Introduction to Modern Difference-in-Differences Methods

Marcelo Ortiz-Villavicencio


EMORY
UNIVERSITY

August 23, 2025

EconThaki - Ecomienza 2025

Introduction

- In many applications, researchers does not have access to experimental data (it could be very expensive or just unethical to run)
- Without access to experimental data, we need to rely on *observational data*.
- Thus, the only way to make progress is to rely on *assumptions* to identify treatment effects.
- In causal inference, different methods rely on different assumptions and some times even different *estimands*.
- Our goal is to assess the *plausibility* of these assumptions in addressing the questions that matter to us.
- This lecture aims to navigate the latest advancements in *Difference-in-Differences* (DiD), a key approach for analyzing *observational panel data*.

Roadmap

1. Motivation
2. Quick Recaps
   - Potential Outcomes Framework
   - Causal Estimands
3. Classical DiD with 2×2 setup
4. DiD with variation in treatment timing
   - Callaway and Sant'Anna (2021)
5. Overcoming violation in PT assumptions?
   - Sensitivity Analysis
   - Triple Differences
6. Appendix

Papers we'll cover in this lecture

Journal of Economic Literature

ISSN 0022-0515 (Print) | ISSN 2328-8175 (Online)

About the JEL | Articles and Issues | Information for Authors

Difference-in-Differences Designs: A Practitioner's Guide

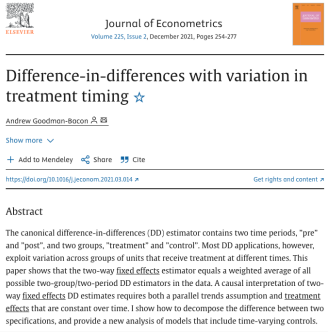
Andrew Baker
Brantly Callaway
Scott Cunningham
Andrew Goodman-Bacon
Pedro H. C. Sant'Anna

JOURNAL OF ECONOMIC LITERATURE (FORTHCOMING)

Abstract

Difference-in-differences (DiD) is arguably the most popular quasi-experimental research design. Its canonical form, with two groups and two periods, is well understood. However, empirical practices can be ad hoc when researchers go beyond that simple case. This article provides an organizing framework for discussing different types of DiD designs and their associated DiD estimators. It discusses covariates, weights, handling multiple periods, and staggered treatments. The organizational framework, however, applies to other extensions of DiD methods as well.

Baker et al. (2025)



Journal of Econometrics

Volume 225, Issue 2, December 2021, Pages 254-277

Difference-in-differences with variation in treatment timing ☆

Andrew Goodman-Bacon

Show more

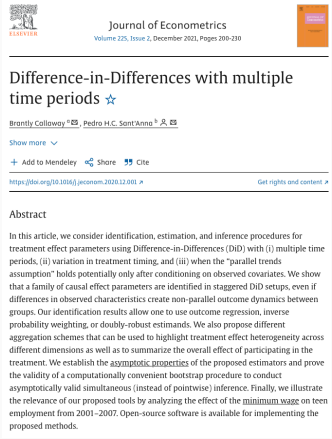
Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.jeconom.2021.03.014 | Get rights and content

Abstract

The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.

Goodman-Bacon (2021)



Journal of Econometrics

Volume 225, Issue 2, December 2021, Pages 200-230

Difference-in-Differences with multiple time periods ☆

Brantly Callaway, Pedro H.C. Sant'Anna

Show more

Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.jeconom.2020.12.001 | Get rights and content

Abstract

In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the validity of a computationally convenient bootstrap procedure to conduct asymptotically valid simultaneous (instead of pointwise) inference. Finally, we illustrate the relevance of our proposed tools by analyzing the effect of the minimum wage on teen employment from 2001–2007. Open-source software is available for implementing the proposed methods.

Callaway & Sant'Anna (2021)

Papers we'll cover in this lecture

JOURNAL ARTICLE FEATURED

A More Credible Approach to Parallel Trends

Get access

Ashesh Rambachan , Jonathan Roth

The Review of Economic Studies, Volume 90, Issue 5, October 2023, Pages 2555–2591, https://doi.org/10.1093/restud/rdad018

Published: 15 February 2023 **Article history**

A correction has been published: *The Review of Economic Studies*, Volume 90, Issue 5, October 2023, Page 2674, https://doi.org/10.1093/restud/rdad056

Cite Permissions Share

Abstract

This paper proposes tools for robust inference in difference-in-differences and event-study designs where the parallel trends assumption may be violated. Instead of requiring that parallel trends holds exactly, we impose restrictions on how different the post-treatment violations of parallel trends can be from the pre-treatment differences in trends ("pre-trends"). The causal parameter of interest is partially identified under these restrictions. We introduce two approaches that guarantee uniformly valid inference under the imposed restrictions, and we derive novel results showing that they have desirable power properties in our context. We illustrate how economic knowledge can inform the restrictions on the possible violations of parallel trends in two economic applications. We also highlight how our approach can be used to conduct sensitivity analyses showing what causal conclusions can be drawn under various restrictions on the possible violations of the parallel trends assumption.

Rambachan & Roth (2023)

Better Understanding Triple Differences Estimators*

Marcelo Ortiz-Villavicencio † Pedro H. C. Sant'Anna‡

July 21, 2025

Abstract

Triple Differences (DDD) designs are widely used in empirical work to relax parallel trends assumptions in Difference-in-Differences (DiD) settings. This paper highlights that common DDD implementations—such as taking the difference between two DiDs or applying three-way fixed effects regressions—are generally invalid when identification requires conditioning on covariates. In staggered adoption settings, the common DiD practice of pooling all not-yet-treated units as a comparison group can introduce additional bias, even when covariates are not required for identification. These insights challenge conventional empirical strategies and underscore the need for estimators tailored specifically to DDD structures. We develop regression adjustment, inverse probability weighting, and doubly robust estimators that remain valid under covariate-adjusted DDD parallel trends. For staggered designs, we demonstrate how to effectively utilize multiple comparison groups to obtain more informative inferences. Simulations and three empirical applications highlight bias reductions and precision gains relative to standard approaches. A companion R package is available.

JEL: C10; C14; C21; C23.
Keywords: Triple Differences; Difference-in-Differences; Difference-in-Difference-in-Differences; Parallel Trends; Doubly Robustness; Staggered Adoption.

Ortiz–Villavicencio & Sant'Anna (2025)

Additional Resources

A bunch of valuable online resources there exist out there to learn about and stay updated in this rapidly growing literature. Some of my favorites are:

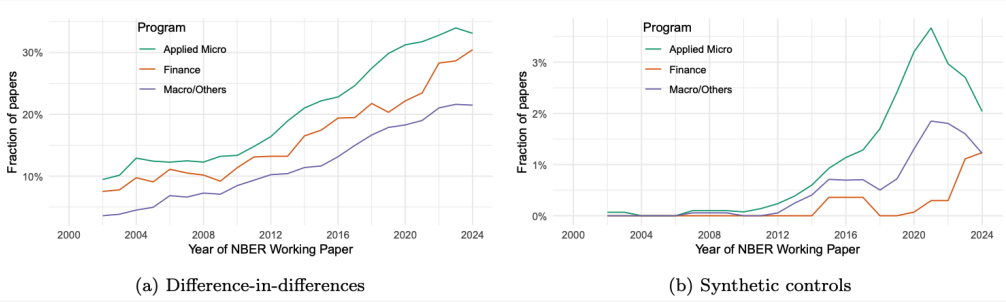
- *DiD Resources* by Pedro Sant'Anna.
- *Difference-in-Differences Workshop* by Brant Callaway.
- *DiD Resources* by Jonathan Roth.
- *Short Course on Causal Inference with Panel Data* by Yiqing Xu.
- *Scott's Mixtape Substack* by Scott Cunningham.
- *DiD Digest* by Bia Gietner. Very useful if you want to stay informed with the latest releases in the DiD literature!

Indeed, this lecture borrows insights and valuable material from these resources!

Motivation

DiD Boom

- Over the past few years, we have seen significant advancements in our understanding of the Difference-in-Differences (DiD).
- The main reason is that DiD methods are easy to implement and interpret, fostering the *credibility revolution*.
- Goldsmith-Pinkham (2024) documents the popularity of DiD among other methods on NBER working papers



(a) Difference-in-differences

(b) Synthetic controls

Why DiD is so popular?

- **Maps to real policy rollouts.** Policies often start in some places first (or at different times). DiD's treated vs. comparison and before vs. after structure matches that reality.
- **Relative low data demands.** Works with panels or even repeated cross-sections; doesn't require an instrument or sharp cutoff like RDD.
- **Software availability.** Traditional and modern DiD tools are available in Stata, R, and Python, including the very recent ones.
- **Simple, transparent identification**. The core assumption—parallel trends—is easy to state, plot, and debate its plausibility. Event-study graphs make designs legible.
- **Works with heterogeneous and dynamic effects.** By utilizing event-study aggregations, we can gain a clearer insight into the evolution of treatment effects in relation to the duration since treatment initiation.

Quick Recaps

Potential Outcomes Framework

- We adopt the *Rubin Causal Model* by employing the potential outcomes framework.
- This is not the only available framework. E.g., in computer science, causal diagrams and Pearl's Causal Inference Framework are widely adopted.
- The potential outcomes framework focuses on *counterfactuals*, defining potential outcomes for each individual under different treatment scenarios, while causal diagrams *visually* represent relationships between variables and help identify potential sources of bias in causal analysis.
- When dealing with *panel data*, it is logical to define potential outcomes based on the *initial treatment time* of the units.

Basic Notation

- We assume we have access to a sample of n units indexed by $i = 1, \ldots, n$, and T periods indexed by $t = 1, \ldots, T$.
- Some units are exposed to a binary (*absorbing*) treatment in any time period $t > 1$.
- Absorbing treatments are not the only patterns permitted in DiD designs. Treatments can actually be switched on or off at various time periods, or multiple treatments might be activated simultaneously. However, to simplify, we focus on a single binary (absorbing) treatment.
- Let $G_i \in \mathcal{G} \subset \{1, \ldots, T\} \cup \{\infty\}$ denote the time unit i is first-treated, with the notion that if a unit is *never-treated*, we set $G_i = \infty$. Assume a never-treated group always exists.
- When is needed, we assume that a vector of *pre-treatment covariates* X_i is available, which support is denoted by $\mathcal{X} \subseteq \mathbb{R}^d$.

