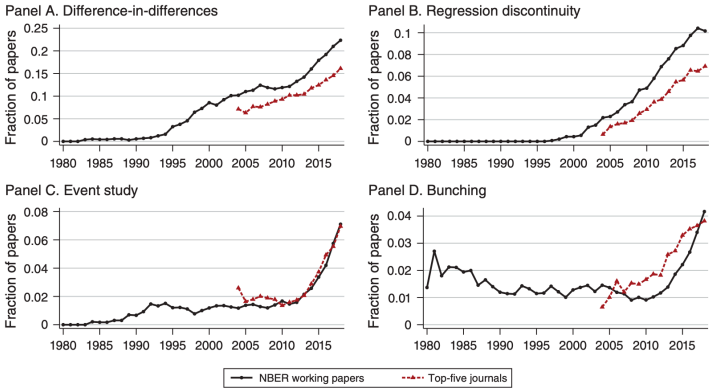


FIGURE 4. QUASI-EXPERIMENTAL METHODS

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.

Structure of the Lectures

- Although DiD is a very applied topic, my two lectures will be somehow very methodological.
- My main goals (probably too ambitious):
  1. Expose everyone about the canonical DiD setup.
  2. Very briefly introduce several research active areas in DiD:
     - 2.1 Role of covariates (link to modern machine-learning econometrics/stats literature)
     - 2.2 DiD setups with variation in treatment timing (problems and solutions)
     - 2.3 Non-Binary treatments
     - 2.4 Potential violations of PT (and how to do sensitivity analysis)
     - 2.5 When is DiD sensitive to functional form assumptions?
     - 2.6 Inference with few clusters (very brief)
  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

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

- 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:

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

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

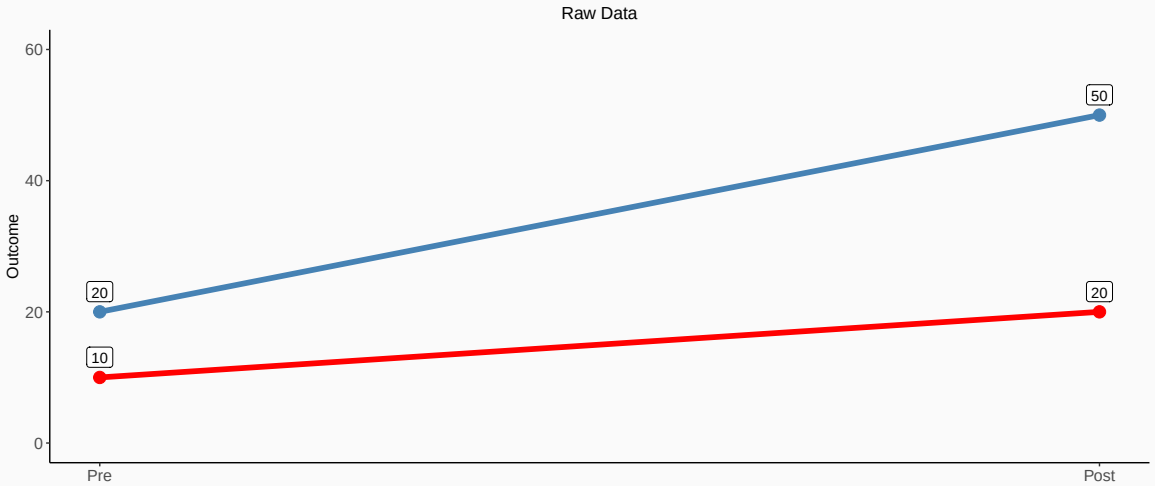
Parallel Trends and the ATT

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

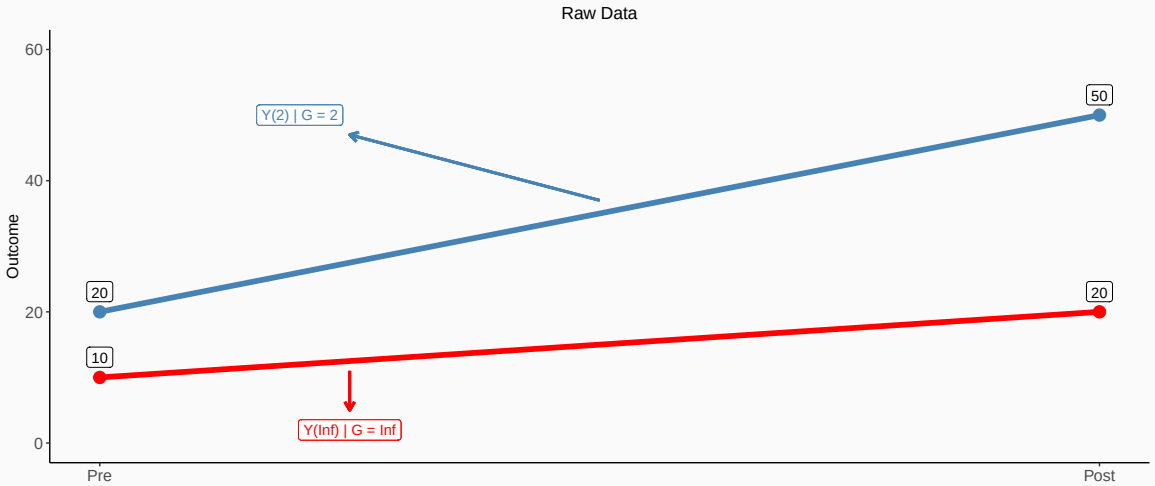
$$
\begin{aligned}
\textbf{ATT} &\equiv \mathbb{E}\left[Y_{i,t=2}\left(2\right)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\
&= \mathbb{E}\left[Y_{i,t=2}|G_i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

