Efficient Difference-in-Differences and Event-Study Estimators

Xiaohong Chen
Yale University

Pedro H. C. Sant'Anna
Emory University

Haitian Xie
Peking University

Seminar at Harvard Medical School, Department of Health Care Policy
April 8, 2026

Causal inference with observational data: What can we do?

- In many real-world applications, we do not have experimental data, and need to rely on **observational data**.
- With observational data alone, we have no choice but to rely on **additional assumptions/models** for the identification of causal parameters.
- Different causal estimation and inference methods, such as DiD, IV, RDD, Synthetic Control, etc., rely on **different identification assumptions**.
- As Econometricians, we aim to provide **better scientific guidance** on using these different causal methods in empirical work.
- Over the last 5 years, we have witnessed important advances in understanding DiD and Event-Study methods in the presence of heterogeneity.

The appeal of Difference-in-Differences: allow for selection on unobservables

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.

- Advantage: Allow for selection on unobservables and for time-trends.
  Identification presumes that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - **a parallel trends assumption**.

- **Data Requirements:** We need data from periods before and after the treatment/intervention to use DiD (and some periods where no unit is treated).

Recent Methodological Advances in DiD-related areas (by no means an exhaustive list)

- Athey and Imbens (2022)
- Borusyak, Jaravel and Spiess (2024)
- Callaway and Sant'Anna (2021)
- de Chaisemartin and D'Haultfœuille (2020, 2024)
- Dube, Girardi, Jordà and Taylor (2023)
- Gardner (2021)
- Goodman-Bacon (2021)
- Lee and Wooldridge (2023)
- Sant'Anna and Zhao (2020)
- Sun and Abraham (2021)
- Wooldridge (2021)

} Either "Reverse Engineer" causal interpretations for TWFE coefs, or "Forward Engineer" new heterogeneous robust DiD estimators.

- Rambachan and Roth (2023)
- Roth (2022)

} Issues with pre-tests and how to handle PT as approximation

- Roth and Sant'Anna (2023b,a)
- Ghanem, Sant'Anna and Wüthrich (2023)
- Marx, Tamer and Tang (2024)

} Better understanding PT and random treatment timing

- ?
- de Chaisemartin, D'Haultfoeuille, Pasquier, Sow and Vazquez-Bare (2024)

} DiD with continuous and multi-valued treatments

- Callaway and Li (2019)
- Tchetgen, Park and Richardson (2024)
- Wooldridge (2023)

} Nonlinear DiD Models

- Arkhangelsky, Athey, Hirshberg, Imbens and Wager (2021)
- Imbens, Kallus and Mao (2021)
- Imbens and Viviano (2023)

} Improve TWFE when T_{pre} is large

Modern DiD, in practice

- Figure 2 of Braghieri, Levy and Makarin (2022): Compare several DiD/ES estimators

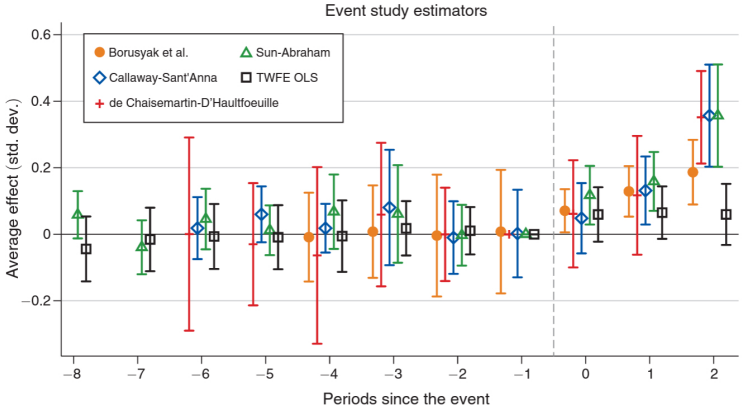
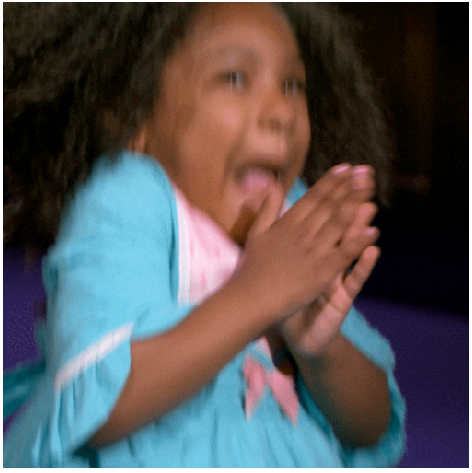


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK

INTRODUCTION



Are there open challenges?

Modern DiD: some open challenges

- How to compare different estimators
  - Different DiD estimators can rely on different identification assumptions, estimate different target parameters, and have different notions of uncertainty.

    How can we compare confidence intervals across estimators?
  - We have already seen many variants of staggered DiD estimators: are there more estimators available for the same target parameter?
  - If so, what is the common component of these families of DiD estimators?
  - Can we unify these procedures from an identification and inferential standpoint?

Modern DiD: some open challenges

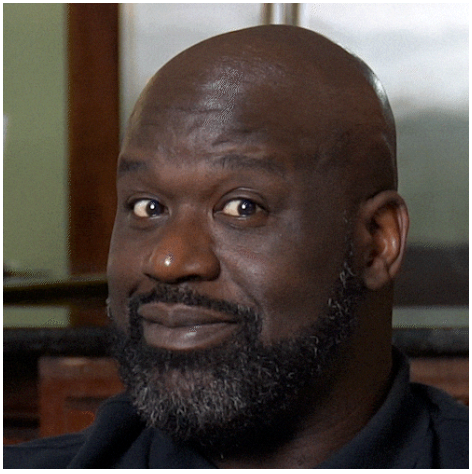
- How to effectively explore pre-treatment periods
  - Some DiD estimators do not explore multiple pre-treatment periods.

    Pre-treatment information not informative about post-treatment: not easy to motivate empirically.
  - Some DiD estimators implicitly assume that all pre-treatment information is equally important.

    Data from 10 periods ago is always as informative as from one period ago: not easy to motivate empirically.
  - There must be a better way to do this, but how to adequately choose the form of the weights?

Modern DiD: some open challenges

- **How to improve efficiency and make more informative inferences**
  - When # of units is small, modern DiD estimators may suffer from limited power.
    How should we think of DiD with state-level data?
  - Is it better to assume treatment-effect heterogeneity away and stick to TWFE estimators?
    This is an argument often made in empirical research.
  - Are power-related issues an inherited problem with all heterogeneous robust DiD estimators?
  - In terms of power, is the fundamental choice heterogeneous robust DiD versus TWFE?
  - How can we improve power within DiD setups with fixed-T?



What we do in the paper

What we do in the paper

- Characterize the DiD model as a model based on sequential conditional moment restrictions.
  - Allows us to bridge DiD and more traditional econometric models (Ai and Chen, 2003, 2012)
  - Allows one to explore different estimation procedures that explore the same information.
  - Highlight that DiD models are often overidentified in a nonparametric sense (Chen and Santos, 2018).

What we do in the paper

- Characterize the DiD model as a model based on sequential conditional moment restrictions.
  - Allows us to bridge DiD and more traditional econometric models (Ai and Chen, 2003, 2012)
  - Allows one to explore different estimation procedures that explore the same information.
  - Highlight that DiD models are often overidentified in a nonparametric sense (Chen and Santos, 2018).

- Derive the efficient influence functions (EIF) and the semiparametric efficient bounds for ATT and ES type parameters, given the DiD identification assumptions.
  - Characterize the smallest asymptotic variance or tightest possible (asymptotic) confidence intervals, among all estimators that are root-*n* consistent and asymptotically normal, without making additional assumptions.
  - Provide rigorous benchmarks for comparisons across ATT and ES estimators in the DiD framework.
  - Reveal how one should aggregate information about pre-treatment periods when the goal is to gain inference precision.

What we do in the paper

- Propose semiparametric efficient estimators that are easy to compute and dominate the existing estimators in terms of power.
- Propose simple Hausman tests for the plausibility of parallel trends when the model is overidentified.
- Our results are derived for a variety of setups:
  - Single treatment date and staggered treatment adoption;
  - Covariates may or may not play an important role for identification;
  - Extension to instrumented DiD setups.