Potential Outcomes

- Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively.
- Let $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ denote the potential outcome for unit i at time t if unit i is first treated at time g; and $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$.
- For the sake of simplicity, $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and use $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ to denote *never-treated* potential outcomes.
- Observed outcome data in time period t for unit i is given by

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g).$$

Quick Example

To put this in context, let's try a quick example using the previous notation. Suppose we are in the *canonical 2x2 DiD setup*.

- "2x2" means 2 groups, 2 time periods. Then, $t \in \{1, 2\}$. In the first period $t = 1$, nobody is treated.
- Some units can start the treatment at time $g = 2$, whereas the rest of the units will remain untreated at time $t = 2$. Therefore, the support of "first-treated periods" is reduced to $\mathcal{G} = \{2, \infty\}$.
- The map from potential outcomes to observed outcomes at any time t is given by

$$Y_{i,t} = 1\{G_i = 2\} \times Y_{i,t}(2) + 1\{G_i = \infty\} \times Y_{i,t}(\infty).$$

- For those units with $g = \infty$ (*never-treated*), we observe $Y_{i,1}(\infty)$ and $Y_{i,2}(\infty)$.
- For those units with $g = 2$ (*treated-units*), we observe $Y_{i,1}(2)$ and $Y_{i,2}(2)$.

If you are familiar with the potential outcomes framework with cross-sectional data, notice that we some extra notation $Y_{i,t}(2) \equiv Y_{i,t}(0, 1)$ and $Y_{i,t}(\infty) \equiv Y_{i,t}(0, 0)$.

Fundamental Problem of Causal Inference

- In the simplest scenario, we aim to observe how an outcome changes following the implementation of a policy.
- This can be simply expressed in terms of potential outcomes as

$$Y_{i,t}(2) - Y_{i,t}(\infty)$$

- This difference represents the causal effect of the treatment on the outcome of unit i at time t.
- Problem: We cannot observe both potential outcomes at time t!

Fundamental Problem of Causal Inference

		Data			
Unit	G_i	$Y_{i,t=1}(2)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$
1	∞	?	?	(✓) $Y_{i=1,t=1}$	(✓) $Y_{i=1,t=2}$
2	2	(✓) $Y_{i=2,t=1}$	(✓) $Y_{i=2,t=2}$	?	?
3	∞	?	?	(✓) $Y_{i=3,t=1}$	(✓) $Y_{i=3,t=2}$
4	2	(✓) $Y_{i=4,t=1}$	(✓) $Y_{i=4,t=2}$	?	?
⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	2	(✓) $Y_{i=n-1,t=1}$	(✓) $Y_{i=n-1,t=2}$	?	?
n	2	(✓) $Y_{i=n,t=1}$	(✓) $Y_{i=n,t=2}$	?	?

✓ Observed data ? Unobserved counterfactual (Missing data)

Target Parameters

- In general, individualized treatment effects (unit-specific) are very hard to recover without strong assumptions.
- Our focus is going to be on treatment effects in the *average* sense.
- Unless otherwise noted, we assume that we have a *balanced panel data* which is a random sample of $(Y_{t=1}, \ldots, Y_{t=T}, G, X)$.
- Our target is to identify the treatment effect in the *post-treatment* periods.

Definition (Average Treatment Effect on the Treated)

In particular, most DiD designs target the average treatment effect on the treated at time t, or $ATT(t)$:

$$\begin{aligned} ATT(t) &= \mathbb{E}\left[Y_{i,t}(2) - Y_{i,t}(\infty) \mid G_i = 2\right] \\ &= \mathbb{E}\left[Y_{i,t} \mid G_i = 2\right] - \mathbb{E}\left[Y_{i,t}(\infty) \mid G_i = 2\right]. \end{aligned}$$

- Then, our identification problem comes from the fact that we never observe $\mathbb{E}\left[Y_{i,t}(\infty) \mid G_i = 2\right]$ in $t = 2$.

Target Parameters

- When we have multiple periods and multiple groups starting treatment at different times, we can extend the $ATT(t)$ notion to a group-time average treatment effect (Callaway and Sant'Anna, 2021).

Definition (Group-time Average Treatment Effect)

Average treatment effect of being first-treated in period g compared to never-being treated, among the units first-treated in period g, at time period t is defined by

$$\begin{aligned} ATT(g,t) &= \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = g] \\ &= \mathbb{E}[Y_{i,t} | G_i = g] - \mathbb{E}[Y_{i,t}(\infty) | G_i = g] \end{aligned}$$

- Similarly, our identification problem comes from the fact that we never observe $\mathbb{E}[Y_{i,t}(\infty) | G_i = g]$ in $t \geq g$.

Other parameters

- Although $ATT(g,t)$ and $ATT(t)$ are typically the focus in DiD designs, more complex treatment designs require to adapt the estimands of interest. A few examples:
- *Reversible Treatments*: Here we care about the history of treatment $\mathbf{d}$ versus no treatment. E.g., $\mathbf{d} = (0,1,1,0)$.

$$ATT(\mathbf{d},t) = \mathbb{E}[Y_{i,t}(\mathbf{d}) - Y_{i,t}(\infty)|D_i = \mathbf{d}]$$

- *Multi-valued Treatments*: The treatment d can assume more than two distinct discrete values. E.g., $d \in \{1,2,3,\ldots,k\} \cup \{\infty\}$. For any two levels d, d':

$$ATT(d|d') = \mathbb{E}[Y_{i,t}(d) - Y_{i,t}(d')|D_i = d]$$

- *Continuous Treatments*: Treatment d represents a dose and can take continuous values. E.g., $d \in \mathcal{D}_+ \subseteq (0,\infty)$. We extend the concept to average causal derivatives:

$$ACD(d|d') = \left.\frac{\partial \mathbb{E}\left[Y_{i,t}(s) \mid D = d'\right]}{\partial s}\right|_{s=d}$$

Classical DiD with 2×2 setup

Setup for 2×2 DiD setup

Using our previously introduced notation, in this section we are going to focus in the 2-period$\times$2-groups DiD setup.

Data Structure:

- 2 periods: $t \in \{1, 2\}$
- No one treated at period $t = 1$.
- Some units are treated at period $t = 2$.
- $\mathcal{G} = \{2, \infty\}$.

Potential Outcomes: $Y_{i,t}(2)$ and $Y_{i,t}(\infty)$.
Observed Outcomes: $Y_{i,t=1}$ and $Y_{i,t=2}$.

SUTVA + No Anticipation Assumptions

Assumption (SUTVA)

The map from potential outcomes to observed outcomes at any time t is given by

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$$

Why we need it? By imposing SUTVA, we are ruling out interference across units. That means, if we have take two units i and j, the potential outcomes for unit i, $Y_{i,t}(\cdot)$, are not affected by the treatment status of unit j, G_j.

Assumption (No-anticipation)

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

Why we need it? This assumption restricts the behavior of unit-specific treatment effects but *does not* restrict treatment effect heterogeneity in periods $t \geq g$ (i.e., post-treatment periods). It is plausible when the policy is not announced well in advance, preventing units from manipulating their treatment status once the policy takes effect. Additionally, it is convenient because it allows us to simplify the notation.

By applying the two prior assumptions, our *missing data problem* becomes "simpler".

SUTVA + No Anticipation Assumptions

		Data			
Unit	G_i	$Y_{i,t=1}(2)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$
1	∞	(✓) $Y_{i=1,t=1}(\infty)$	?	(✓) $Y_{i=1,t=1}$	(✓) $Y_{i=1,t=2}$
2	2	(✓) $Y_{i=2,t=1}$	(✓) $Y_{i=2,t=2}$	?	?
3	∞	(✓) $Y_{i=3,t=1}(\infty)$	?	(✓) $Y_{i=3,t=1}$	(✓) $Y_{i=3,t=2}$
4	2	(✓) $Y_{i=4,t=1}$	(✓) $Y_{i=4,t=2}$	?	?
⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	2	(✓) $Y_{i=n-1,t=1}$	(✓) $Y_{i=n-1,t=2}$	?	?
n	2	(✓) $Y_{i=n,t=1}$	(✓) $Y_{i=n,t=2}$	?	?

SUTVA + No Anticipation Assumptions

		Data		
Unit	G_i	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(2)$	$Y_{i,t=2}(\infty)$
1	∞	(✓) $Y_{i=1,t=1}(\infty)$	?	(✓) $Y_{i=1,t=2}$
2	2	(✓) $Y_{i=2,t=1}$	(✓) $Y_{i=2,t=2}$	?
3	∞	(✓) $Y_{i=3,t=1}(\infty)$	?	(✓) $Y_{i=3,t=2}$
4	2	(✓) $Y_{i=4,t=1}$	(✓) $Y_{i=4,t=2}$	?
⋮	⋮	⋮	⋮	⋮
$n-1$	2	(✓) $Y_{i=n-1,t=1}$	(✓) $Y_{i=n-1,t=2}$	?
n	2	(✓) $Y_{i=n,t=1}$	(✓) $Y_{i=n,t=2}$	?

Identification Problem

We are interested in the effect of the treatment in *post-treatment periods*, i.e., $t = 2$. Our target parameter is then

$$ATT(2) = \mathbb{E}[Y_{i,2}(2) - Y_{i,2}(\infty)|G = 2]$$

How can we recover this parameter? Can just take the difference in outcomes at $t = 2$ between treated and control groups?

$$\begin{aligned}
\mathbb{E}[Y_{i,2}|G = 2] - \mathbb{E}[Y_{i,2}|G = \infty] &= \mathbb{E}[Y_{i,2}(2)|G = 2] - \mathbb{E}[Y_{i,2}(\infty)|G = \infty] \\
&= \mathbb{E}[Y_{i,2}(2)|G = 2] - \mathbb{E}[Y_{i,2}(\infty)|G = 2] + \mathbb{E}[Y_{i,2}(\infty)|G = 2] \\
&\quad - \mathbb{E}[Y_{i,2}(\infty)|G = \infty] \\
&= \mathbb{E}[Y_{i,2}(2) - Y_2(\infty)|G = 2] \\
&\quad + (\mathbb{E}[Y_{i,2}(\infty)|G = 2] - \mathbb{E}[Y_{i,2}(\infty)|G = \infty]) \\
&= ATT(2) + \text{Selection Bias}
\end{aligned}$$

A simple difference in means at $t = 2$ does not recover $ATT(2)$ due to the presence of *selection bias*!