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

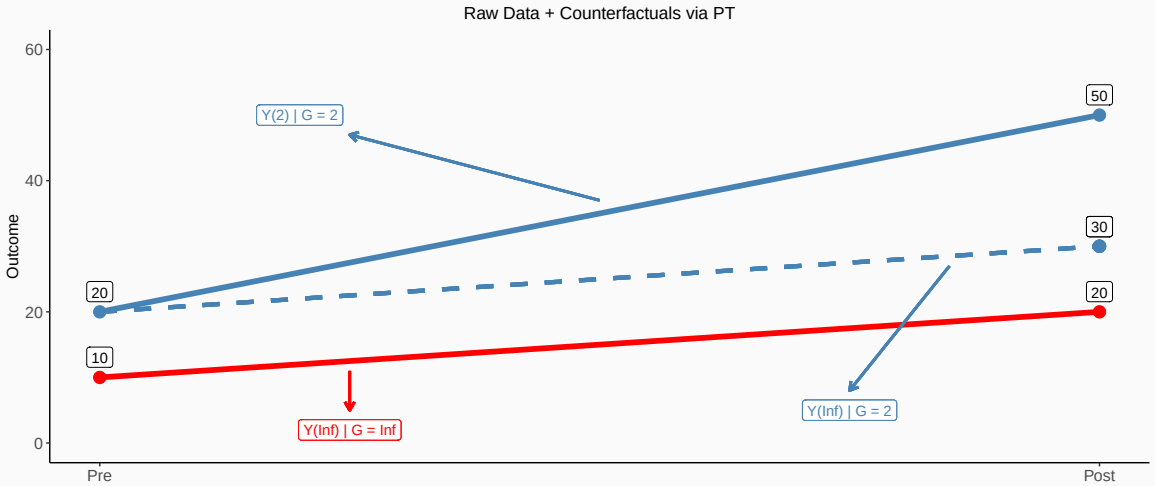
Parallel Trends via graphs



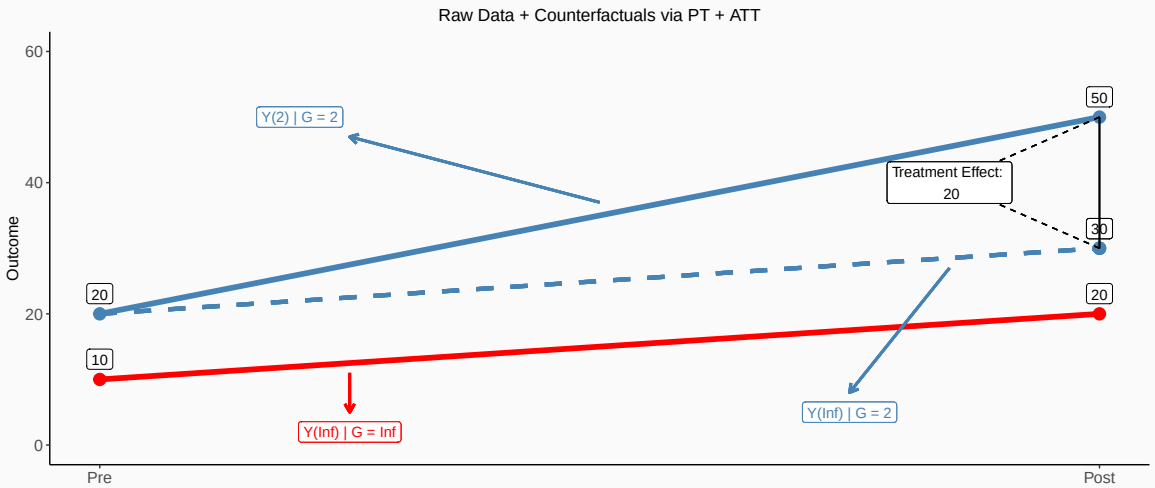
Parallel Trends via graphs



Parallel Trends via graphs



Parallel Trends via graphs



Under the already invoked assumptions, can we identify the ATE?

Why?

ATT vs. ATE vs. ATU

- So far we have focused on the Average Treatment Effect among Treated units

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

- Our assumptions allowed us to identify $\mathbb{E}\left[Y_{t=2}\left(\infty\right)|G=2\right]$.

- Our assumptions, however, does not allow us to identify the Average Treatment Effect among Untreated units

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

  because they do not allow us to identify $\mathbb{E}\left[Y_{t=2}\left(2\right)|G=\infty\right]$.

- Thus, unless we impose additional assumptions, we cannot identify the ATE:

$$ATE = ATT \times \mathbb{P}(G=2) + ATU \times \mathbb{P}(G=\infty).$$

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?

“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 estimators asymptotic linear representation (influence function), but not many people like that.

⏩ Show all derivations using “brute force” DiD estimator

‘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;
  - Treatment turn on and off;
  - Non-binary treatments;
  - Parallel trends may not hold exactly.
  - Only few treated and untreated clusters are available

Let’s groups these advances in “big themes”

See **Roth, Sant’Anna, Bilinski and Poe (2021)** for more details

What is the role played by covariates in DiD setups?

Recent Boom of New DiD Methods: Some big themes

- What is the role played by covariates in DiD setups?