Practical takeaway messages

Our practical takeaway messages

- DiD models are usually nonparametrically overidentified:
  off-the-shelf DML/DR/reg using $Y = Y_{t=\text{post}} - Y_{t=\text{pre}}$ as outcome is generally not semipar. efficient
- Our EIFs are available in closed form and are automatically orthogonal moments:
  - Provide a blueprint for efficient DiD estimators
- Gains in efficiency can be of first-order according to our simulations and empirical application
- Optimal to weight pre-treatments and comparison groups in a non-uniform manner
- When covariates are required for identification, the efficient weights also depend on covariate values.
  - For example, optimal weights for men may differ from those for women, as the outcomes for these two covariate groups may have heterogeneous correlations over time.

Some remarks

- We focus on “fixed-T, large n” panels.
- We expect our semiparametric efficient estimators to work best when $\sqrt{n} >> T$.
  - PT may not be as plausible when T is large.
  - When T is large, there are probably better tools available.
- The degree of semiparametric efficiency gains depends on the degree of overidentifying restrictions, such as the number of parallel trends across periods.
  - We develop a simple Hausman-style overidentification test to assess the plausibility of PT assumptions. We also have some visual tools
  - You can incorporate linear, quadratic, and other known unit-specific trends, too, if you trust those parametric assumptions.

What we will add to the paper in a few months

- **New practical implications to highlight:**
  - Precision gains from exploiting over-identification
  - Sensitivity analysis across different comparison groups
  - Incremental tests for selecting the comparison group
  - Connection between PT and concerns about spillovers & compositional changes

    Event-time balancing, subsetting to $[t-5, t+5]$ *— our framework makes this easy and assessable*

- **New connections:**
  - Our *J*-test as a diagnostic for PT violations.
  - Our efficient DiD $\longrightarrow$ Armstrong, Kline and Sun (2025)

- **Possible Extension:** Unbalanced panels & repeated cross-sections

  No-compositional-changes assumption $\Rightarrow$ *additional sources of over-identification affecting ES weights*

An Empirical Application

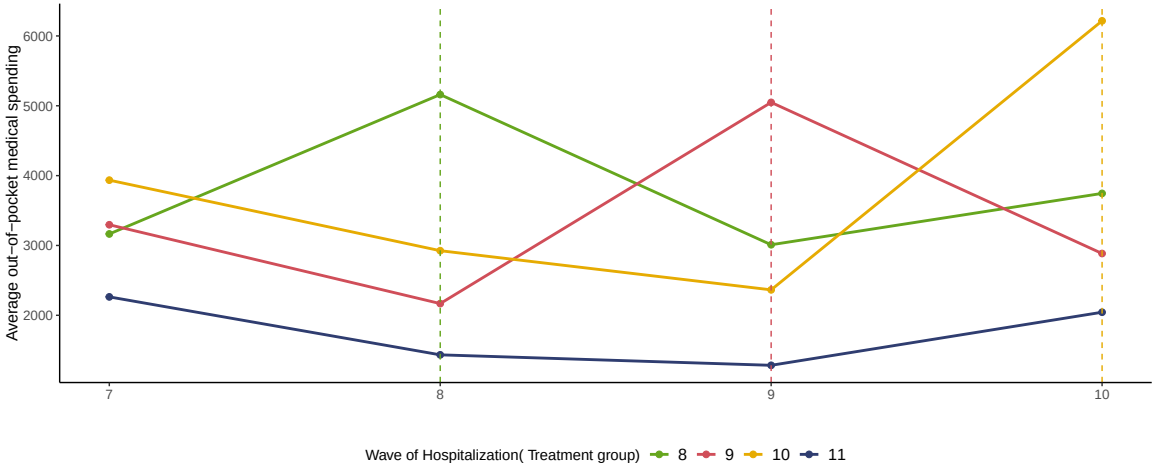
What is causal effect of hospitalization on out-of-pocket medical spending?

- Dobkin, Finkelstein, Kluender and Notowidigdo (2018) study the effect of hospitalization on out-of-pocket medical spending and several other outcomes.
- They explore the variation in hospitalization time across individuals and use a DiD and ES strategy to estimate the causal effects of hospitalization.
- We follow the sample construction of Sun and Abraham (2021), which explores the publicly available survey data from the Health and Retirement Study (HRS) from the replication package of Dobkin et al. (2018):
  - Adults hospitalized at ages 50–59, excluding pregnancy-related admissions.
  - Balanced panel spanning waves 7–11 (2004–2012).
  - Final Sample: 652 individuals in 4 treatment groups
  - $G_i = 8$ (252), $G_i = 9$ (176), $G_i = 10$ (163), $G_i = \infty$ (65; re-labeled the $G_i = 11$ as "never-treated").

Parameters of Interest and Different DiD Estimators

- Parameters of Interest: Post-treatment effects $ATT(g,t)$ and aggregated event-study measures $ES(e)$, ES_{avg}.
- Estimates calculated using multiple DiD methods:
  - Our Efficient DiD estimator (EDiD)
  - Callaway and Sant'Anna (2021) and Sun and Abraham (2021) estimator using never-treated as comparison group (CS-SA)
  - Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2020) estimators using not-yet-treated as comparison group (CS-dCDH)
  - Borusyak et al. (2024), Gardner (2021), and Wooldridge (2021) imputation estimators (BJS-G-W).
- As PT is plausible in this context (Dobkin et al., 2018 and Sun and Abraham, 2021), all point estimates are expected to be similar.
- Our efficient DiD estimator is expected to deliver gains in efficiency.

Visualize the data for each cohort



Point estimates are indeed similar across estimators

Estimator	$ATT(8,8)$	$ATT(8,9)$	$ATT(8,10)$	$ATT(9,9)$	$ATT(9,10)$	$ATT(10,10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	3072	1112	1038	3063	90	2908	3024	692	1038	1585
	(806)	(637)	(817)	(690)	(641)	(894)	(486)	(471)	(816)	(521)
CS-SA	2826	825	800	3031	107	3092	2960	530	800	1430
	(1035)	(909)	(1008)	(702)	(651)	(995)	(539)	(585)	(1008)	(647)
CS-dCDH	3029	1248	800	3324	107	3092	3134	779	800	1571
	(913)	(861)	(1008)	(959)	(651)	(995)	(536)	(570)	(1008)	(566)
BJS-G-W	3029	1285	1021	3239	77	2758	3017	788	1021	1609
	(916)	(767)	(851)	(862)	(729)	(957)	(555)	(587)	(851)	(582)

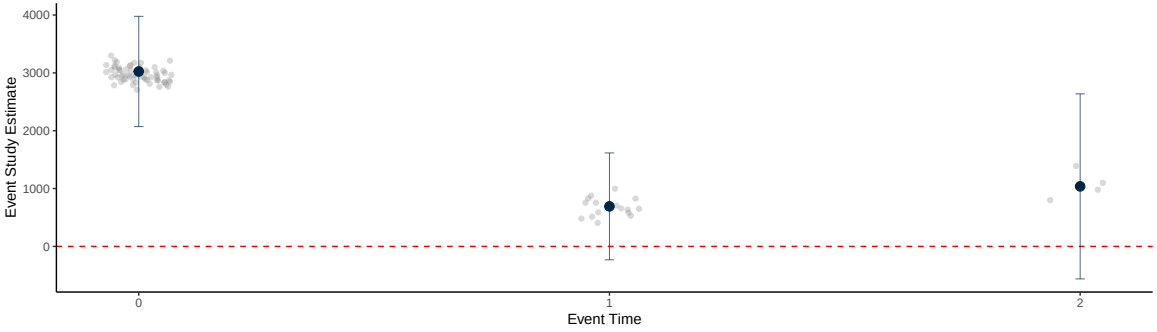
Our Efficient DiD provides substantial gains in efficiency

- Estimates of the asymptotic relative efficiency (ARE) of our proposed efficient DiD estimator with respect to the other available DiD estimators.
- Heuristically, ARE provides a relative measure of sample size needed for other DiD estimators to achieve the same precision as our efficient DiD estimator.

Estimator	$ATT(8,8)$	$ATT(8,9)$	$ATT(8,10)$	$ATT(9,9)$	$ATT(9,10)$	$ATT(10,10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CS-SA	1.65	2.04	1.52	1.04	1.03	1.24	1.23	1.54	1.52	1.54
CS-dCDH	1.28	1.83	1.52	1.93	1.03	1.24	1.21	1.46	1.52	1.18
BJS-G-W	1.29	1.45	1.09	1.56	1.29	1.15	1.30	1.55	1.09	1.25

- Not possible to rank the other DiD estimators in terms of ARE in the application:
  - BJS-G-W: Efficient under homoskedasticity and no residual serial correlation (Wooldridge, 2021; Borusyak et al., 2024).