Parallel Trends Assumption

As mentioned previously, our challenge in identifying the treatment effect arises because we cannot observe $\mathbb{E}[Y_{i,t}(\infty) \mid G_i = 2]$ at $t = 2$. Additionally, we demonstrated that taking the difference in means at $t = 2$ is insufficient to recover the treatment effect. This is where the *parallel trends assumption* comes into play!

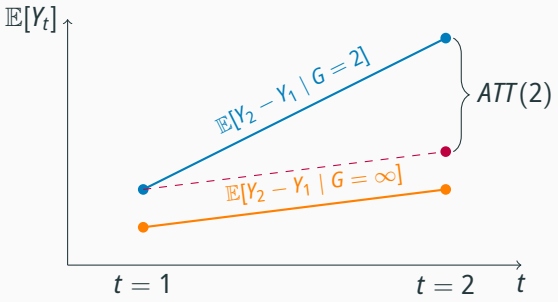
Assumption (Parallel Trends Assumption)

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

What does it mean? We are saying that, *in the absence of treatment*, the evolution of the outcome among treated units (i.e., $G_i = 2$) is, *on average*, the same as the evolution of the outcome among the control units (i.e., $G_i = \infty$).

ok so... how the PTA could help us?

DiD mechanics



From *PTA*, we know that:

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

By *SUTVA* + *No-anticipation*, we impose that:

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

Finally, we can identify the treatment effect by:

$$ATT(2) = (E[Y_{i,2}|G_i = 2] - E[Y_{i,1}|G_i = 2]) - (E[Y_{i,2}|G_i = \infty] - E[Y_{i,1}|G_i = \infty])$$

how can we estimate $ATT(2)$ in practice?

TWFE or Plug-in estimator

- The most straightforward approach to estimation is the *plug-in* estimator. In simple words, replace population expectations by sample means.

$$\widehat{ATT(2)} = \left(\bar{Y}_{g=2,t=2} - \bar{Y}_{g=2,t=1}\right) - \left(\bar{Y}_{g=\infty,t=2} - \bar{Y}_{g=\infty,t=1}\right)$$

  where $\bar{Y}_{g=g',t=t'} = \frac{\sum_{i=1}^{n} 1\{G_i=g',t=t'\}Y_{i,t'}}{\sum_{i=1}^{n} 1\{G_i=g',t=t'\}}$.

- Example:

	$\bar{Y}_{g=g',t=t'}$		
	$g'=2$	$g'=\infty$	Diff
$t'=2013$	419.2	474.0	-54.8
$t'=2014$	428.5	483.1	-54.6
Diff	*9.3*	*9.1*	0.2

TWFE or Plug-in estimator

- Although the plug-in estimator is very intuitive, it lack generality.
- In practice, researchers prefer to estimate the *ATT* in DiD using regression methods.
- Probably the most popular approach (questionable in some contexts though) is the *two-way fixed effects* (TWFE) regression specification:

$$Y_{i,t} = \alpha_0 + \beta_{twfe} 1\{G_i = 2\}\{T_i = 2\} + \beta_1 1\{G_i = 2\} + \beta_2 \{T_i = 2\} + \varepsilon_{i,t}$$

with $\mathbb{E}[\varepsilon_{i,t}|G_i, T_i] = 0$.

- One can play with the terms of this regression to show that β_{twfe} recovers the parameter of interest, $ATT(2)$.
- Recall we show that $ATT(2)$ is identified by
  $ATT(2) = (E[Y_{i,2}|G_i = 2] - E[Y_{i,1}|G_i = 2]) - (E[Y_{i,2}|G_i = \infty] - E[Y_{i,1}|G_i = \infty])$.
- Set of moment restrictions where $\mathbb{E}[Y_{i,t'}|G_i = g'] \equiv \mathbb{E}[Y_{i,t}|G_i = g', T_i = t']$ are given by:

$$\begin{aligned}
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1] &= \alpha_0 \\
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2] &= \alpha_0 + \beta_2 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1] &= \alpha_0 + \beta_1 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2] &= \alpha_0 + \beta_1 + \beta_2 + \beta_{twfe}
\end{aligned}$$

- These imply that

$$ATT(2) = \left[\left(\alpha_0 + \beta_1 + \beta_2 + \beta_{twfe}\right) - \left(\alpha_0 + \beta_1\right)\right] - \left[\left(\alpha_0 + \beta_2\right) - \left(\alpha_0\right)\right] = \beta_{twfe}$$

TWFE regression

- The main benefit of using TWFE regressions is its familiarity to economists, allowing us to conveniently perform *asymptotically valid inference*.
- However, TWFE regressions can become more complex by:
  - Introducing additional covariates.
  - Dealing with multiple time periods.
  - Variation in treatment timing (e.g., staggered treatment adoption)
  - Complex treatment patterns.
- Recent literature on DiD highlights the limitations of TWFE regressions in addressing *treatment effect heterogeneity* and more complex scenarios.
- However, it is important to note that the limitations are not inherently TWFE regressions' fault. Ultimately, TWFE is just an *estimation tool*, and it is the responsibility of the researcher to ensure that the results are contextually relevant.

What can go wrong with *TWFE regressions*?

Conditional Parallel Trends Assumption

- First, let us consider the inclusion of covariates in the analysis.
- Including covariates in the analysis can sometimes help to relax the PT assumption.
- **Question**: what happens if units with different observed characteristics were to evolve differently in the absence of treatment?
- The PT assumption may be questioned if there is an *imbalance* in the *pre-treatment characteristics* related to the outcome variable dynamics between the treated and control groups.
- We want to allow for *covariate-specific trends*!

Conditional Parallel Trends Assumption

We relax the previously introduced PTA by assuming that it hold *only after* conditioning on a vector of ovserved *pre-treatment covariates*.

Assumption (Conditional Parallel Trends Assumption)

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

What does it mean? We are saying that, *in the absence of treatment*, *conditional on X*, the evolution of the outcome among treated units (i.e., $G_i = 2$) is, *on average*, the same as the evolution of the outcome among the control units (i.e., $G_i = \infty$).

Strong Overlap

Having introduced covariates in the analysis, we will need to introduce an additional assumption stating that every units has a *strictly positive* probability of being in the *untreated* group.

Assumption (Strong Overlap Assumption)

The conditional probability of belonging to the treatment group, given observed characteristics X, is uniformly bounded away from 1.
That is, for some $\epsilon > 0, \mathbb{P}[G = 2 \mid X] < 1 - \epsilon$ *almost surely.*

Why we need it? Without overlap, especially in the context of observational data, it becomes difficult or impossible to find comparable individuals in the control group for every individual in the treated group. This lack of overlap forces reliance on models to *extrapolate* beyond the observed data, leading to unreliable and potentially biased causal estimates.

So can we use *TWFE regressions* again?

TWFE Regression with covariates

- Under unconditional PTA, we have shown that we can use a TWFE regression specification to recover the $ATT(2)$.
- If covariates are available, one might intuitively assume that simply adding them to the previous specification is enough.

$$Y_{i,t} = \tilde{\alpha}_0 + \tilde{\beta}_{twfe} 1\{G_i = 2\}\{T_i = 2\} + \tilde{\beta}_1 1\{G_i = 2\} + \tilde{\beta}_2\{T_i = 2\} + X_i'\Gamma + ui, t$$

  with $\mathbb{E}[u_{i,t}|G_i, T_i, X_i] = 0$.
- Is $\tilde{\beta}_{twfe}$ recovering the ATT?

- By following the same reasoning as before, adding the Law of Iterated Expectations and assuming Conditional PTA, Strong Overlap, it is straightforward to demonstrate that:

$$
\begin{aligned}
ATT(2) &= \mathbb{E}\Big[\left(\mathbb{E}\left[Y_{i,2}|G_i = 2, X\right] - \mathbb{E}\left[Y_{i,1}|G_i = 2, X\right]\right) \\
&\quad - \left(\mathbb{E}\left[Y_{i,2}|G_i = \infty, X\right] - \mathbb{E}\left[Y_{i,1}|G_i = \infty, X\right]\right) |G_i = 2\Big] \\
&= \mathbb{E}[ATT(2, X)|G_i = 2] \\
&= \left(\mathbb{E}\left[Y_{i,2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,1}|G_i = 2\right]\right) - \mathbb{E}\left[\left(\mathbb{E}\left[Y_{i,2}|G_i = \infty, X\right] - \mathbb{E}\left[Y_{i,1}|G_i = \infty, X\right]\right) |G_i = 2\right]
\end{aligned}
$$

TWFE Regression with covariates

- Set of moment restriction where $\mathbb{E}[Y_{i,t'}|G_i = g', X] \equiv \mathbb{E}[Y_{i,t}|G_i = g', T_i = t', X]$ are given by

$$\begin{aligned}
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1, X_i] &= \tilde{\alpha}_0 + X_i'\Gamma \\
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2, X_i] &= \tilde{\alpha}_0 + \tilde{\beta}_2 + X_i'\Gamma \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1, X_i] &= \tilde{\alpha}_0 + \tilde{\beta}_1 + X_i'\Gamma \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2, X_i] &= \tilde{\alpha}_0 + \tilde{\beta}_1 + \tilde{\beta}_2 + \tilde{\beta}_{twfe} + X_i'\Gamma
\end{aligned}$$

- Notice that $\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2, X_i] - \mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1, X_i] = \tilde{\beta}_{twfe} + \tilde{\beta}_2$
- Similarly, $\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2, X_i] - \mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1, X_i] = \tilde{\beta}_2$
- Therefore, it implies that $ATT(2, X) = \tilde{\beta}_{twfe}$
- In words: This implies that we are imposing that average treatment effects are *homogeneous* between *covariate subpopulations*!

Then, what should we do?

Alternatives

- Our primary challenge is allowing for covariate-specific trends within each group.
- However, TWFE regression appears to be too inflexible for this purpose.
- Several alternatives have been proposed in the literature for DiD designs with covariates that are more flexible:
  - **Regression Adjustments**: Proposed by Heckman, Ichimura and Todd (1997), it focuses on modeling the *outcome regression* process for each group over time.
  - **IPW**: Proposed originally by Abadie (2005), this method focuses on modeling the *propensity score*.
  - **AIPW**: Proposed by Sant'Anna and Zhao (2020), the main idea is to provide an additional layer of robustness by integrating the other two approaches.