Recent Boom of New DiD Methods: Some big themes

- What is the role played by covariates in DiD setups?
  - Three different modelling approaches available: Outcome Regression (Heckman et al., 1997), IPW (Abadie, 2005) and doubly robust (**Sant'Anna and Zhao, 2020**).
  - **Sant'Anna and Zhao (2020)** also derive semiparametric efficiency bounds for the 2*x*2 setup with covariates.
  - Adding covariates linearly into the TWFE will not give you the *ATT*.
  - **Sant'Anna and Zhao (2020)** also discuss the efficiency loss of observing repeated cross-sections instead of balanced panel data.
  - **Chang (2020)** discusses how to use data-adaptive (aka machine learning) procedures in DiD setups using doubly robust methods

Role of Covariates in 2x2 DiD Setups

- Let's give a bit more details on these interesting questions.

- Outcome regression estimator a la Heckman et al. (1997):

$$\widehat{ATT}^{reg,p} = \bar{Y}_{g=2,t=2} - \left[\bar{Y}_{g=2,t=1} + n_{treat}^{-1} \sum_{i|G_i=2} \left(\widehat{m}_{t=2}^{G=\infty}(X_i) - \widehat{m}_{t=1}^{G=\infty}(X_i) \right) \right].$$

- IPW estimator a la Abadie (2005):

$$\widehat{ATT}^{ipw,p} = \frac{1}{\mathbb{E}_n\left[1\{G_i = 2\}\right]} \cdot \mathbb{E}_n \left[\frac{1\{G_i = 2\} - \widehat{p}(X)}{1 - \widehat{p}(X)} \left(Y_{t=2} - Y_{t=1} \right) \right],$$


OUTCOME REGRESSION
PROPENSITY SCORE
JAKE-CLARK.TUMBLR
imgflip.com

Doubly Robust DiD Procedures

- **Sant'Anna and Zhao (2020):** Combine both outcome regression and IPW approaches to form more robust estimators.

- Estimators are **Doubly Robust consistent**: they are consistent for the ATT if either (but not necessarily both)
  - Regression working models for outcome dynamics are correctly specified
  - Propensity score working model is correctly specified

Doubly Robust DiD procedure with Panel

Sant'Anna and Zhao (2020) considered the following doubly robust estimand when panel data are available:

$$ATT^{dr,p} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)\left((Y_{t=2} - Y_{t=1}) - \left(m_{t=2}^{G=\infty}(X) - m_{t=1}^{G=\infty}(X)\right)\right)\right],$$

- In the paper, we also discuss how to tailor estimators that are also **Doubly Robust for inference**, a much more demanding task!
- **Chang (2020)** discusses how you can use **double machine learning** procedures based on this DR DiD formulation.

What if we have staggered treatment adoption?

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!**
- In my opinion, Goodman-Bacon (2021) explains this in the clearest way.

Recent Boom of New DiD Methods: TWFE Diagnostics

- What if we want to learn about TE dynamics?

Recent Boom of New DiD Methods: TWFE Diagnostics

- What if we want to learn about TE dynamics?
- Common practice: use variants of the TWFE ES specification

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

  with $D_{i,t}^k = 1\{t - G_i = k\}$ are the "event-time" dummies.

- Do γ_k^{lead}'s and γ_k^{lags}'s recover any interesting causal parameter of interest?
  - **Sun and Abraham (2021)**: γ_k^{lead}'s should not be used to "pre-test" the PTA as they can be "contaminated" (they are not guaranteed to be zero even when all assumptions hold).
  - **Sun and Abraham (2021)**: γ_k^{lags}'s are not appropriate measures of TE dynamics when TE are heterogeneous.

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.**

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 never-treated/last-treated cohort as comparison group. However, this paper has many other results about pitfalls of TWFE that are not in CS.
  - **Gardner (2021)**, **Borusyak, Jaravel and Spiess (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 rule 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 (2021b)**: 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.

Non-binary treatments

Recent Boom of New DiD Methods: Continuous and Multi-valued Treatments

- What if treatment is multi-valued or continuous?
- **Callaway, Goodman-Bacon and Sant'Anna (2021):** Make some advances on this problem (still in progress).
- We can measure treatment effect "in levels":

$$ATT(a|b) = \mathbb{E}[Y_t(a) - Y_t(0)|D = b] \quad \text{and} \quad ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)].$$

- But we can also measure treatment effects in "increments":

$$ACRT(d|d) = \left.\frac{\partial \mathbb{E}[Y_t(l)|D = d]}{\partial l}\right|_{l=d} \quad \text{and} \quad ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}.$$

or

$$ACRT(d_j|d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})|D = d_j] \quad \text{and} \quad ACR(d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})].$$

- Discuss problems with TWFE and how to fix some of these (more to come soon!)

Importance of being careful about parameter of interest

- With binary treatments and staggered adoption, the literature has somehow stressed the pitfalls of using variants of the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}.$$

- Issue is that, under some assumptions,

$$\beta = \sum_{t,g} w_{t,g} \cdot ATT(g,t),$$

but the weights $w_{t,g}$ are not guaranteed to be convex, i.e., they can be negative; see, e.g., Athey and Imbens (2021), Borusyak et al. (2021), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2021).

Importance of being careful about parameter of interest

- With binary treatments and staggered adoption, the literature has somehow stressed the pitfalls of using variants of the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}.$$

- Issue is that, under some assumptions,

$$\beta = \sum_{t,g} w_{t,g} \cdot ATT(g,t),$$

but the weights $w_{t,g}$ are not guaranteed to be convex, i.e., they can be negative; see, e.g., Athey and Imbens (2021), Borusyak et al. (2021), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2021).

- What if the weights were convex? Would this be "fine"?

Importance of being careful about parameter of interest

- With binary treatments and staggered adoption, the literature has somehow stressed the pitfalls of using variants of the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}.$$

- Issue is that, under some assumptions,

$$\beta = \sum_{t,g} w_{t,g} \cdot ATT(g,t),$$

but the weights $w_{t,g}$ are not guaranteed to be convex, i.e., they can be negative; see, e.g., Athey and Imbens (2021), Borusyak et al. (2021), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2021).