Visualizing Event-Study stability



Understanding the Efficient Weights for ATT(g,t)'s

Let’s dig into the HOW

DiD Framework

- "Short" panel data $\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i)'\}_{i=1}^n$ for n large and T finite fixed.
- Treatment is binary, may have different starting dates, and is an absorbing state.
  - Treatment does not "turn off".
- $Y_{i,t}(\mathbf{0_{g-1}}, \mathbf{1_{T-g+1}}) \equiv Y_{i,t}(g)$: Potential outcomes indexed by treatment starting date g.
  - Potential outcome for unit i and time t if they were treated for the first time in period g.
- G_i denotes the time unit i is first-treated, with $G_i = \infty$ if they stay untreated by time T.
  - $G_i \in \mathcal{G} \subseteq \{2, \ldots, T, \infty\}$: no unit treated in period $t = 1$
  - With single treatment date at period $t = g$, $G_i = g$ (treated units) or $G_i = \infty$ (untreated units).
  - $\mathcal{G}_{\text{trt}} = \mathcal{G} \setminus \{\infty\}$ denote the support of G among eventually treated units.
  - Let $G_g = 1\{G = g\}$.

Target Parameters

- The group-time average treatment effects for the treated:

$$ATT(g,t) := \mathbb{E}[Y_t(g) - Y_t(\infty)|G = g]$$

Average treatment effect at time period t of being first treated in period g compared to never being treated, among units that indeed start treatment at time g.

- $ATT(g,t)$: can track down how treatment effect varies with elapsed treatment time (aka event-time), $e = t - g$, for each group g.

- With multiple groups, can sum over the $ATT(g,t)$'s to get an aggregated event-study

$$ES(e) := \mathbb{E}\left[ATT\left(G, G+e\right) \middle| G + e \in [1,T]\right] = \sum_{g \in \mathcal{G}_{\text{trt}}} \mathbb{P}(G = g | G + e \in [1,T]) ATT(g, g+e).$$

$$ES_{\text{avg}} := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} ES(e), \qquad \text{with } |\mathcal{E}| \text{ the number of elements in } \mathcal{E} = t - G, t \geq G.$$

Maintained Assumption: Sampling, Overlap, and No-anticipation

Assumption (Maintained Assumption (M))

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

(ii) *(O) For each* $g \in \mathcal{G}$, $\mathbb{E}[G_g|X] \in (0,1)$ *almost surely (a.s.).*

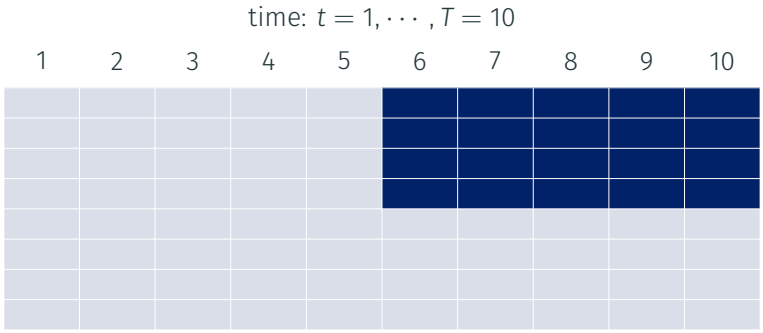
(iii) *(NA) For every* $g \in \mathcal{G}_{trt}$, *and every pre-treatment periods* $t < g$, $\mathbb{E}[Y_{i,t}(g)|G = g, X] = \mathbb{E}[Y_{i,t}(\infty)|G = g, X]$ *almost surely.*

- The maintained Assumption M is not enough to identify ATT or ES type parameters.
- DiD methods impose parallel trends assumptions to identify these parameters: we will discuss them in a bit.

DiD with Single Treatment Time

No variation in treatment timing

- Single treatment period at time g: $G_i = g$ (treated) or $G_i = \infty$ (untreated).
- $ES(e) = ATT(g, g + e)$.



Parallel Trends Assumption: Post-treatment periods

Assumption (PT in the post-treatment periods)

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

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = \infty, X] \text{ a.s.}$$

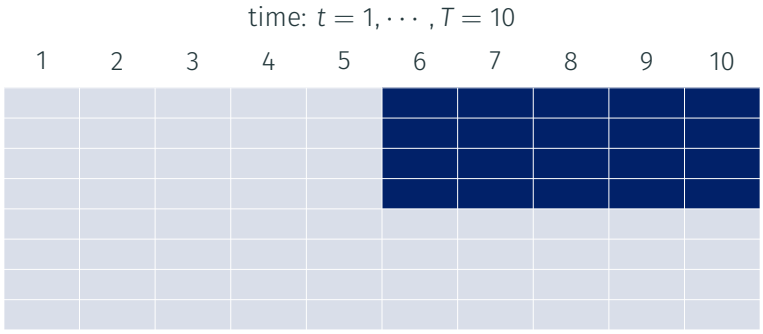
- Impose parallel trends only for post-treatment periods.
- Uses only period $g - 1$ as the baseline.
- Without covariates, easy to show that, for any $t \geq g$,

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{g-1}|G = g] - \mathbb{E}[Y_t - Y_{g-1}|G = \infty].$$

- Limitation: If we were gifted 10 more pre-treatment data periods, the estimand for $ATT(g, t)$ would not be allowed to use any of that.

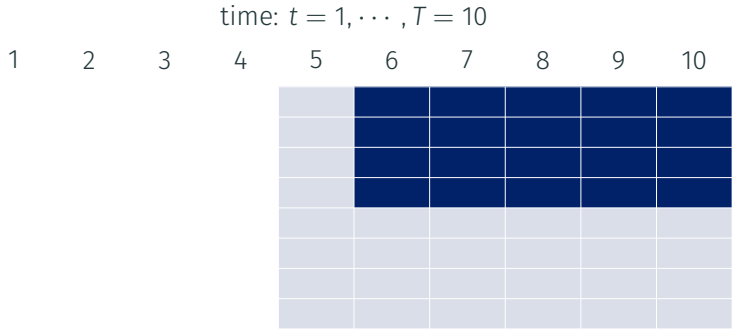
PT-Post: Implications

- $ES(e) = ATT(g, g+e)$ for any $e \geq 0$



PT-Post: Implications

- $ES(e) = ATT(g, g+e)$ for any $e \geq 0$



Parallel Trends Assumption: All periods

Assumption (PT in all periods)

For each $t \in \{2, \ldots, T\}$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = \infty, X] \text{ a.s.}$$

- Impose parallel trends in all periods.
- Allows us to use any pre-treatment period as a baseline.
- Without covariates, easy to show that for any $t \geq g$ and any $t' < g$,

$$ATT(g,t) = \mathbb{E}[Y_t - Y_{t'}|G = g] - \mathbb{E}[Y_t - Y_{t'}|G = \infty].$$

- If we were gifted 10 more pre-treatment periods of data, we could easily use all of them to compute $ATT(g,t)$.
- How to do that efficiently, such that we maximize precision?

Characterization of the DiD model based on seq. conditional moment restrictions

Lemma (Moment-restrictions for over-identified DiD with single treatment time)

The family of prob. dist. of $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ *satisfying Assumptions M and PT-All-g are observationally equivalent to the family of prob. dist. of* $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ *satisfying Assumption M(i)(ii), and the set of moment restrictions: for all* $t \in \{g, \ldots, T\}$*, with prob. one,*

$$\mathbb{E}[G_g(ATT(g,t) - CATT(g,t,X))] = 0,$$

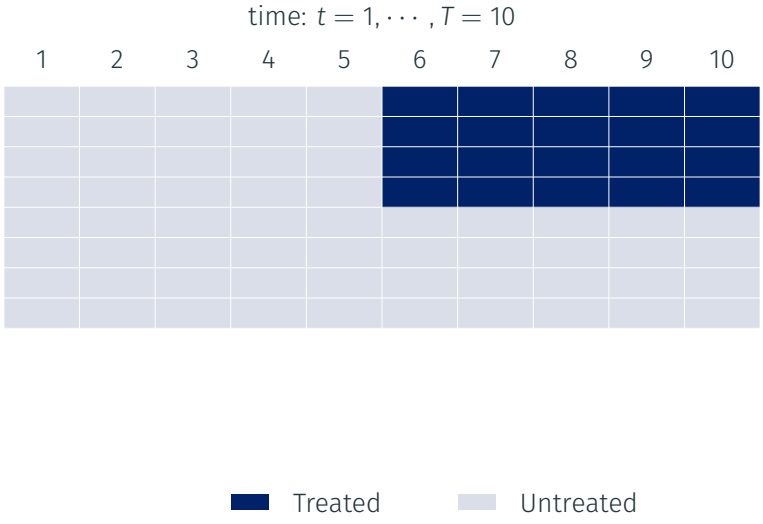
$$\mathbb{E}\left[CATT(g,t,X) - \frac{G_g(Y_t - Y_{g-1})}{p_g(X)} + \frac{G_\infty(Y_t - Y_{g-1})}{p_\infty(X)} \middle| X\right] = 0,$$

$$\mathbb{E}\left[\frac{G_g(Y_{t'} - Y_1)}{p_g(X)} - \frac{G_\infty(Y_{t'} - Y_1)}{p_\infty(X)} \middle| X\right] = 0, \text{ for all } 2 \leq t' \leq g-1,$$