1) Regression Adjustment

- The idea is to directly utilize the identifying assumption previously discussed.

$$
\begin{aligned}
ATT(2) &= \mathbb{E}[ATT(2,X)|G_i = 2] \\
&= \mathbb{E}[(\underbrace{\mathbb{E}\left[Y_{i,2}|G_i = 2, X\right]}_{=m_{t=2}^{G=2}(X)} - \underbrace{\mathbb{E}\left[Y_{i,1}|G_i = 2, X\right]}_{=m_{t=1}^{G=2}(X)}) \\
&\quad - (\underbrace{\mathbb{E}\left[Y_{i,2}|G_i = \infty, X\right]}_{=m_{t=2}^{G=\infty}(X)} - \underbrace{\mathbb{E}\left[Y_{i,1}|G_i = \infty, X\right]}_{=m_{t=1}^{G=\infty}(X)})|G_i = 2] \\
&= \mathbb{E}\left[\left(m_{t=2}^{G=2}(X) - m_{t=1}^{G=2}(X)\right) - \left(m_{t=2}^{G=\infty}(X) - m_{t=1}^{G=\infty}(X)\right) \mid G_i = 2\right]
\end{aligned}
$$

- Estimate the unknown *regression functions* $m_t^G(X)$ using your preferred method, e.g., OLS.
- Note that with access to *panel data*, the identification problem simplifies to:

$$ATT_{reg} = \mathbb{E}[Y_{i,2} - Y_{i,1}|G_i = 2] - \mathbb{E}[m_{\Delta}^{G=\infty}(X)|G_i = 2]$$

where $m_{\Delta}^{G=\infty}(X) := \mathbb{E}[Y_{i,2} - Y_{i,1}|G_i = \infty, X]$.

Ok, but I still need to model $m_t^G(X)$ accurately, right?

2) Inverse Probability Weighting

- Another approach would avoid to directly modeling the outcome regression process $m_t^G(X)$.
- Instead, it models the *propensity score*, i.e., $\pi(x) := P(1\{G_i = 2\}|X)$

$$ATT_{ipw} = \mathbb{E}\left[\left(\frac{1\{G_i = 2\}}{\mathbb{E}[1\{G_i = 2\}]} - \frac{\frac{\pi(X)(1-1\{G_i=2\})}{1-\pi(X)}}{\mathbb{E}\left[\frac{\pi(X)(1-1\{G_i=2\})}{1-\pi(X)}\right]}\right)(Y_{i,2} - Y_{i,1})\right]$$

- Again, we need to estimate the unknown *propensity score* function $\pi(x)$ using your preferred method, e.g., logit.
- To get an estimator, plug in the fitted propensity score values into ATT_{ipw}, and replace the population expectations with sample means.

Can we combine both procedures?

3) Augmented Inverse Probability Weighting

- The two previous procedures rely on researcher ability to model correctly either $m_t^G(x)$ or $\pi(x)$.
- A natural question should be: can we combine both approaches to have an extra layer of *robustness*?
- Indeed, Sant'Anna and Zhao (2020) proposes estimators that are *Doubly Robust consistent*: they are consistent for the $ATT(2)$ if either (we just need one to hold):
  - $m_t^G(x)$ for outcome evolution are correctly specified,
  - $\pi(x)$ is correctly specified.

$$ATT_{dr} = \mathbb{E}\left[\left(\frac{1\{G_i = 2\}}{\mathbb{E}[1\{G_i = 2\}]} - \frac{\frac{\pi(X)(1-1\{G_i=2\})}{1-\pi(X)}}{\mathbb{E}\left[\frac{\pi(X)(1-1\{G_i=2\})}{1-\pi(X)}\right]}\right)\left((Y_{i,2} - Y_{i,1}) - m_{\Delta}^{G=\infty}(X)\right)\right]$$

- If both $m_t^G(x)$ and $\pi(x)$ are correctly specified, the DR DiD estimator for the ATT_{dr} is "the most precise estimator" (*minimum asymptotic variance*) among all (regular) estimators that does not rely on additional functional form restrictions!

What have we learned so far?

- Although TWFE regressions are “user-friendly”, they can sometimes produce misleading results.
- For instance, if we aim to incorporate covariate-specific trends, the estimate of β_{twfe} can be *significantly biased*. Why?
- Alternative approaches to handle covariate-specific trends includes Regression Adjustments, IPW, or AIPW.
- Covariates should be pre-treatment variables. There are methods to address time-varying covariates, but they are beyond the scope of this lecture. See e.g., Caetano and Callaway (2023)

DiD with variation in treatment timing

Does TWFE regressions work in setups with *variation in treatment timing*?

TWFE regression with multiple time periods

- Another typical setup in DiD designs occurs when researchers have access to multiple time periods and different units implement the treatment at different times.
- **Example:** States that raise their minimum wage above the federal level are classified as *treated*. Furthermore, states can choose to deviate from the federal minimum wage at any time throughout the analysis period, while some states may not make any changes to their minimum wage policy.
- This type of setting is referred to as *staggered designs*, where the treatment is assumed to be in an *absorbing state*; once a unit is treated, it remains treated permanently.

TWFE with staggered adoption

- We have shown that in the 2×2 case, we can recover the $ATT(2)$ by β_{twfe} from the following TWFE specification

$$Y_{i,t} = \alpha_0 + \beta_{twfe} 1\{G_i = 2\}\{T_i = 2\} + \beta_1 1\{G_i = 2\} + \beta_2 \{T_i = 2\} + \varepsilon_{i,t}$$

- We may want to "generalize" that TWFE specification to tackle multiple time periods setups:

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

where dummies $D_{i,t} = 1\{t - G_i \geq 0\}$, where G_i indicates the period unit i is first treated (Group) and $D_{i,t}$ is an indicator for unit i being ever-treated by period t.

- But, what is the parameter β actually recovering?

TWFE with staggered adoption

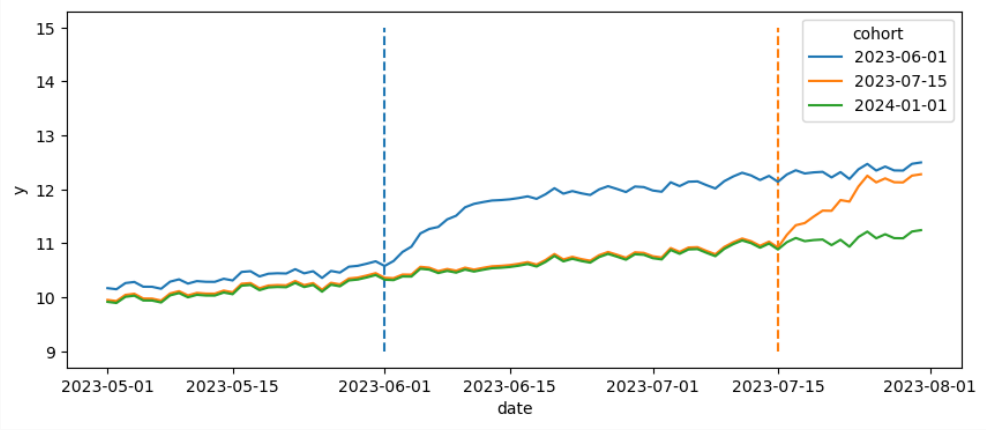
- Several studies have addressed this question: Goodman-Bacon (2021); Athey and Imbens (2018); Borusyak, Jaravel and Spiess (2023); De Chaisemartin and d'Haultfoeuille (2020)
- The key message from these studies is that when treatment effects are *heterogeneous/dynamic*, the parameter β is not straightforward to interpret.
- The parameter β represents a weighted average of ATTs across different groups and periods. However, some of these weights may be *negative* (De Chaisemartin and d'Haultfoeuille, 2020).
- Even when weights are non-negative, they are not really intuitive because involve making *forbidden comparisons* (Goodman-Bacon, 2021).

Issue with TWFE

- TWFE as weighted-average of 2x2 comparisons (Goodman-Bacon, 2021)
  1. Newly treated vs Never treated;
  2. Newly treated vs Not-yet treated;
  3. Newly treated vs Earlier treated.
- The TWFE regression specification does not respect the identification assumptions and uses *already-treated* units as a comparison group for *later-treated* units.

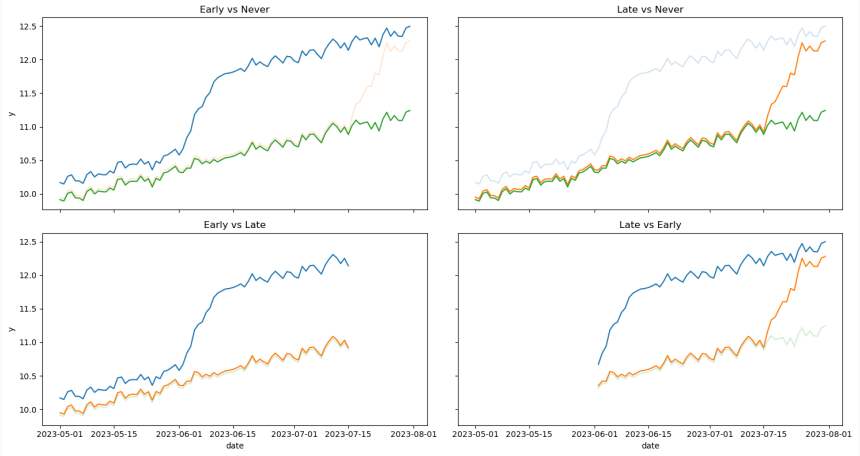
Numerical Example

Figure 1: DiD with different timing - Simulated DGP



Numerical Example

Figure 2: Decomposition of multiple 2*x*2 DiD



- TWFE estimator is an average of well-understood 2x2 DiD estimators, like those plotted in the previous figure, with weights based on subsample shares.