- What if the weights were convex? Would this be "fine"?

- LATE and MTE IV literature have being debating this issue for the last 20 years: **What is the causal question of interest? That should help us picking "good"**

What if treatment is continuous?

- With continuous treatments, this becomes even more important, as discussed in Callaway et al. (2021)

- Even with two periods, with no units being treated in period $t = 1$, some units remaining untreated at period $t = 2$, and the others receiving different dosages d, the β from the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}$$

  can have **very different** causal interpretations!

What if treatment is continuous?

- Under a "strong parallel trends" assumption, we have:
  - If we were to use "slope effects" as "building blocks":

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l) ACR(l)\, dl + w_0 \frac{ATE(d_L)}{d_L},$$

    where $ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}$, and all weights are non-negative and integrate to one.

  - If we were to use "level effects" as "building blocks":

$$\beta^{twfe} = \int_{\mathcal{D}_+} w_1^{\text{alt}}(l) \frac{ATE(l)}{l}\, dl$$

    where $ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)]$, and the weights integrate to one but are non-convex (i.e., can be negative).

What if treatment is continuous?

- Under a "strong parallel trends" assumption, we have:
  - If we were to use "slope effects" as "building blocks":

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l) ACR(l)\, dl + w_0 \frac{ATE(d_L)}{d_L},$$

    where $ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}$, and all weights are non-negative and integrate to one.
  - If we were to use "level effects" as "building blocks":

$$\beta^{twfe} = \int_{\mathcal{D}_+} w_1^{\text{alt}}(l) \frac{ATE(l)}{l}\, dl$$

    where $ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)]$, and the weights integrate to one but are non-convex (i.e., can be negative).
- Same estimator and same assumptions, but sharply different interpretations!

In my view, whenever it is possible, we should be clear about the causal parameter of interest from the very beginning!

Recent Boom of New DiD Methods: Continuous and Multi-valued Treatments

- What about fuzzy DiD setups?
- de Chaisemartin and D'Haultfœuille (2018): fantastic paper showing how one can handle setups where treatment is binary (say at unit level) but one is willing to impose parallel trends at a more aggregate level (say state level).
- The aggregation step leads to non-binary "treatments", and potentially all "clusters" are exposed to treatment in all periods (but with different intensity).
- de Chaisemartin and D'Haultfœuille (2018) shows that the "Wald-estimand" has a LATE interpretation when the effect of the treatment is stable over time, and if the effect of the treatment is the same in the treatment and in the control group.
- Since these assumptions are strong, the authors also propose alternative estimators that build on Athey and Imbens (2006) and do not rely on these assumptions.

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 (2021):** 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.

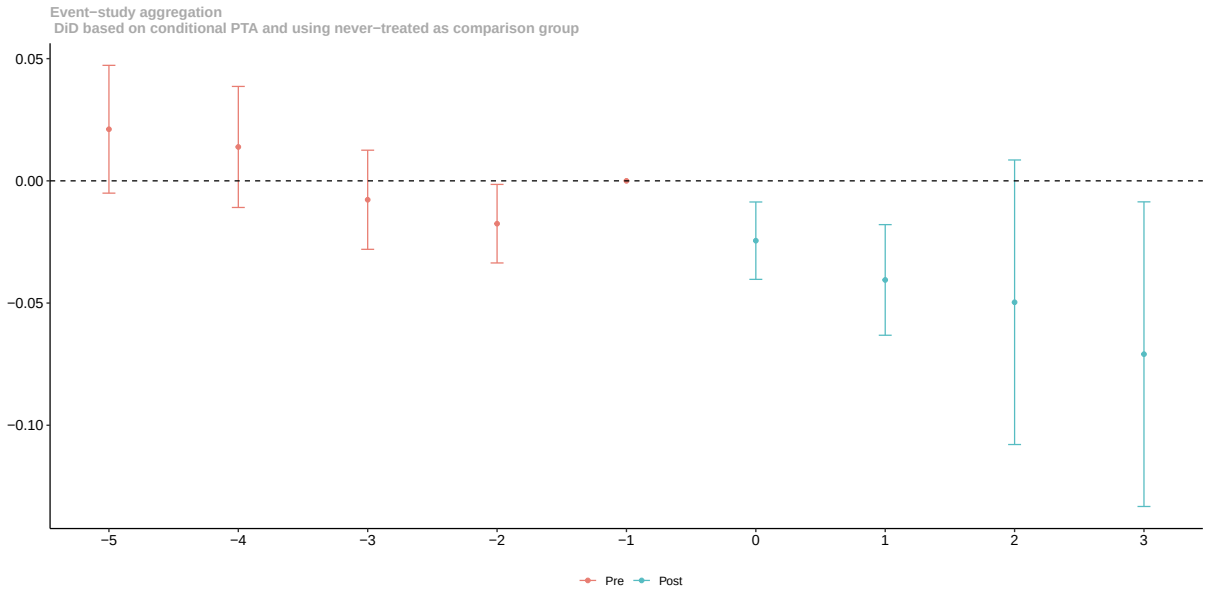
- However, as a good econometrician, instead of sitting in our Ivory Tower, we need to seek several practical, easy-to-use tools that can alleviate some of these problems.

- This is what Rambachan and Roth (2021) do!
- In my view, the sensitivity analysis procedures in Rambachan and Roth (2021) are fundamental to improve 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 (2021), respectively.