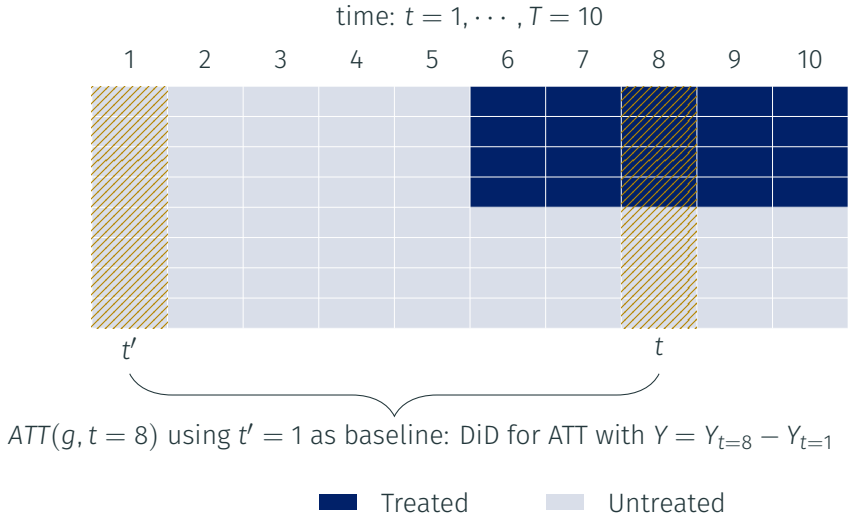
$$\mathbb{E}[G_g - p_g(X)|X] = 0.$$

- Semiparametric efficient bound for $ATT(g,t)$: apply Ai and Chen (2012) for seq. moments.

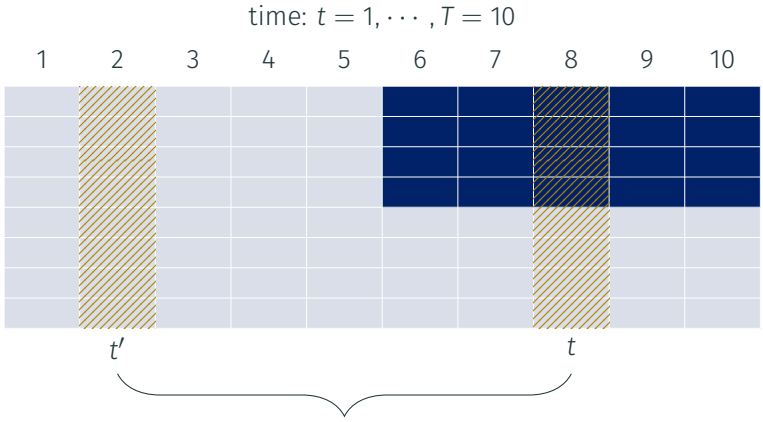
Understanding the sources of over-identification



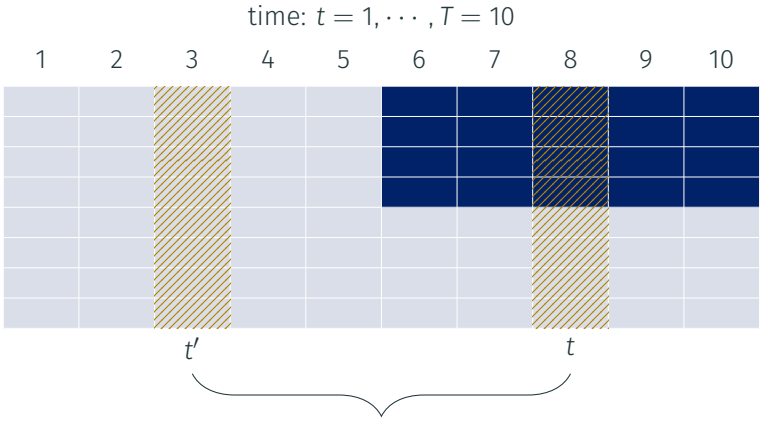
Understanding the sources of over-identification



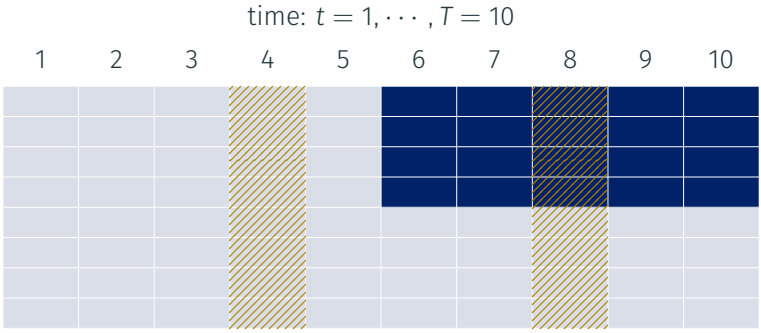
Understanding the sources of over-identification



Understanding the sources of over-identification



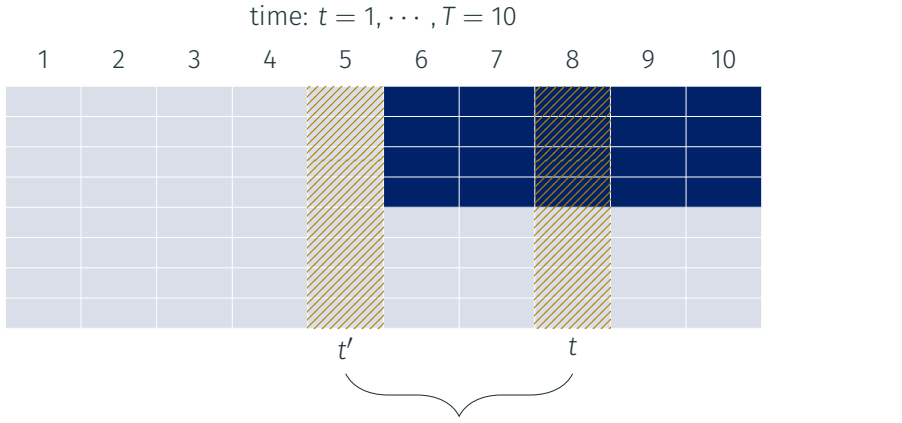
Understanding the sources of over-identification



$ATT(g, t = 8)$ using $t' = 4$ as baseline: DiD for ATT with $Y = Y_{t=8} - Y_{t=4}$

Treated Untreated

Understanding the sources of over-identification



Intuition on how to get semiparametric efficiency

- For each $1 \leq t' \leq g-1$, we fix the baseline period at t', and compute the "efficient influence function" for $ATT(g,t)$ as-if there were only 2 groups, $G = g$ and $G = \infty$, and two periods, t (post-treatment) and t' (pre-treatment)
  - Akin to compute the "DR scores" in DML language.
- Stack all the non-collinear influence functions into a vector, $\mathbb{IF}^{att(g,t)}$.
- Compute the covariance of $\mathbb{IF}^{att(g,t)}$ given covariates, $V_{gt}(X) = \text{Cov}\left(\mathbb{IF}^{att(g,t)}|X\right)$.
- Efficient Influence Function for $ATT(g,t)$ is given by

$$EIF^{att(g,t)} = \frac{1'V_{gt}(X)^{-1}}{1'V_{gt}(X)^{-1}1}\mathbb{IF}^{att(g,t)}.$$

- Next, we explore these results to obtain EIF-based estimands for $ATT(g,t)$, which serve as a blueprint for efficient estimation.