Theorem (Difference-in-Differences Decomposition Theorem)

Assume that the data contain $k = 1, \ldots, K$ timing groups of units ordered by the time when they receive a binary treatment, $k \in (1, T]$. There may be one timing group, U, that includes units that never receive treatment. The OLS estimate, $\hat{\beta}^{TWFE}$, in a two-way fixed-effects regression is a weighted average of all possible two-by-two DiD estimators:

$$\hat{\beta}^{TWFE} = \sum_{k \neq U} s_{kU} \hat{\beta}^{2\times 2}_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^{k}_{k\ell} \hat{\beta}^{2\times 2,k}_{k\ell} + s^{\ell}_{k\ell} \hat{\beta}^{2\times 2,\ell}_{k\ell} \right].$$

where the 2×2 DiD estimators are:

$$\begin{aligned}
\hat{\beta}^{2\times 2}_{kU} &\equiv \left(\bar{y}^{\text{POST}(k)}_{k} - \bar{y}^{PRE(k)}_{k} \right) - \left(\bar{y}^{\text{POST}(k)}_{U} - \bar{y}^{\text{PRE}(k)}_{U} \right), \\
\hat{\beta}^{2\times 2,k}_{k\ell} &\equiv \left(\bar{y}^{\text{MID}(k,\ell)}_{k} - \bar{y}^{PRE(k)}_{k} \right) - \left(\bar{y}^{\text{MID}(k,\ell)}_{\ell} - \bar{y}^{PRE(k)}_{\ell} \right), \\
\hat{\beta}^{2\times 2,\ell}_{k\ell} &\equiv \left(\bar{y}^{POST(\ell)}_{\ell} - \bar{y}^{\text{MID}(k,\ell)}_{\ell} \right) - \left(\bar{y}^{POST(\ell)}_{k} - \bar{y}^{\text{MID}(k,\ell)}_{k} \right).
\end{aligned}$$

Let's go back to first principles!

- We have already seen that the traditional TWFE specifications would not work well for us.
- Let's go back to basic principles: Ensure that you only make the comparisons you want to
  - Fix the parameter of interest.
  - State your identification assumptions
  - Get an identification results (estimand) which should guide the choice of the estimation method by plug-in principle.

Callaway and Sant'Anna (2021) in a nutshell

- Consider a random sample

$$\left\{ \left(Y_{i,1}, Y_{i,2}, \ldots, Y_{i,\mathcal{T}}, D_{i,1}, D_{i,2}, \ldots, D_{i,\mathcal{T}}, X_i \right) \right\}_{i=1}^{n}$$

  where $D_{i,t} = 1$ if unit i is treated in period t, and 0 otherwise
- $G_{i,g} = 1$ if unit i is first treated at time g, and zero otherwise ("Treatment start-time dummies")
- $C = 1$ is a *never-treated* comparison group
- Staggered treatment adoption: $D_{i,t} = 1 \implies D_{i,t+1} = 1$, for $t = 1, 2, \ldots, \mathcal{T}$.

- Limited Treatment Anticipation: There is a known $\delta \geq 0$ s.t.

$$\mathbb{E}\left[Y_{i,t}(g) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_{i,t}(\infty) \mid X, G_g = 1\right]$$

for all $g \in \mathcal{G}, t \in 1, \ldots, \mathcal{T}$ with $\underbrace{t < g - \delta}_{\text{before effective starting date}}$.

- Generalized propensity score uniformly bounded away from 1 :

$$p_{g,t}(X) = P\left(G_g = 1 \mid X, G_g + (1 - D_t)\left(1 - G_g\right) = 1\right) \leq 1 - \epsilon$$

Definition (Group-time average treatment effect)

Average Treatment Effect at time t of starting treatment at time g, among the units that indeed started treatment at time g is defined as

$$\text{ATT}(g,t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(\infty) \mid G_g = 1\right], \text{ for } t \geq g - \delta$$

PT assumption based on "never treated" group

Assumption (Conditional Parallel Trend based on a "never treated")

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

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \mid X, C = 1\right] \text{ a.s.}$$

PT assumption based on "not-yet-treated" group

Assumption (Conditional Parallel Trend based on a "not-yet-treated")

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

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \mid X, D_s = 0, G_g = 0\right] \text{ a.s.}$$

Identification: Never treated as comparison group

If one invokes the Conditional PTA based on *never treated* units, Callaway and Sant'Anna (2021) shows that, for all g and t such that $g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \mathcal{T}\}, t \in \{2, \ldots \mathcal{T}-\delta\}$ and $t \geq g - \delta$, $\text{ATT}(g,t)$ is nonparametrically identified by the DR estimand

$$\text{ATT}^{\text{nev}}_{dr}(g,t;\delta) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1} - m^{nev}_{g,t,\delta}(X)\right)\right]$$

where $m^{nev}_{g,t,\delta}(X) = \mathbb{E}\left[Y_t - Y_{g-\delta-1} \mid X, C=1\right]$

- The same logic as the 2×2 DiD with covariates applies from this point.

Identification: not-yet-treated as comparison group

If one invokes the Conditional PTA based on *not-yet-treated* units, Callaway and Sant'Anna (2021) shows that, for all g and t such that $g \in \mathcal{G}_\delta, t \in 2, \dots \mathcal{T} - \delta$ and $t \geq g - \delta$,

$$\text{ATT}^{ny}_{dr}(g,t;\delta) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\frac{p_{g,t+\delta}(X)(1-D_{t+\delta})}{1-p_{g,t+\delta}(X)}}{\mathbb{E}\left[\frac{p_{g,t+\delta}(X)(1-D_{t+\delta})}{1-p_{g,t+\delta}(X)}\right]}\right)\left(Y_t - Y_{g-\delta-1} - m^{ny}_{g,t,\delta}(X)\right)\right]$$

where $m^{ny}_{g,t,\delta}(X) = \mathbb{E}\left[Y_t - Y_{g-\delta-1} \mid X, D_{t+\delta} = 0, G_g = 0\right]$

- The same logic as the 2×2 DiD with covariates applies from this point.

Summarizing $ATT(g,t)$: Event-studies

- $ATT(g,t)$'s act as *building blocks*, which can be used to construct more informative parameters.
- A strategy commonly employed by applied researchers is to report *event-study* plots.
- The effect of a policy intervention may depend on the length of exposure to it.
- This represents the average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods

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

- However, this is not the only one. Other weighting schemes are discussed in Callaway and Sant'Anna (2021).

Overcoming violation in PT assumptions?

Violations of PT

- We have concentrated on the limitations of TWFE and explored some recent advancements in DiD that address these limitations.
- However, these new developments still focus on a type of PTA.
- The next logical question is: what occurs when PTA fails?
- There are several reason why PTA may fail, and different strands of the literature have been focus on how to address these potential problems.
  - PTA valid after conditioning of covariates.
  - Testing for violations of PTA.
  - *Sensitivity Analysis*.
  - *Designs that overcome violations in PTA*.
- In this final part of the lecture, we will discuss some of these topics (Note that this is not an exhaustive review of the existing literature on this matter).

Sensitivity Analysis

- Sensitivity analysis evaluates how causal effect estimates might be biased when the identifying assumption are *violated* in *specific* and *meaningful* ways.
- Sensitivity analysis is distinct from testing as identifying assumptions are often inherently *non-testable*, serving more as an *"insurance" check*.
- Sensitivity analysis in causal inference originated from the Hill-Fisher debate on the link between smoking and lung cancer, and it was first formalized by Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin and Wynder (1959).
- In DiD, if we knew the true *pre-trends*, that would be informative about the counterfactual *post-treatment* differences in trends.

A More Credible Approach to Parallel Trends

- Rambachan and Roth (2023) proposes a method to set bounds by using information from *pre-trends* to constrain the counterfactual differences in trends.
- The key idea: counterfactual difference in trends cannot be *too different* than observed pre-trends.
- Formally, Rambachan and Roth (2023) allows us to determine how different the *counterfactual trend* would need to be from the *pre-trends* to overturn a conclusion (i.e., alter your initial conclusion).

Imposing restrictions on violations of PTA

- Consider a setup where there are 3 time periods: $t \in \{-1, 0, 1\}$ where treatment occurs at $t = 1$. Then $\mathcal{G} = \{1, \infty\}$.
- We parametrized the violation of PTA by:

$$\delta_1 = \mathbb{E}[Y_{i,1}(\infty) - Y_{i,0}(\infty) | G_i = 1] - \mathbb{E}[Y_{i,1}(\infty) - Y_{i,0}(\infty) | G_i = \infty]$$

- Although we cannot identify δ_1, we can employ similar logic to determine its *pre-treatment counterpart*, denoted as δ_1:

$$\delta_{-1} = \mathbb{E}[Y_{i,-1}(\infty) - Y_{i,0}(\infty) | G_i = 1] - \mathbb{E}[Y_{i,-1}(\infty) - Y_{i,0}(\infty) | G_i = \infty]$$

- Thus, the idea is to restrict the possible values of δ_1 given δ_{-1}.

Restrictions on δ_1

- **Bounds on relative magnitudes:** Set a value $\bar{M} \geq 0$ such that $|\delta_1| \leq \bar{M}\,|\delta_{-1}|$. In general, if we have more than 3 periods, say $\delta_{\text{pre}} = \left(\delta_{-\underline{T}}, \ldots \delta_{-1}\right)'$ and $\delta_{post} = (\delta_1, \ldots \delta_{\bar{T}})'$.

$$\Delta^{RM}(\bar{M}) = \left\{ \delta : \forall t \geq 0, |\delta_{t+1} - \delta_t| \leq \bar{M} \cdot \max_{s<0} |\delta_{s+1} - \delta_s| \right\}$$

  $\Delta^{RM}(\bar{M})$ bounds the *maximum post-treatment violation* of parallel trends between consecutive periods by $\bar{M}$ times the *maximum pre-treatment violation* of parallel trends.
- **Smoothness restriction:** We impose that the differential trends evolve smoothly over time by bounding the extent to which its slope may change across consecutive periods. That is, δ_1 can deviate from a *linear extrapolation* of the pre-trend: $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$ for $M \geq 0$. In general, for multiple periods

$$\Delta^{\text{SD}}(M) := \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq M, \forall t\}.$$

- Context-specific restrictions.