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

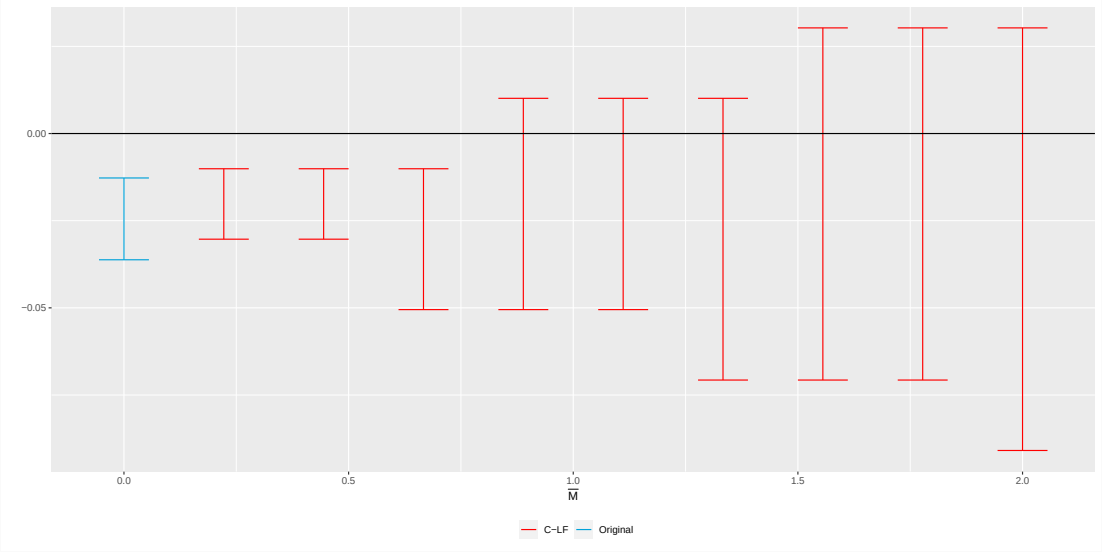
```
# Install the packages (I used the Github versions)
devtools::install_github("bcallaway11/did");
devtools::install_github("asheshrambachan/HonestDiD")

#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
0.05
0.00
−0.05
−0.10
−5
−4
−3
−2
−1
0
1
2
3
Pre
Post

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

```
# Brant has 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(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



When is PT sensitive to functional form?

Recent Boom of New DiD Methods: When is PT sensitive to functional form?

- When is PT sensitive to functional form?

Recent Boom of New DiD Methods: When is PT sensitive to functional form?

- When is PT sensitive to functional form?
- **Roth and Sant'Anna (2021a):** Provide necessary and sufficient conditions for DiD estimators to be insensitive to functional form restrictions.
- This holds if and only if PT holds in a distributional sense.
- This is testable - can cast it as a test of monotonicity!
- Also provides some "microfoundations" of how PT can hold in this particular distributional sense.

Inference with few treated clusters

Recent Boom of New DiD Methods: What if we have a handful of clusters only?

- What if we have a handful of clusters only?

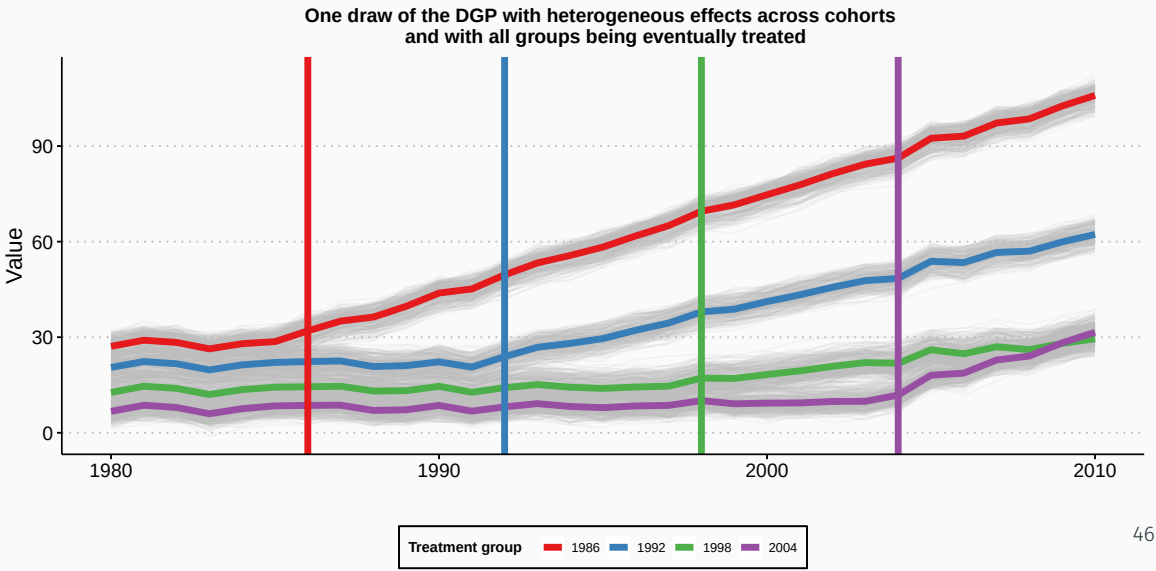
Recent Boom of New DiD Methods: What if we have a handful of clusters only?

- What if we have a handful of clusters only?
- The literature has tackled this question using different restrictions on potential outcomes and/or treatment effect heterogeneity.
- This is a hard problem, especially when we do not want to impose restrictions on the time-dependency of the potential outcomes.
- Most of the literature adopts a “regression-view” of the problem, which, in my view, hide important implications of the required assumptions for these solutions to work.
- For the interest of time, I refer to Section 5 of Roth et al. (2021) for more details.