Using EIF as a blueprint for estimating ATT(g,t)

- The key is to explore that $\mathbb{E}[EIF^{att(g,t)}] = 0$ to get IF-based estimand:

$$ATT(g,t) = \mathbb{E}\left[\frac{\mathbf{1}' V_{gt}(X)^{-1}}{\mathbf{1}' V_{gt}(X)^{-1}\mathbf{1}} \theta_{g,t}(W)\right], \tag{1}$$

where $p_g(X) = \mathbb{E}[G_g|X]$, $\theta_{g,t}(W) = (\theta_{g,t,1}(W), \ldots, \theta_{g,t,g-1}(W))'$ is a $(g-1)\times 1$ column vector with

$$\theta_{g,t,t'}(W) = \frac{1}{\mathbb{P}(G=g)}\left(G_g - \frac{(1-G_g)p_g(X)}{1-p_g(X)}\right)(Y_t - Y_{t'} - \mathbb{E}[Y_t - Y_{t'}|G=\infty, X]).$$

- Here, $\theta_{g,t}(W)$ is a vector of DR DiD "integrands", each being computed pretending we were in the 2×2 DiD setup of Sant'Anna and Zhao (2020).

- Efficient DiD estimators: apply plug-in principle, or do DML.

DiD with Single Treatment Time

What about the weights?

Understanding the weights across pre-treatment periods

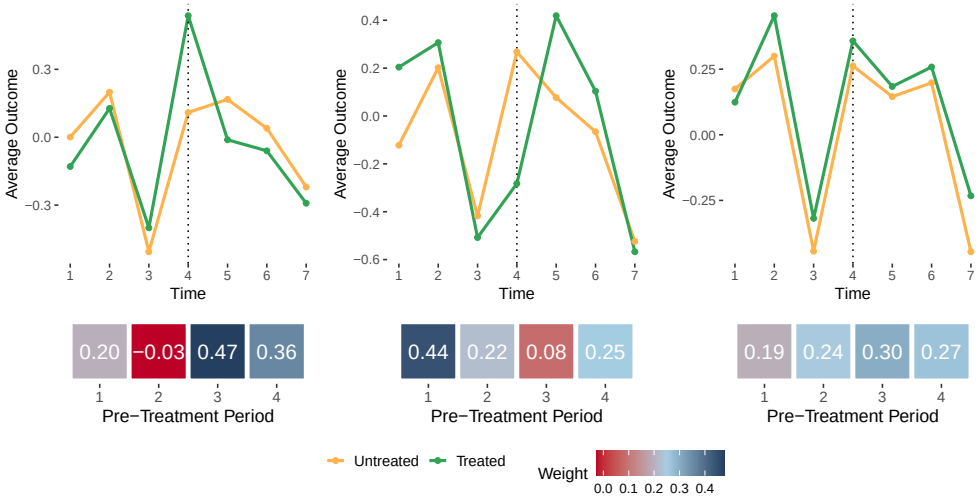
- The optimal way to aggregate across pre-treatment periods to learn about $ATT(g,t)$ is given by the inverse of $V_{gt}(X)$, the conditional covariance of the influence functions.
- This is OK, but it does not really help us better understand these weights.
- In the paper, we show that one can use the "easier-to-understand" matrix $V^*_{gt}(X)$ to form the optimal weights, where $dim(V^*_{gt}(X)) = dim(V_{gt}(X))$ and the (j,k)-th element being $V^*_{gt}(X)$ is given by

$$\frac{1}{p_g(X)} \text{Cov}(Y_t - Y_j, Y_t - Y_k | G = g, X) + \frac{1}{1 - p_g(X)} \text{Cov}(Y_t - Y_j, Y_t - Y_k | G = \infty, X). \tag{2}$$

Some important takeaways

- Weights depend on the covariance of outcome changes for each treatment group;
- Weights vary with the time of the ATT of interest, t.
- Weights are covariate-dependent (different from GMM)
- Weights are generally not constant across pre-treatment periods.
- EIF is obtained by optimally weighting the individual IFs
- Most DiD and ES estimators are not semiparametrically efficient:
  - Either only looks at a single pre-treatment t':
    DTWFE, de Chaisemartin and D'Haultfœuille (2020, 2024), Callaway and Sant'Anna (2021), Sun and Abraham (2021).
  - or taking simple average over $t' < g$:
    TWFE, Wooldridge (2021), Gardner (2021),Borusyak et al. (2024), Lee and Wooldridge (2023)

Plots for the three different DGPs



DiD with Single Treatment Time

What if PT holds only in post-treatment periods?

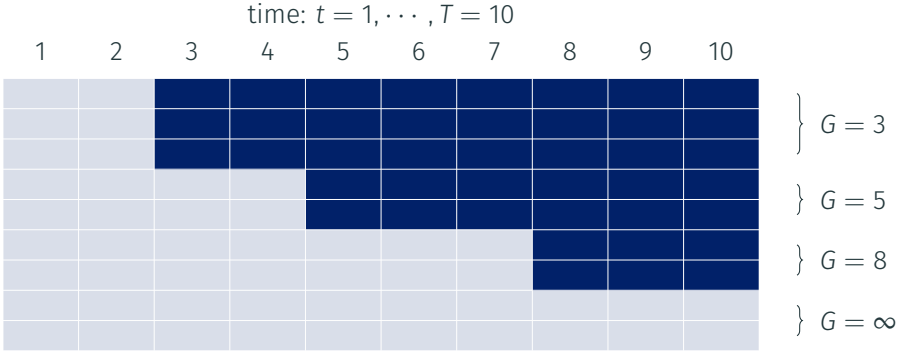
Corollary 1: PT holding only in post-treatment periods

- Under PT-Post, we can only use the period immediately before treatment $t' = g - 1$.
- The model is just-identified; the efficient influence function is simply $\mathbb{IF}_{g-1}^{att(g,t)}$.
- This generalizes the EIF in Sant'Anna and Zhao (2020), who focused on the much simpler two-period model.

DiD with Staggered Treatment Adoption

Staggered Adoption

- Multiple treatment starting periods, leading to several treatment groups defined by treatment starting date.
- Each group has their $ATT(g, t)$.



PT for all periods

Assumption (Parallel Trends for all groups and periods (PT-All))

For each $t \in \{2, \ldots, T\}$ *and* $(g, g') \in \mathcal{G}_{trt} \times \mathcal{G}$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = g', X] \text{ a.s.}$$

i.e., conditional on covariates, the average evolution of untreated potential outcomes is the same across treatment groups, in all available periods.

- Two sources of nonparametric over-identification (in the sense of Chen and Santos (2018)): multiple baseline periods and multiple comparison groups.

Characterization of the DiD model based on seq. conditional moment restrictions

Lemma (Moment-restrictions for over-identified staggered DiD)

The family of prob. dist. of $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ *satisfying Assumptions M and PT-All are observationally equivalent to the family of prob. dist. of* $(Y_{t=1} \cdots, Y_{t=T}, X', G)$ *satisfying Assumption M(i)(ii) and the set of moment restrictions:*
for all $g, g' \in \mathcal{G}_{trt} \times \mathcal{G}_{trt}$ *and post-treatment periods* $t \in \{g, \ldots, T\}$*, with probability one,*

$$\mathbb{E}[G_g(ATT(g,t) - CATT(g,t,X))] = 0,$$

$$\mathbb{E}\left[CATT(g,t,X) - \frac{G_g(Y_t - Y_{g-1})}{p_g(X)} + \frac{G_\infty(Y_t - Y_{g-1})}{p_\infty(X)}\Big|X\right] = 0,$$

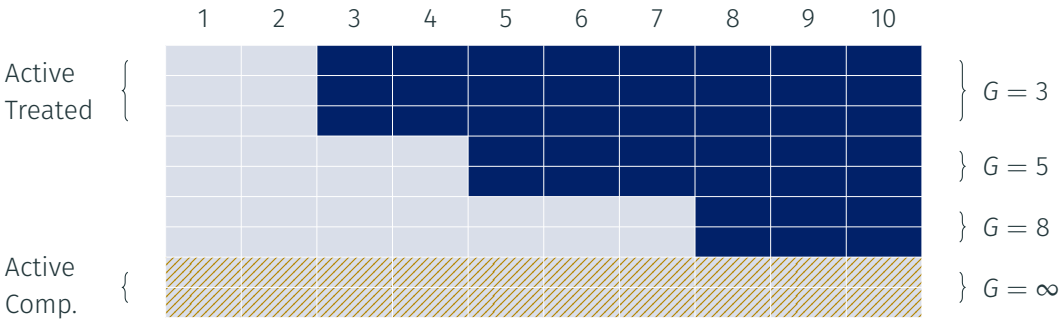
$$\mathbb{E}\left[\frac{G_{g'}(Y_{t'} - Y_1)}{p_{g'}(X)} - \frac{G_\infty(Y_{t'} - Y_1)}{p_\infty(X)}\Big|X\right] = 0, \text{for all } 2 \le t' \le g' - 1,$$

$$\mathbb{E}[G_g - p_g(X)|X] = 0.$$

- Semiparametric efficient bound for $ATT(g,t)$: apply Ai and Chen (2012) for seq. moments.