But… what happens if there a *time-varying confounder?*

But… what happens if there a *time-varying confounder?*

***Triple Differences* to the rescue!**

Triple Differences

- Triple Differences (DDD) extend to cases where the PTA in DiD *may not hold*.
  - **Ex**: PTA can be violated due to the presence of a *time-varying confounder* that *changes differently across states*.
- When the PTA is questionable, researchers often augment the design by adding another *placebo* comparison group to ”clean” the bias introduced by the confounder.
- DDD designs address this issue by finding a *within-state comparison group* that **is not exposed** to the treatment but is **affected** by the time-varying confounder.

Example 1: Garthwaite, Gross and Notowidigdo (2014)

- **Program**: Impact of public health insurance on labor supply using a *large public health insurance disenrollment*
- **Institutional background**:
  - In 2005, Tennessee discontinued its expansion of *TennCare*, the state's Medicaid system.
  - $\sim 170,000$ adults lost public health insurance coverage over three months.
- **DDD to the rescue**:
  - DiD-PTA could be affected by Tennessee-specific shocks other than 2005 TennCare disenrollment
  - The disenrollment primarily targeted *childless adults*. Other adults *are not exposed* to the policy but are *affected* by unobserved shocks happening in TN.

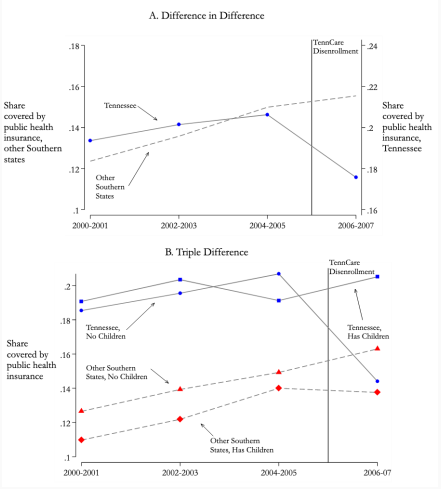


Fig 1. Share Publicly Insured

Example 2: Muralidharan and Prakash (2017)

- **Program**: Impact of giving bicycles on girls' *secondary school enrollment* in Bihar, India.
- **Institutional background**:
  - Enrollment declines as the distance to school increases.
  - Higher attrition rates for women compared to men before the program.
- **DDD to the rescue**:
  - DiD-PTA challenged by concurrent economic growth and *increased education spending* in Bihar, unlike the control state, Jharkhand.
  - Boys in Bihar *are not exposed* to the policy but are *affected* by the expansion in education spending.

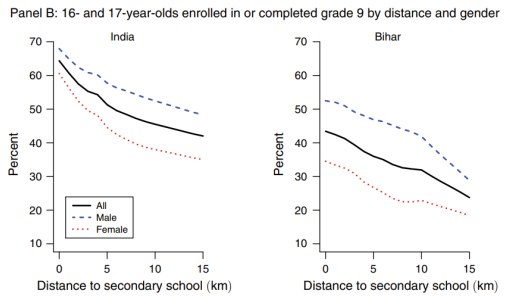
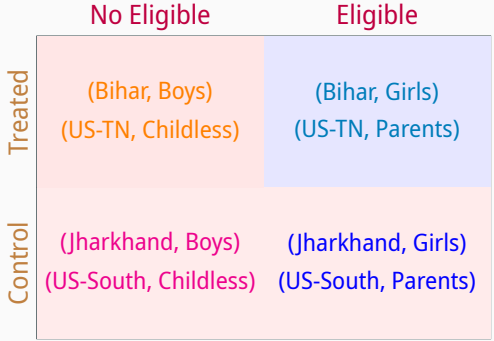


Fig 2. Distance on Female Enrollment

What's the appeal of DDD compared to DiD?

Putting everything together, in DDD we allow to use all the information to control for location-specific trends and partition-specific trends, which otherwise would arise questionable results using DiD.



What do we do in the paper?

We know very little about DDD design, even though it is widely used in empirical work. The key question in this paper is: **How can we leverage our DiD knowledge to improve our understanding about DDD**?

- We study identification, estimation, and inference procedures for ATT and ES type parameters in DDD settings.
- Our results are derived for several setups:
  - Covariates may play or not a role for identification.
  - Single treatment date or staggered treatment adoption.
- Challenging the folks' wisdom:
  - *Regression-based* procedures may enforce strict restrictions that result in *bias*.
  - DDD estimators **cannot** be expressed as the *difference between two DiD* estimators when DDD-type PTA only hold after *conditioning on covariates*.
  - In staggered DDD settings, pooling all *not-yet-treated units*—as is often done in DiD applications—can lead to substantial *bias*.

Recovering the ATT using 3WFE Regression

- Let's start with *two time periods* and single treatment date. Thus, $S = \{2, \infty\}$ and $Q = \{0, 1\}$.
- In the canonical DDD with *no covariates*, by Olden and Møen (2022), it's straightforward to show that: Difference of 2 DiDs

$$\beta_{\text{3wfe}} = \underbrace{\left[\left(\mathbb{E}\left[Y_2 - Y_1 | S = 2, Q = 1\right]\right) - \left(\mathbb{E}\left[Y_2 - Y_1 | S = 2, Q = 0\right]\right)\right]}_{\text{DiD estimand among } S=2}$$

$$- \underbrace{\left[\left(\mathbb{E}\left[Y_2 - Y_1 | S = \infty, Q = 1\right]\right) - \left(\mathbb{E}\left[Y_2 - Y_1 | S = \infty, Q = 0\right]\right)\right]}_{\text{DiD estimand among } S=\infty}$$

$$= ATT(2, 2)$$

- In general, the following 3WFE regression specification can be used to recover the $ATT(2, 2)$:

$$Y_{i,t} = \gamma_i + \gamma_{s,t} + \gamma_{q,t} + \beta_{\text{3wfe}} D_{i,t} + \varepsilon_{i,t},$$

In general, one can recover $ATT(2,2)$ in the canonical DDD either (i) by a 3WFE regression or (ii) by a difference of 2 DiDs.

In general, one can recover $ATT(2,2)$ in the canonical DDD either (i) by a 3WFE regression or (ii) by a difference of 2 DiDs.

What happens when *covariates* play an important role or we have a *staggered design*?

Notation

We have access to a sample of n units available, $i = 1, 2, \cdots, n$

- T time periods: $t = 1, 2, \ldots, T$. Each unit may be exposed to a binary (absorbing) treatment in any time period $t > 1$.
- Unit i belongs to a group that enables treatment for the first time in period $g > 1$. Let $S_i \in \mathcal{S} \subset \{2, ..., T\} \cup \{\infty\}$ be a variable that indicates the first time the policy/treatment was enabled.
- Within each group g, each unit belongs to a *population partition* that qualifies (or is eligible) for the treatment or not. Let $Q_i = 1$ if unit i qualifies, $Q_i = 0$ otherwise.
- Let $D_{i,t} = 1\{t \geq S_i, Q_i = 1\}$ be an indicator for whether unit i receives treatment in period t, and let $G_i = \min\{t : D_{i,t} = 1\}$ be the earliest period at which unit i has received treatment. **Ex**: Units in $S_i = 2$ with $Q_i = 1$ are treated at time $G_i = 2$, otherwise the unit remains untreated ($G_i = \infty$).
- A vector of *pre-treatment covariates* X_i, whose support is denoted by $\mathcal{X} \subseteq \mathbb{R}^d$ is available.

Potential Outcomes + Parameters of Interest

- Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively.
- Let $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ denote the potential outcome for unit i at time t if unit i is first treated at time g; and $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$.
- For the sake of simplicity, $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and use $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ to denote *never-treated* potential outcomes. We observe $Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$

Definition (Group-time Average Treatment Effect on the Treated)

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | S_i = g, Q_i = 1]$$

Potential Outcomes + Parameters of Interest

- Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively.
- Let $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ denote the potential outcome for unit i at time t if unit i is first treated at time g; and $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$.
- For the sake of simplicity, $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and use $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ to denote *never-treated* potential outcomes. We observe $Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$

Definition (Group-time Average Treatment Effect on the Treated)

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | S_i = g, Q_i = 1]$$

- Then, our identification problem comes from the fact that we never observe $\mathbb{E}\left[Y_{i,t}(\infty) | S_i = g, Q_i = 1\right]$ in $t \geq g$.

Potential Outcomes + Parameters of Interest

- Let $\mathbf{0}_s$ and $\mathbf{1}_s$ be s-dimensional vectors of zeros and ones, respectively.
- Let $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ denote the potential outcome for unit i at time t if unit i is first treated at time g; and $Y_{i,t}(\mathbf{0}_T)$ the outcome if untreated by time $t = T$.
- For the sake of simplicity, $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and use $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ to denote *never-treated* potential outcomes. We observe $Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g)$

Definition (Group-time Average Treatment Effect on the Treated)

$$ATT(g,t) \equiv \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | G_i = g] = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty) | S_i = g, Q_i = 1]$$

- Then, our identification problem comes from the fact that we never observe $\mathbb{E}\left[Y_{i,t}(\infty) | S_i = g, Q_i = 1\right]$ in $t \geq g$.

Definition (Aggregate Event Study Parameter)

$$ES(e) \equiv \mathbb{E}\left[ATT(G, G+e) \,\middle|\, G + e \in [2,T]\right] = \sum_{g \in \mathcal{G}_{\text{trt}}} \mathbb{P}(G = g | G + e \in [2,T]) ATT(g, g+e)$$

Assumptions

- Adding covariates linearly in the above 3WFE specification would imply additional restrictions to the DGP:
  - *Homogeneous* treatment effects in covariates.
  - Rule out *covariate-specific trends* in both the treated and comparison groups.

Assumptions

- Adding covariates linearly in the above 3WFE specification would imply additional restrictions to the DGP:
  - *Homogeneous* treatment effects in covariates.
  - Rule out *covariate-specific trends* in both the treated and comparison groups.
- Our goal is to introduce an estimator for $ATT(g,t)$ in DDD designs under the condition that DDD PTA-type is valid after controlling for covariates, i.e., $X_i \in \mathcal{X}$.

Assumptions

- Adding covariates linearly in the above 3WFE specification would imply additional restrictions to the DGP:
  - *Homogeneous* treatment effects in covariates.
  - Rule out *covariate-specific trends* in both the treated and comparison groups.
- Our goal is to introduce an estimator for $ATT(g,t)$ in DDD designs under the condition that DDD PTA-type is valid after controlling for covariates, i.e., $X_i \in \mathcal{X}$.