Let’s now zoom into staggered setups and why TWFE is not “clean”

Illustration of TWFE problems with staggered treatment adoptions

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.

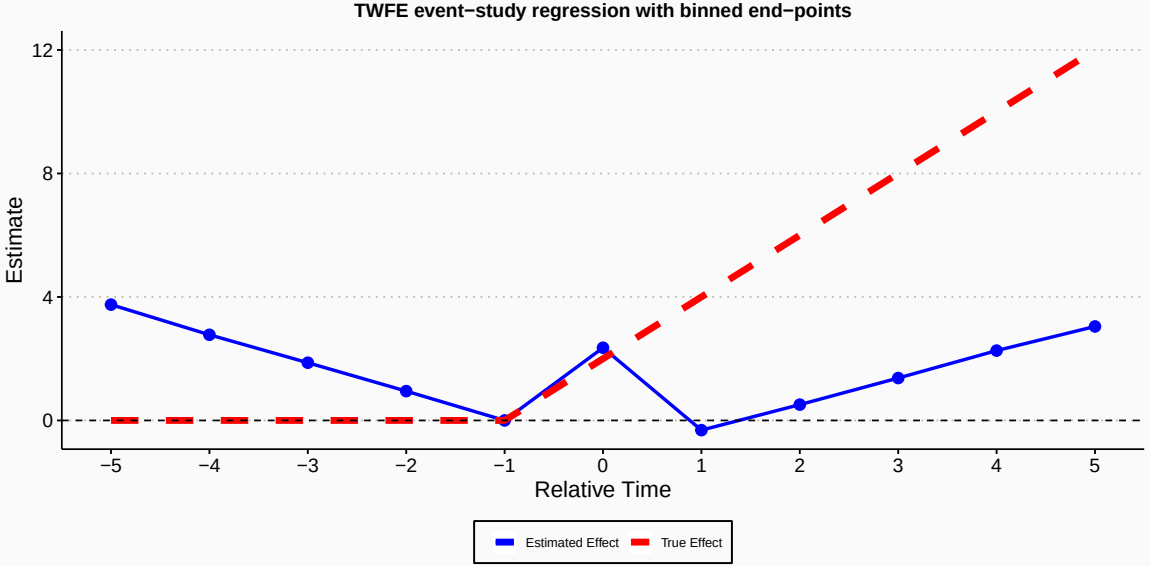
Traditional methods: TWFE event-study regression

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

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

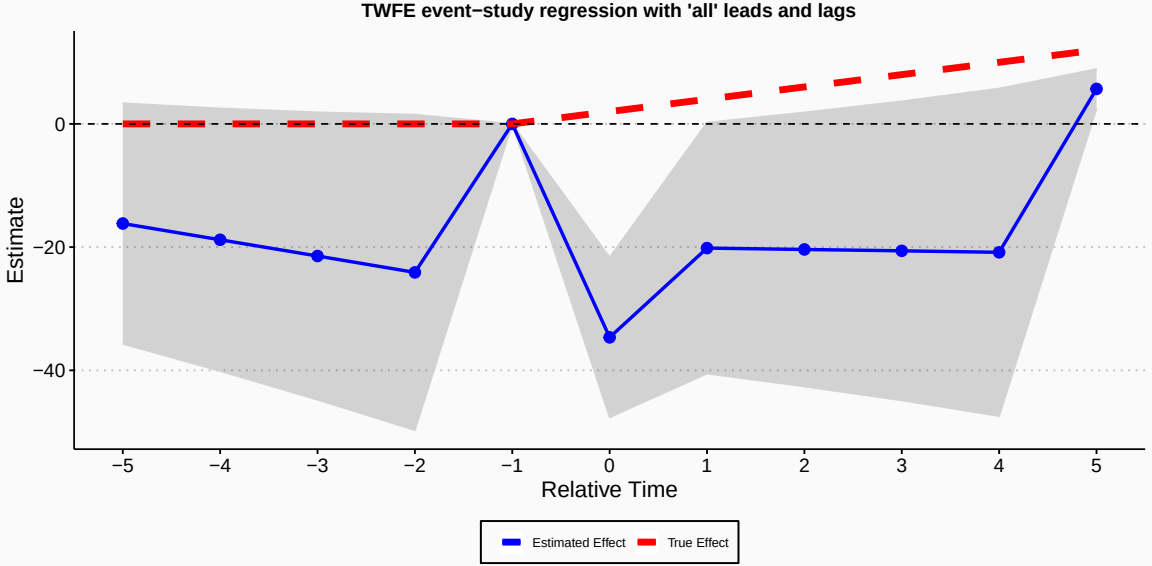
  with K and L to be equal to 5 ?