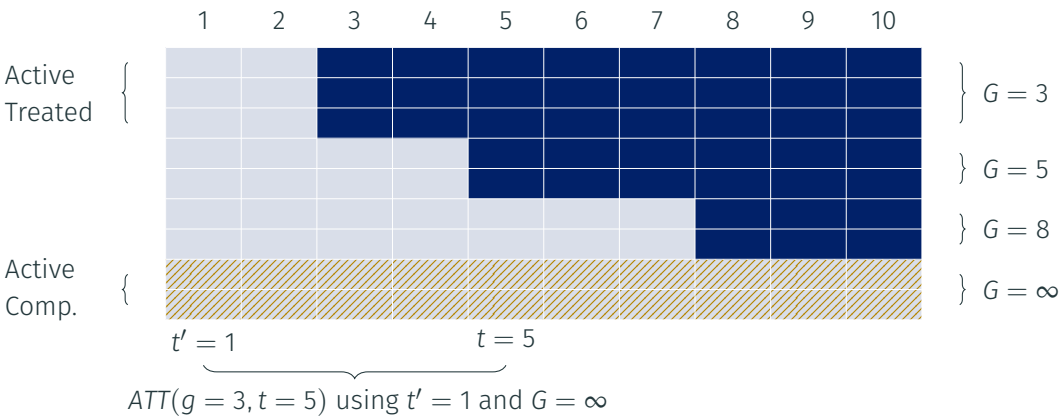
Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



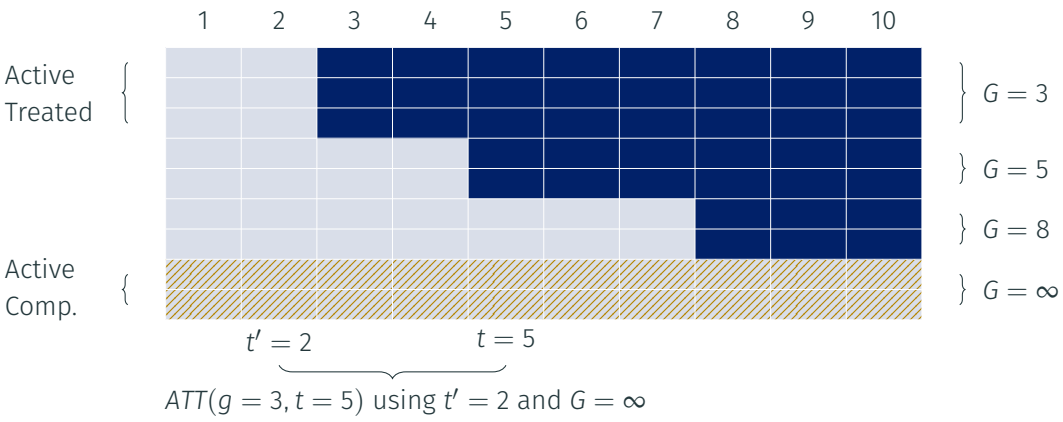
Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



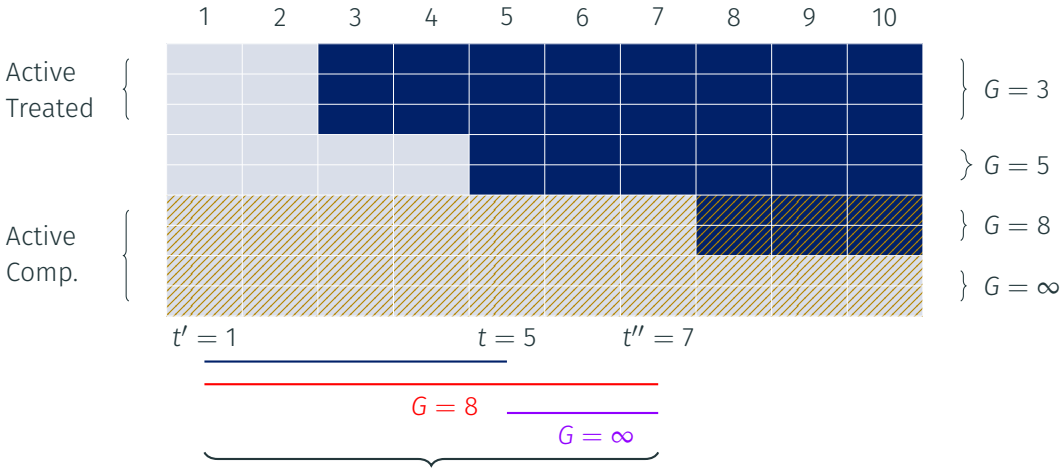
Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using one active comparison group $G = \infty$:



Understanding the sources of over-identification

$ATT(g = 3, t = 5)$, using two active comparison groups $G = 8$, and $G = \infty$:



Lemma (Over-identification in staggered designs)

Under Assumptions M and PT-All, for every group $(g, g') \in \mathcal{G}_{trt} \times \mathcal{G}_{trt}$ *and time periods* $(t, t', t'') \in \mathcal{T} \times \mathcal{T} \times \mathcal{T}$ *such that* $t \geq g$, $g > t'$, *and* $g' > \max\{t', t''\}$, *with probability one,*

$$CATT(g,t,X) = \underbrace{\mathbb{E}[Y_t - Y_{t'} | G = g, X]}_{\equiv m_{g,t,t'}(X)} - \Big(\underbrace{\mathbb{E}[Y_t - Y_{t''} | G = \infty, X]}_{\equiv m_{\infty,t,t''}(X)} + \underbrace{\mathbb{E}[Y_{t''} - Y_{t'} | G = g', X]}_{\equiv m_{g',t'',t'}(X)} \Big), \tag{3}$$

and, as a consequence,

$$ATT(g,t) = \mathbb{E}\big[Y_t - Y_{t'} \big| G = g\big] - \mathbb{E}\big[\left(m_{\infty,t,t''}(X) + m_{g',t'',t'}(X)\right) \big| G = g\big]. \tag{4}$$

More generally, for any covariate-specific weights $w^{g,t}_{g',t',t''}(X)$ *that sum up to one, we have that*

$$ATT(g,t) = \mathbb{E}\left[\sum_{(g',t',t'') \in \mathcal{H}^{g,t}} w^{g,t}_{g',t',t''}(X) \left[m_{g,t,t'}(X) - \left(m_{\infty,t,t''}(X) + m_{g',t'',t'}(X)\right)\right] \Big| G = g \right].$$

Semiparametric Efficient Staggered DiD

- We will now fix $t' = 1$ for simplicity (no loss of generality).
- For $g' \in \mathcal{G}_{\text{trt}}$ and $1 \leq t'' \leq g' - 1$, let $\pi_g = \mathbb{E}[G_g]$ and

$$\begin{aligned}\mathbb{IF}^{att(g,t)}_{g',t''} = &\frac{1}{\pi_g}\Big(G_g((m_{g,t,1}(X) - m_{\infty,t,t''}(X) - m_{g',t'',1}(X)) - ATT(g,t))\Big) \\ &+ \frac{G_g}{\pi_g}(Y_t - Y_1 - m_{g,t,1}(X)) \\ &- \frac{G_\infty}{\pi_g}\frac{p_g(X)}{p_\infty(X)}(Y_t - Y_{t''} - m_{\infty,t,t''}(X)) - \frac{G_{g'}}{\pi_g}\frac{p_g(X)}{p_{g'}(X)}(Y_{t''} - Y_1 - m_{g',t'',1}(X)). \qquad (5)\end{aligned}$$

- We then stack all the non-collinear IF terms to form $\mathbb{IF}^{att(g,t)}_{\text{stg}}$.