Assumption (DDD-Conditional Parallel Trends)

For each $g \in \mathcal{G}_{trt}$, $g' \in \mathcal{S}$ and time periods t such that $t \geq g$ and $g' > \max\{g,t\}$, with probability one,

$$\begin{aligned} &\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 1, X\right] - \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g, Q = 0, X\right] \\ &= \\ &\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g', Q = 1, X\right] - \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) | S = g', Q = 0, X\right]. \end{aligned}$$

Assumption (Random Sampling)

$\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i, S_i, Q_i)'\}_{i=1}^n$ *is a random sample from* $(Y_{t=1}, \ldots, Y_{t=T}, X', G, S, Q)'$.

Assumption (No Anticipation)

For every $g \in \mathcal{G}_{trt}$, *and every pre-treatment period* $t < g$,
$\mathbb{E}[Y_{i,t}(g)|S = g, Q = 1, X] = \mathbb{E}[Y_{i,t}(\infty)|S = g, Q = 1, X]$ *with probability one.*

Assumption (Strong Overlap)

For every $(g, q) \in \mathcal{S} \times \{0, 1\}$ *and for some* $\epsilon > 0$, $\mathbb{P}[S = g, Q = q|X] > \epsilon$ *with probability one.*

Doubly Robustness and Efficient Influence Function

EMORY UNIVERSITY

- Under the previous assumptions, the $ATT(2,2)$ can be identified via *Regression Adjustments* or *IPW* or any convex combination between them. RA & IPW
- However, this depends on the researcher's ability to accurately model outcome regression or propensity scores.
  - How would you choose this combination if the goal was to achieve **Doubly Robustness** (DR)?
  - How would you choose this combination if the goal was **efficiency**?
- We tackle these questions by deriving the *efficient influence function* for the ATT in DDD setups.
- That usually leads to DR estimands, too.

Lemma (Efficient Influence Function for DDD)

Suppose that conditional PT, no-anticipation, and strong overlap assumptions are satisfied, and balanced panel data is available. Let $\tau(\Delta Y, X) := \Delta Y - m^{S=2,Q=0}_{Y_2-Y_1}(X) - m^{S=\infty,Q=1}_{Y_2-Y_1}(X) + m^{S=\infty,Q=0}_{Y_2-Y_1}(X)$, $S := (\Delta Y, S, Q, X)$. *Then, the efficient influence function for the ATT*$(2,2)$ *is given by*

$$\begin{aligned}\eta_{eff}(S) = w^{S=2,Q=1}_{trt} \cdot \Big(\tau(\Delta Y, X) - ATT(2,2)\Big) \\ - w^{S=2,Q=0}_{comp}(X) \cdot \Big(\Delta Y - m^{S=2,Q=0}_{Y_2-Y_1}(X)\Big) \\ - w^{S=\infty,Q=1}_{comp}(X) \cdot \Big(\Delta Y - m^{S=\infty,Q=1}_{Y_2-Y_1}(X)\Big) \\ + w^{S=\infty,Q=0}_{comp}(X) \cdot \Big(\Delta Y - m^{S=\infty,Q=0}_{Y_2-Y_1}(X)\Big).\end{aligned}$$

Weights Notation

DR DDD as a function of 3 DiDs

- We can take the expected value of $\eta_{\text{eff}}(S)$ and isolate $ATT(2,2)$ given that any influence function has mean zero.
- By simple manipulation, we get the following *DR-DDD* estimand for the $ATT(2,2)$,

$$
\begin{aligned}
ATT_{\text{dr}}(2,2) = \; & \mathbb{E}\left[\left(w_{\text{trt}}^{S=2,Q=1} - w_{\text{comp}}^{S=2,Q=0}(X)\right)\left(Y_2 - Y_1 - m_{Y_2-Y_1}^{S=2,Q=0}(X)\right)\right] \\
& + \mathbb{E}\left[\left(w_{\text{trt}}^{S=2,Q=1} - w_{\text{comp}}^{S=\infty,Q=1}(X)\right)\left(Y_2 - Y_1 - m_{Y_2-Y_1}^{S=\infty,Q=1}(X)\right)\right] \\
& - \mathbb{E}\left[\left(w_{\text{trt}}^{S=2,Q=1} - w_{\text{comp}}^{S=\infty,Q=0}(X)\right)\left(Y_2 - Y_1 - m_{Y_2-Y_1}^{S=\infty,Q=0}(X)\right)\right]
\end{aligned}
$$

- For estimation, we follow a two-step approach:
  - (1) estimate the nuisance functions $p(\cdot)$ and $m(\cdot)$,
  - (2) plug-in into the sample counterpart of the previous estimand to get an estimator for $ATT(2,2)$.

DR DDD estimator for $ATT(g,t)$ with any g_c comparison group

- Let our nuisances functions be

$$m^{S=g,Q=q}_{Y_t - Y_{t'}}(X) := \mathbb{E}\left[Y_t - Y_{t'} | S = g, Q = q, X\right];$$
$$p^{S=g,Q=1}_{g',q'}(X) := \mathbb{P}[S = g, Q = 1 | X, (S = g, Q = 1) \cup (S = g', Q = q')].$$

- For any $g_c \in \mathcal{S}$ such that $g_c > \max\{g,t\}$, and any *post-treatment* period $t \geq g$, let the doubly robust DDD estimand for the $ATT(g,t)$ be given by

$$\begin{aligned} ATT_{\text{dr},g_c}(g,t) = &\mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g,0}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g,Q=0}_{Y_t - Y_{g-1}}(X)\right)\right] \\ &+ \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,1}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=1}_{Y_t - Y_{g-1}}(X)\right)\right] \\ &- \mathbb{E}\left[\left(w^{S=g,Q=1}_{\text{trt}}(S,Q) - w^{S=g,Q=1}_{g_c,0}(S,Q,X)\right)\left(Y_t - Y_{g-1} - m^{S=g_c,Q=0}_{Y_t - Y_{g-1}}(X)\right)\right] \end{aligned}$$

- Notice that if we set $g_c = \infty$, we are using *never-treated* units as comparison group!

Weights Notation and additional estimators

$ATT(g,t)$ is (non-parametrically) identified for all post-treatment periods

Theorem

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\ldots,T\}$, and $g_c \in \mathcal{S}$ such that $t \geq g$ and $g_c > t$,

$$ATT(g,t) = ATT_{dr,g_c}(g,t) = ATT_{ra,g_c}(g,t) = ATT_{ipw,g_c}(g,t).$$

Corollary

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$ and $t \in \{2,\ldots,T\}$ such that $t \geq g$, and any set of weights $w^{g,t}$ that sum up to one over $\mathcal{G}_c^{g,t}$,

$$ATT(g,t) = \sum_{\in \mathcal{G}_c^{g,t}} w^{g,t}\, ATT_{dr,g_c}(g,t).$$

$ATT(g,t)$ is (non-parametrically) identified for all post-treatment periods

Theorem

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\dots,T\}$, and $g_c \in \mathcal{S}$ such that $t \geq g$ and $g_c > t$,

$$ATT(g,t) = ATT_{dr,g_c}(g,t) = ATT_{ra,g_c}(g,t) = ATT_{ipw,g_c}(g,t).$$

Corollary

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$ and $t \in \{2,\dots,T\}$ such that $t \geq g$, and any set of weights $w^{g,t}$ that sum up to one over $\mathcal{G}_c^{g,t}$,

$$ATT(g,t) = \sum_{\in \mathcal{G}_c^{g,t}} w^{g,t}\, ATT_{dr,g_c}(g,t).$$

- These results extend the DiD identification results in Callaway and Sant'Anna (2021) for DDD setups.

$ATT(g,t)$ is (non-parametrically) identified for all post-treatment periods

Theorem

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$, $t \in \{2,\dots,T\}$, and $g_c \in \mathcal{S}$ such that $t \geq g$ and $g_c > t$,

$$ATT(g,t) = ATT_{dr,g_c}(g,t) = ATT_{ra,g_c}(g,t) = ATT_{ipw,g_c}(g,t).$$

Corollary

Let previous assumptions hold. Then, for all $g \in \mathcal{G}_{trt}$ and $t \in \{2,\dots,T\}$ such that $t \geq g$, and any set of weights $w^{g,t}$ that sum up to one over $\mathcal{G}_c^{g,t}$,

$$ATT(g,t) = \sum_{\in \mathcal{G}_c^{g,t}} w^{g,t}\, ATT_{dr,g_c}(g,t).$$

- These results extend the DiD identification results in Callaway and Sant'Anna (2021) for DDD setups.
- **Pooling all not-yet-treated would not work, though!**

Over-identification?

- Our previous result suggest that the DDD model is *over-identified*, as we can use multiple not-yet-treated enabling groups g_c as valid comparison groups.
- **Example:** To identify the $ATT(2,2)$ in a setup with $S \in \{2,3,\infty\}$, we can set $g_c = 3$ or $g_c = \infty$ as valid comparison groups and would give me the same valid result under our assumptions.
- Any weighted sum of these estimands that use different g_c's will also lead to the $ATT(g,t)$, as long as the weights sum up to one.
- We propose *combining all available options* leading to the asymptotically most precise (minimum variance) DDD estimator for the $ATT(g,t)$'s a la *Generalized Methods of Moments* (GMM).

Optimal DDD (GMM-based)

- Our two-step GMM procedure involves *stacking* the influence functions $\mathbb{IF}_{\text{dr},g_c}(g,t)$ for the available comparison cohorts where $g_c > \max\{g,t\}$.
- Let $\mathbb{RIF}_{\text{dr},g_c}(g,t) = \mathbb{IF}_{\text{dr},g_c}(g,t) + ATT_{\text{dr},}(g,t)$ denote its re-centered influence function, and denote the $k_{g,t} \times 1$ vector of all $\mathbb{RIF}_{\text{dr},g_c}(g,t)$ for $\in \mathcal{G}_c^{g,t}$ by $\mathbb{RIF}_{\text{dr}}(g,t)$.
- Since influence functions are mean zero, and that $ATT(g,t) = ATT_{\text{dr},g_c}(g,t)$ for any $\in \mathcal{G}_c^{g,t}$, we have the vector of moment conditions $\mathbb{E}[\mathbb{RIF}_{\text{dr}}(g,t) - \theta^{g,t}] = 0$, with $\theta^{g,t} = ATT(g,t)$
- From standard GMM results (Newey and McFadden, 1994), it follows that, under mild regularity conditions, the optimal (population) GMM estimator for $\theta^{g,t}$ is given by

$$\theta^{g,t}_{\text{opt}} = \frac{\mathbf{1}'\Omega^{-1}_{g,t}}{\mathbf{1}'\Omega^{-1}_{g,t}\mathbf{1}}\,\mathbb{E}[\mathbb{RIF}_{\text{dr}}(g,t)] = \frac{\mathbf{1}'\Omega^{-1}_{g,t}}{\mathbf{1}'\Omega^{-1}_{g,t}\mathbf{1}} ATT_{\text{dr}}(g,t),$$

where the last equality follows from $\mathbb{E}[\mathbb{IF}_{\text{dr}}(g,t)] = 0$ and $\Omega_{g,t} \equiv \mathbb{E}[\mathbb{IF}_{\text{dr}}(g,t) \cdot \mathbb{IF}_{\text{dr}}(g,t)']$ being positive definite.