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

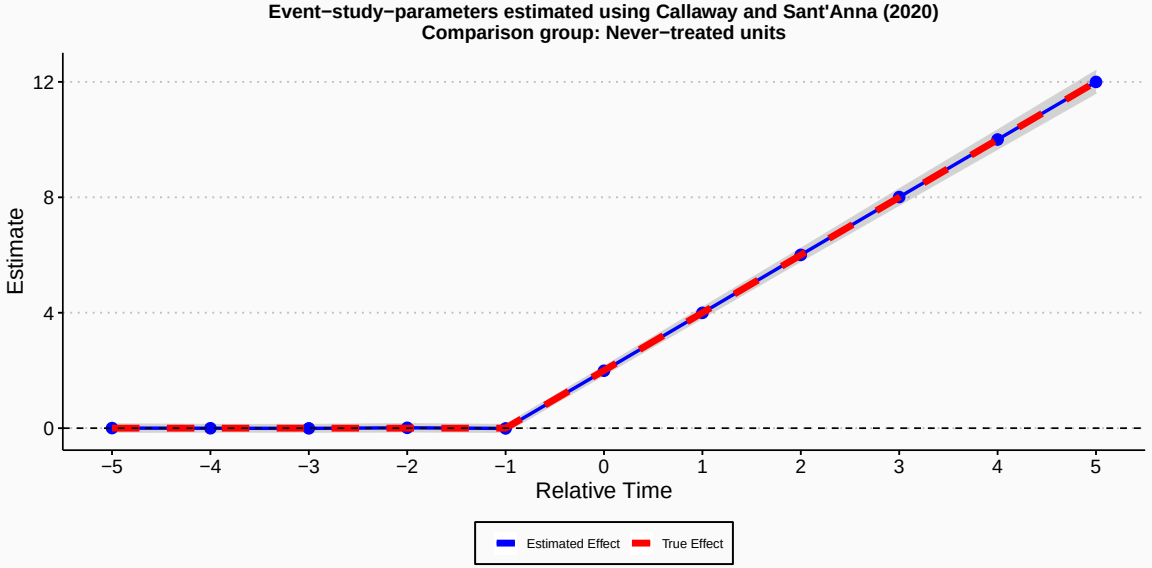

TWFE event−study regression with binned end−points
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
Estimated Effect
True Effect

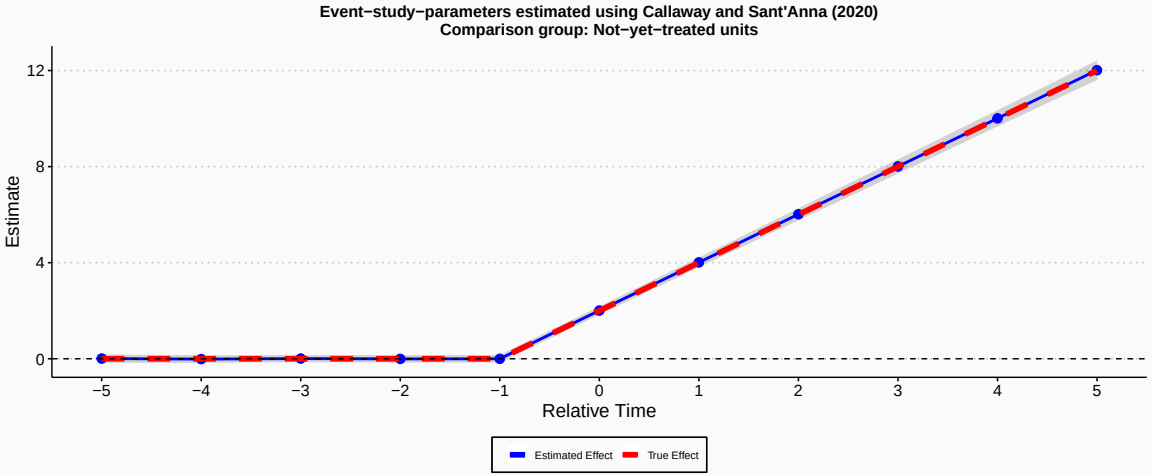
Traditional methods: TWFE event-study regression

- 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 $D_{i,t}^{\leq -K}$ and of $D_{i,t}^{\geq L}$ 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
Estimated Effect
True Effect

Is there hope?


Event−study−parameters estimated using Callaway and Sant'Anna (2020)
Comparison group: Never−treated units
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
Estimated Effect
True Effect


Event−study−parameters estimated using Callaway and Sant'Anna (2020)
Comparison group: Not−yet−treated units
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
Estimated Effect
True Effect

Callaway and Sant'Anna (2020)

Main take-way message

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

Simplification

For simplicity, let's focus on the case without covariates X

Identification

Let’s talk about identification

Callaway and Sant'Anna (2020)

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 $ATT(g,t)$

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

- 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

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

- 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:

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

- 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 $g = t$.

Callaway and Sant'Anna (2020)

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, calendar time.
- Practioners 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:

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

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

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

and

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

- 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

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

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.

Third step: Estimation and Inference

Callaway and Sant'Anna (2020)

Estimation and Inference

Estimation

- Identification results suggest a simple plug-in estimation procedure.
- Replace population expectations with their empirical analogues.
- Callaway and Sant'Anna (2021) allows for covariates and provides high-level conditions that first-step estimators have to satisfy.
    - 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.

- **Solution:** Use bootstrap to do simultaneous inference.

- Details are on the paper (and also on slides available on my webpage).

Thanks!!!

Assumption (Panel Data Sampling Scheme)

The data $\{Y_{i,t=1}, Y_{i,t=2}, G_i\}_{i=i}^{n}$ *is a random sample of the population of interest.*

- We observe data at periods $t = 1$ and $t = 2$ for the same units.
- We can also consider the case with repeated cross-sections, though notation is different.

- Recall that our DiD estimator is

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

- In the panel data case, we can simplify this a bit further:

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty},$$

where $\overline{\Delta Y}_{g=d}$ is the sample mean of $\Delta Y_i \equiv Y_{i,t=2} - Y_{i,t=1}$ for units in group d,

$$\overline{\Delta Y}_{g=d} = \frac{\sum_{i:G_i=d} \Delta Y_i}{n_{G=d}} = \frac{n^{-1}\sum_{i=1}^n \Delta Y_i 1\{G_i = d\}}{n^{-1}\sum_{i=1}^n 1\{G_i = d\}} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = d\}\right]}{\mathbb{E}_n\left[1\{G = d\}\right]},$$

and $n_{G=d} = \sum_{i=1}^n 1\{G = d\}$ is the sample size of group $G = d$.