Efficient Staggered DiD

- Here follows the efficient staggered DiD estimand

$$ATT(g,t) = \mathbb{E}\left[\frac{\mathbf{1}'\Omega_{gt}(X)^{-1}}{\mathbf{1}'\Omega_{gt}(X)^{-1}\mathbf{1}}\theta_{\text{stg}}^{\text{att(g,t)}}(W)\right], \tag{6}$$

where $\Omega_{gt}(X) = \text{Cov}(\mathbb{1}\mathbb{F}_{\text{stg}}^{att(g,t)}|X)$, $\theta_{\text{stg}}^{\text{att(g,t)}}(W)$ is the column vector

$$\theta_{\text{stg}}^{\text{att(g,t)}}(W) = (\theta_{g'}^{\text{att(g,t)}}(W)', g' \in \mathcal{G}_{\text{trt}})', , \tag{7}$$

such that, for $g' = g$,

$$\theta_{g'}^{\text{att(g,t)}}(W) = (\theta_{g,1}^{\text{att(g,t)}}(W), \ldots, \theta_{g,g-1}^{\text{att(g,t)}}(W))',$$

and, for $g' \neq g$,

$$\theta_{g'}^{\text{att(g,t)}}(W) = (\theta_{g',2}^{\text{att(g,t)}}(W), \ldots, \theta_{g',g'-1}^{\text{att(g,t)}}(W))',$$

with

$$\begin{aligned}\theta_{g',t''}^{\text{att(g,t)}}(W) = &\frac{1}{\pi_g}G_g(Y_t - Y_1 - m_{\infty,t,t''}(X) - m_{g',t'',1}(X)) \\ &- \left(\frac{G_\infty}{\pi_g}\frac{p_g(X)}{p_\infty(X)}(Y_t - Y_{t''} - m_{\infty,t,t''}(X)) + \frac{G_{g'}}{\pi_g}\frac{p_g(X)}{p_{g'}(X)}(Y_{t''} - Y_1 - m_{g',t'',1}(X))\right).\end{aligned} \tag{8}$$

Semiparametric efficiency result

Theorem (Efficient DiD with staggered treatment adoption)

Under Assumptions M and PT-All, the efficient influence function for $ATT(g,t)$, $t \geq g$, is given by

$$\mathbb{E}\mathbb{I}\mathbb{F}_{stg}^{att(g,t)} = \frac{1'\Omega_{gt}(X)^{-1}}{1'\Omega_{gt}(X)^{-1}1}\mathbb{I}\mathbb{F}_{stg}^{att(g,t)}.$$

The semiparametric efficient variance bounds are obtained as the second moments of the efficient influence functions, provided they are finite.

See our paper for the efficient influence function of a $ES(e)$, $e \geq 0$.

Conclusion

Our practical takeaway messages

- DiD models are usually nonparametrically overidentified:
  off-the-shelf DML/DR/reg using $Y = Y_{t=\text{post}} - Y_{t=\text{pre}}$ as outcome is generally not semipar. efficient
- Our EIFs are available in closed form and are automatically orthogonal moments:
  - Provide a blueprint for efficient DiD estimators
- Gains in efficiency can be of first-order according to our simulations and empirical application
- Optimal to weight pre-treatments and comparison groups in a non-uniform manner
- When covariates are required for identification, the efficient weights also depend on covariate values.
  - For example, optimal weights for men may differ from those for women, as the outcomes for these two covariate groups may have heterogeneous correlations over time.

Thanks!

pedro.santanna@emory.edu

psantanna.com

@pedrohcgs

@pedrosantanna.bsky.social

Appendix

Monte Carlo Simulations

Monte Carlo Simulations

Simulations calibrated with CPS data

Empirically-calibrated simulations using CPS data

- We consider an empirically-calibrated DGP that builds on Arkhangelsky et al. (2021), which is a setup with single treatment date.
- We differ from Arkhangelsky et al. (2021) in four aspects:
  1. We consider heterogeneous treatment effects across units, though $ATT = 0$;
  2. Consistent with traditional DiD setup with "large n" and fixed T, we consider short panels with $T = 7$—four pre-treatment and three post-treatment periods;
  3. We do not limit the maximum number of treated units in a given simulation;
  4. All our outcomes are measured in log to avoid violating support restrictions in the simulation.
- Apart from these differences, the construction of our DGP follows from Arkhangelsky et al. (2021).
- Target parameter: $ES_{\text{avg}} = \dfrac{ATT(5,5) + ATT(5,6) + ATT(5,7)}{3}$.

Details of the DGP

CPS simulations: Alternative DiD estimators

- We compare our efficient DiD estimator with several alternatives.
- Let $E_{i,t} = t - G_i$. For $G_i = \infty$, $E_{i,t} = 0$.
- OLS estimates from the standard TWFE specification (TWFE)

$$Y_{i,t} = \alpha_t + \eta_i + D_{i,t}\beta + \varepsilon_{i,t} \tag{9}$$

- Average of post-treatment OLS estimates of β_e from a Dynamic TWFE specification (DTWFE):

$$Y_{i,t} = \alpha_t + \eta_i + \sum_{e \neq -1} 1[E_{i,t} = e]\beta_e^{avg} + \varepsilon_{i,t}, \tag{10}$$

- Synthetic DiD (SDiD) of Arkhangelsky et al. (2021): Weight untreated units and time periods more flexibly,

$$\hat{\beta}^{sdid} = \arg \min_{\mu,\alpha,\eta,\beta} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - (\mu + \alpha_t + \eta_i + \beta D_{it})\right)^2 \hat{w}_i^{sdid} \hat{\lambda}_t^{sdid} \right\}. \tag{11}$$

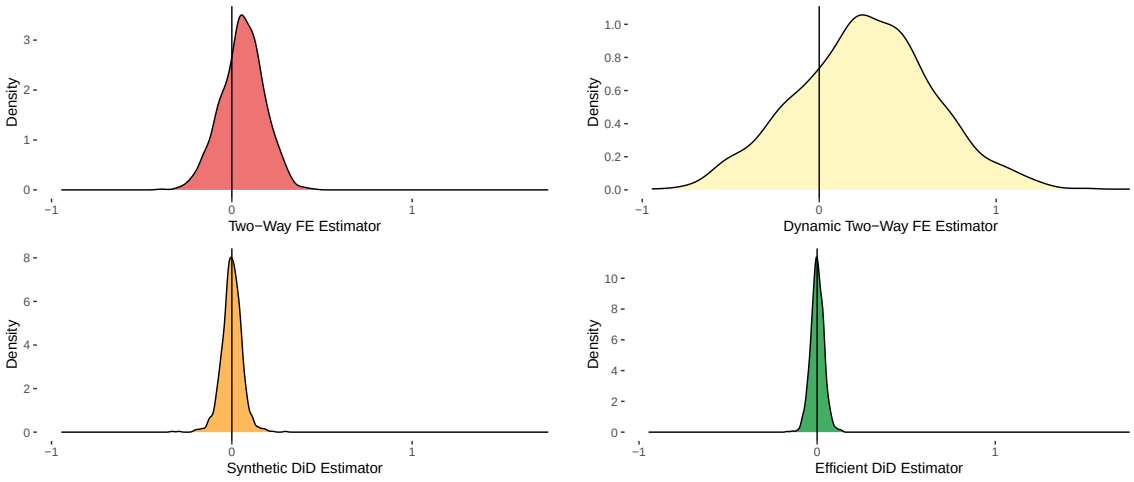
  Implemented using the default in their **R** package.

Monte Carlo results: Relative RMSE and Bias

		Relative RMSE				Bias ($\times 10$)			
	Sample size	EDiD	TWFE	DTWFE	SDiD	EDiD	TWFE	DTWFE	SDiD
1. Baseline	50	1	3.57	12.46	1.53	0.01	0.60	2.58	0.00
	200	1	2.32	3.37	1.95	0.00	-0.01	0.00	0.00
Outcome Model									
2. No corr	50	1	3.52	12.33	1.45	0.02	0.59	2.46	0.00
	200	1	2.27	3.33	1.95	-0.01	0.00	-0.01	0.00
3. No M	50	1	3.67	13.04	1.49	-0.02	0.61	2.45	0.03
	200	1	2.17	3.00	1.68	-0.01	0.02	0.00	0.01
4. No F	50	1	1.47	1.88	1.42	0.00	-0.02	-0.04	-0.02
	200	1	1.64	2.18	1.63	0.00	0.00	-0.01	0.00
Treatment Assignment									
6. Gun law	50	1	7.27	18.23	4.67	0.00	-0.08	-0.23	-0.11
	200	1	8.70	12.94	6.05	0.00	0.03	0.02	0.02
7. Abortion	50	1	6.99	17.19	4.77	-0.01	0.55	1.75	0.33
	200	1	8.04	12.64	5.23	0.00	-0.01	-0.05	0.00
Outcome Variable									
9. Ln Hours	50	1	1.01	1.92	0.95	-0.34	0.12	1.53	0.02
	200	1	1.24	1.95	1.21	0.01	0.05	0.02	0.07
10. Ln U-rate	50	1	0.82	1.44	0.82	0.73	-0.24	-0.36	-0.26
	200	1	1.03	1.53	1.01	0.03	0.00	0.01	0.00