Highlights

- DDD is widely used in empirical research, but its properties have receive little attention.
- In its basic format, it is equivalent to running two separate DiD estimators and subtracting one from another.
  - This equivalence breaks down when covariates play an important role in the analysis.
- We use semiparametric efficiency theory to forward-engineering DR DDD estimands and circumvent the issues pointed out.
  - If you feel fancy, you can even use machine learning algorithms to estimate your nuisances!
- We can leverage these results to tackle staggered treatment setups, too.
  - However, we need to be clever on how we use the information available.
- If you would like to learn more, please read the draft available on my website or ArXiv.
- An R package that automates all these procedures is available for practitioners! See https://marcelortiz.com/tripledifff/

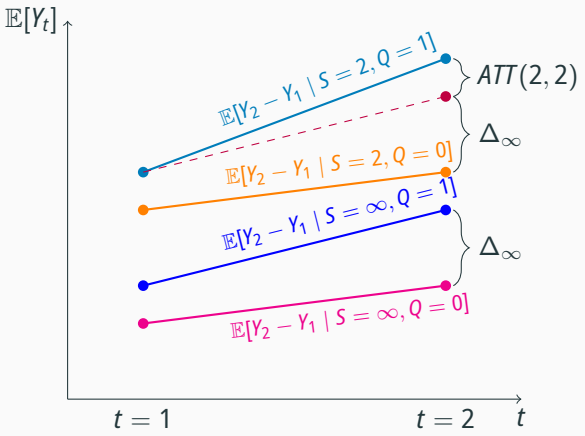
Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

Appendix

2x2x2 DDD without covariates

Let, $g \in S = \{2, \infty\}$ and $Q = \{0, 1\}$.



$$\theta = \Big[\Big(\mathbb{E}[Y_2|S=2,Q=1] - \mathbb{E}[Y_1|S=2,Q=1]\Big) - \Big(\mathbb{E}[Y_2|S=2,Q=0] - \mathbb{E}[Y_1|S=2,Q=0]\Big)\Big] - \Big[\Big(\mathbb{E}[Y_2|S=\infty,Q=1] - \mathbb{E}[Y_1|S=\infty,Q=1]\Big) - \Big(\mathbb{E}[Y_2|S=\infty,Q=0] - \mathbb{E}[Y_1|S=\infty,Q=0]\Big)\Big]$$

- Note that this is the difference of two DiD's: one among $S = 2$ across Q groups, and one among $S = \infty$ across Q groups.

Go Back

Regression Adjustment and IPW identification

We show in the paper that if conditional PT, no-anticipation, and strong overlap assumptions are satisfied and balanced panel data is available, the ATT is identified via *regression adjustments* or *IPW*:

$$ATT(2,2) = ATT_{ra}(2,2) = ATT_{ipw}(2,2),$$

where

$$ATT_{ra}(2,2) := \mathbb{E}\left[Y_2 - Y_1 | S=2, Q=1\right] - \mathbb{E}\left[m_{Y_2-Y_1}^{S=2,Q=0}(X) + \left(m_{Y_2-Y_1}^{S=\infty,Q=1}(X) - m_{Y_2-Y_1}^{S=\infty,Q=0}(X) \right) \middle| S=2, Q=1\right],$$

$$ATT_{ipw}(2,2) := \mathbb{E}\left[\left(\left(w_{trt}^{S=2,Q=1} - w_{comp}^{S=2,Q=0}(X)\right) - \left(w_{comp}^{S=\infty,Q=1}(X) - w_{comp}^{S=\infty,Q=0}(X)\right) \right) \cdot (Y_2 - Y_1)\right].$$

Go Back

Some additional notation

- For $(g, q) \in \{2, \infty\} \times \{0, 1\}$, let $\Delta Y = Y_2 - Y_1$, and

$$m^{S=g,Q=q}_{Y_2-Y_1}(X) := \mathbb{E}\left[Y_2 - Y_1 | S = g, Q = q, X\right], \quad \text{(outcome regression)}.$$
$$p^{S=g,Q=q}(X) := \mathbb{P}[S = g, Q = q | X] \quad \text{(multi-valued propensity score)}.$$

- For $(g_c, q) \in \mathcal{S}_{comp} \equiv \{(\infty, 0), (\infty, 1), (2, 0)\}$, let

$$w^{S=2,Q=1}_{trt} := \frac{1_{\{S=2,Q=1\}}}{\mathbb{E}[1_{\{S=2,Q=1\}}]},$$
$$w^{S=g_c,Q=q}_{comp}(X) := \frac{\frac{1_{\{S=g_c,Q=q\}} \cdot p^{S=2,Q=1}(X)}{p^{S=g_c,Q=q}(X)}}{\mathbb{E}\left[\frac{1_{\{S=g_c,Q=q\}} \cdot p^{S=2,Q=1}(X)}{p^{S=g_c,Q=q}(X)}\right]}$$

- Let $\tau(\Delta Y, X) := \Delta Y - m^{S=2,Q=0}_{Y_2-Y_1}(X) - m^{S=\infty,Q=1}_{Y_2-Y_1}(X) + m^{S=\infty,Q=0}_{Y_2-Y_1}(X)$, $S := (\Delta Y, S, Q, X)$.

Go Back

Regression Adjustment and IPW identification

We show in the paper that if conditional PT, no-anticipation, and strong overlap assumptions are satisfied and balanced panel data is available, the ATT is identified via *regression adjustments* or *IPW*:

$$ATT(2,2) = ATT_{ra}(2,2) = ATT_{ipw}(2,2),$$

where

$$ATT_{ra}(2,2) := \mathbb{E}\left[Y_2 - Y_1 | S=2, Q=1\right] - \mathbb{E}\left[m_{Y_2-Y_1}^{S=2,Q=0}(X) + \left(m_{Y_2-Y_1}^{S=\infty,Q=1}(X) - m_{Y_2-Y_1}^{S=\infty,Q=0}(X) \right) \middle| S=2, Q=1\right],$$

$$ATT_{ipw}(2,2) := \mathbb{E}\left[\left(\left(w_{trt}^{S=2,Q=1} - w_{comp}^{S=2,Q=0}(X)\right) - \left(w_{comp}^{S=\infty,Q=1}(X) - w_{comp}^{S=\infty,Q=0}(X)\right) \right) \cdot (Y_2 - Y_1)\right].$$

Go Back

Analogously, let the RA DDD estimand for the $ATT(g,t)$ be given by

$$ATT_{\text{ra},g_c}(g,t) = \mathbb{E}\left[w_{\text{trt}}^{S=g,Q=1}(S,Q)\left(Y_t - Y_{g-1} - m_{Y_t-Y_{g-1}}^{S=g,Q=0}(X) - m_{Y_t-Y_{g-1}}^{S=g_c,Q=1}(X) + m_{Y_t-Y_{g-1}}^{S=g_c,Q=0}(X)\right)\right]$$

and the IPW estimand be

$$\begin{aligned} ATT_{\text{ipw},g_c}(g,t) = \mathbb{E}&\left[\left(w_{\text{trt}}^{S=g,Q=1}(S,Q) - w_{g,0}^{S=g,Q=1}(S,Q,X)\right)(Y_t - Y_{g-1})\right] \\ &- \mathbb{E}\left[\left(w_{g_c,1}^{S=g,Q=1}(S,Q,X) - w_{g_c,0}^{S=g,Q=1}(S,Q,X)\right)(Y_t - Y_{g-1})\right]. \end{aligned}$$

where the weights w are given by

$$w_{\text{trt}}^{S=g,Q=1}(S,Q) := \frac{1\{S=g,Q=1\}}{\mathbb{E}[1\{S=g,Q=1\}]}, \qquad w_{g',q'}^{S=g,Q=1}(S,Q,X) := \frac{\frac{1\{S=g',Q=q'\}\cdot p_{g',q'}^{S=g,Q=1}(X)}{1-p_{g',q'}^{S=g,Q=1}(X)}}{\mathbb{E}\left[\frac{1\{S=g',Q=q'\}\cdot p_{g',q'}^{S=g,Q=1}(X)}{1-p_{g',q'}^{S=g,Q=q}(X)}\right]}.$$

Go Back

References

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

Athey, Susan and Guido Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," 2018.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2023.

Caetano, Carolina and Brantly Callaway, "Difference-in-Differences when Parallel Trends Holds Conditional on Covariates," 2023. Working Paper.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Cornfield, Jerome, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder, "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *JNCI: Journal of the National Cancer Institute*, 01 1959, *22* (1), 173–203.

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo, "Public Health Insurance, Labor Supply, and Employment Lock," *The Quarterly Journal of Economics*, 03 2014, *129* (2), 653–696.

Goldsmith-Pinkham, Paul, "Tracking the Credibility Revolution across Fields," 2024.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

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

Newey, Whitney K. and Daniel McFadden, "Large sample estimation and hypothesis testing," in "Handbook of Econometrics," Vol. 4, Amsterdam: North-Holland: Elsevier, 1994, chapter 36, pp. 2111–2245.

Olden, Andreas and Jarle Møen, "The triple difference estimator," *The Econometrics Journal*, 2022, *25* (3), 531–553.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 02 2023, *90* (5), 2555–2591.

Sant'Anna, Pedro H.C. and Jun Zhao, "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics*, 2020, *Forthcoming.*