- Henceforth, for a generic variable A,

$$\mathbb{E}_n\left[A\right] \equiv \frac{\sum_{i=1}^n A_i}{n}.$$

DiD estimator in 2x2 setups with Panel Data

Go back

- We then have that

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]}.$$

- We want to know if this estimator is "reliable".
  - As number of units increase, does it converges in probability the true ATT, under our assumptions?
  - How can we conduct reliable inference about the ATT without invoking distributional assumptions?
- We will rely on large sample approximations results.
- All those stats class you took (or teach), can be very handy now!
- We will use LLN + CMT + CLT.

- Since
$$\widehat{\theta}_n^{DiD} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]},$$
consistency follows directly from law of large numbers and continuous mapping theorem.

- **LLN**: with iid + bounded moments (which we are implicitly assuming), sample means converge in probability to population means.

- **Continuous mapping theorem**: continuous functionals preserve limits.

- As a result, we have, as $n \to \infty$,
$$\widehat{\theta}_n^{DiD} \xrightarrow{p} \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]} = \mathbb{E}\left[\Delta Y | G=2\right] - \mathbb{E}\left[\Delta Y | G=\infty\right] \equiv \theta^{DiD},$$
and $\theta^{DiD} = ATT$ under SUTVA + No-Anticipation + PT assumptions.

Asy. Normality of the DiD estimator in 2x2 setups with Panel Data

Go back

- Now, we want to derive the asymptotic distribution of

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) &= \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\ &\quad -\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right).\end{aligned}$$

- To get there, we can use CLT and Delta Method (iid + finite asymptotic variance + denominator bounded away from zero.)

- We will do this slightly different because I want to get the **influence function**.

- Express $\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right)$ as as average of *iid* terms + negligible terms.

- Let's first analyze

$$\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right).$$

- With some manipulation, we can rewrite this as

$$\begin{aligned}
&\frac{1}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right] - \mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]\right) \\
&-\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]^2}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right] - \mathbb{E}\left[1\{G=2\}\right]\right) \\
&+\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right] \cdot \left(\mathbb{E}_n\left[1\{G=2\}\right] - \mathbb{E}\left[1\{G=2\}\right]\right)}{\mathbb{E}\left[1\{G=2\}\right]^2 \cdot \mathbb{E}_n\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right] - \mathbb{E}\left[1\{G=2\}\right]\right) \\
&-\frac{\left(\mathbb{E}_n\left[1\{G=2\}\right] - \mathbb{E}\left[1\{G=2\}\right]\right)}{\mathbb{E}\left[1\{G=2\}\right] \cdot \mathbb{E}_n\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right] - \mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]\right).
\end{aligned}$$

- Red terms converges in probability to zero by LLN
- Blue terms converges in distribution to Normal with finite variance by CLT.
- Then, by Slutsky's Theorem

$$\begin{aligned}
&\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \quad & \frac{1}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right] - \mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]\right) \\
& - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]^2}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right] - \mathbb{E}\left[1\{G=2\}\right]\right) \\
& + o_p(1).
\end{aligned}$$

- Rearranging some terms (and with some abuse of notation), we have

$$\begin{aligned}
&\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \;& \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] - \sqrt{n}\mathbb{E}\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] \\
& - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] - 1\right) + o_p(1) \\
= \;& \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] - \sqrt{n}\mathbb{E}\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] \\
& - \sqrt{n}\mathbb{E}_n\left[\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] + \sqrt{n}\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]} + o_p(1).
\end{aligned}$$

- Continuing the manipulations...

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \; & \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] - \sqrt{n}\mathbb{E}_n\left[\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] + o_p(1) \\
= \; & \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] + o_p(1) \\
= \; & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right] + o_p(1).
\end{aligned}
$$

- Thus, we have that

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \; & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right] + o_p(1) \\
= \; & \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{1\{G_i=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y_i \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right)}_{=\xi_{i,G=2}\ :\text{ influence function}} + o_p(1) \\
= \; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{i,G=2} + o_p(1),
\end{aligned}
$$

- The $\xi_{i,G=2}$ is the **influence function** we were after: it is mean zero, has finite variance and is *iid*.

- Now, following exactly the same steps as we did, we have that

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right) \\
= \; & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=\infty\}}{\mathbb{E}\left[1\{G=\infty\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right)\right] + o_p(1) \\
= \; & \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{1\{G_i=\infty\}}{\mathbb{E}\left[1\{G=\infty\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y_i \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right)\right)}_{=\xi_{i,G=\infty}\ :\text{ influence function}} + o_p(1) \\
= \; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{i,G=\infty} + o_p(1),
\end{aligned}
$$

- Putting these pieces together, it follows that

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) &= \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\ &\quad - \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_{i,G=2} - \xi_{i,G=\infty}) + o_p(1)\end{aligned}$$

- Now, it follows from the CLT that, as $n \to \infty$ (i.e., number of treated and untreated units increase),

$$\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) \xrightarrow{d} N(0, V_p),$$

where

$$V_p = \mathbb{E}\left[(\xi_{G=2} - \xi_{G=\infty})^2\right] = \mathbb{E}\left[\xi_{G=2}^2\right] + \mathbb{E}\left[\xi_{G=\infty}^2\right]$$

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Athey, Susan and Guido Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

_ and _, "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," Technical Report, Working Paper 2021.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, (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, "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, April 2018, *85* (2), 999–1028.

_ and _ , "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.

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

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

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, October 1997, *64* (4), 605–654.

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, "An Honest Approach to Parallel Trends," *Working Paper*, 2021.

Roth, Jonathan and Pedro H. C. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," oct 2021, pp. 1–36.

_ 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.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, November 2020, *219* (1), 101–122.

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.