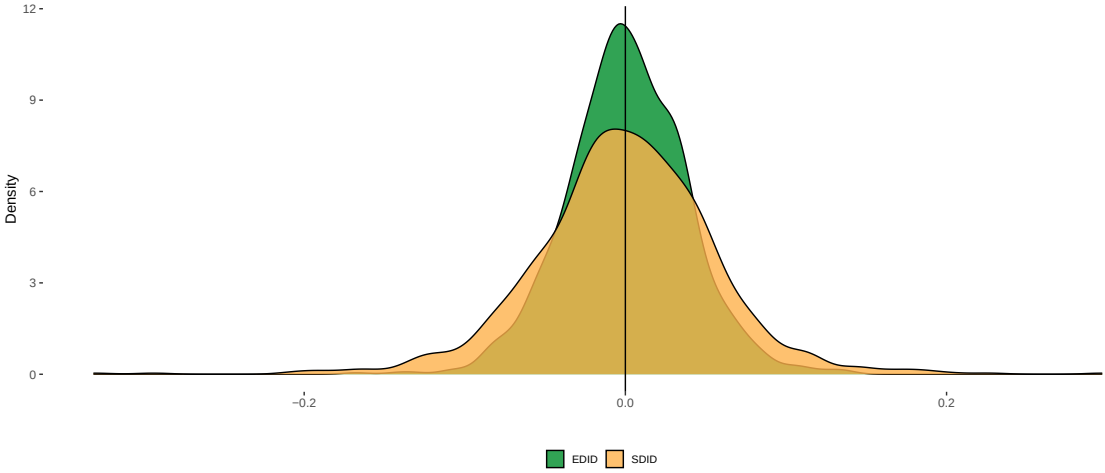
Monte Carlo results: Baseline DGP with $n = 50$

Figure 1: Monte Carlo for Baseline DGP: $n = 50$



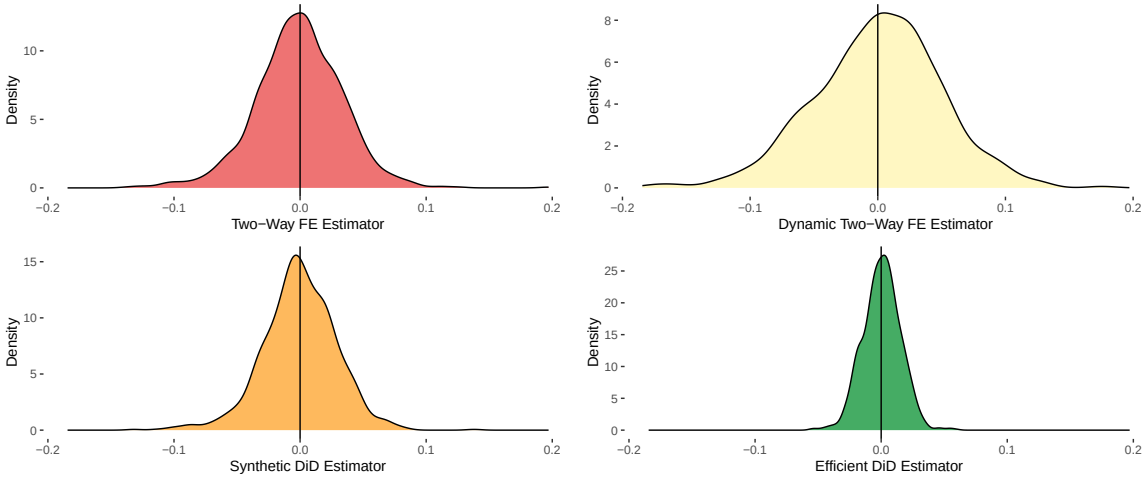
Monte Carlo results: Baseline DGP with $n = 50$

Figure 2: Monte Carlo for Baseline DGP: $n = 50$



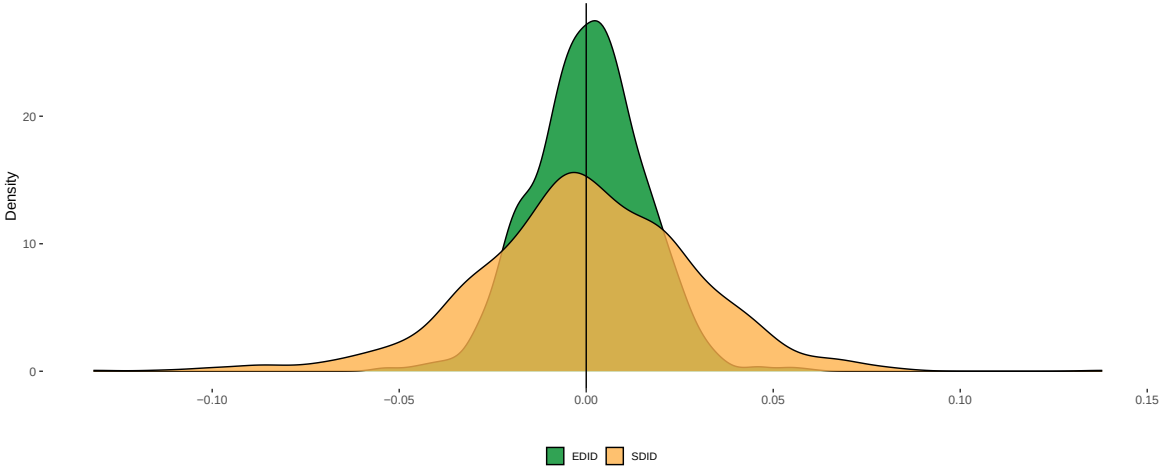
Monte Carlo results: Baseline DGP with $n = 200$

Figure 3: Monte Carlo for Baseline DGP: $n = 200$



Monte Carlo results: Baseline DGP with $n = 200$

Figure 4: Monte Carlo for Baseline DGP: $n = 200$



Monte Carlo results: Coverage

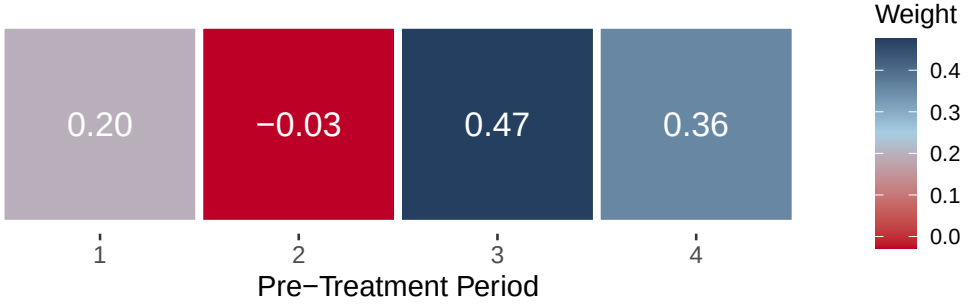
		Bootstrap (300 rep.)				Analytical			
	Sample size	EDiD	TWFE	DTWFE	SDiD	EDiD	TWFE	DTWFE	SDiD
1. Baseline	50	0.94	0.92	0.93	0.93	0.80	0.92	0.92	-
	200	0.96	0.97	0.97	0.97	0.94	0.98	0.97	-
Outcome Model									
2. No corr	50	0.92	0.94	0.94	0.93	0.80	0.93	0.92	-
	200	0.93	0.97	0.97	0.96	0.93	0.97	0.97	-
3. No M	50	0.94	0.93	0.93	0.94	0.82	0.92	0.92	-
	200	0.93	0.97	0.97	0.96	0.91	0.96	0.97	-
4. No F	50	0.94	0.94	0.94	0.95	0.82	0.94	0.92	-
	200	0.95	0.95	0.96	0.96	0.94	0.95	0.96	-
Treatment Assignment									
6. Gun law	50	0.95	0.94	0.94	0.97	0.94	0.94	0.94	-
	200	0.94	0.96	0.96	0.97	0.94	0.96	0.96	-
7. Abortion	50	0.94	0.92	0.91	0.93	0.92	0.93	0.91	-
	200	0.95	0.98	0.98	0.97	0.95	0.97	0.98	-
Outcome Variable									
9. Ln Hours	50	0.93	0.94	0.93	0.95	0.77	0.94	0.92	-
	200	0.94	0.95	0.96	0.96	0.93	0.95	0.95	-
10. Ln U-rate	50	0.93	0.91	0.92	0.92	0.79	0.90	0.91	-
	200	0.95	0.94	0.95	0.95	0.92	0.94	0.94	-

Monte Carlo results: Relative length of confidence intervals

		Bootstrap (300 rep.)				Analytical			
	Sample size	EDiD	TWFE	DTWFE	SDiD	EDiD	TWFE	DTWFE	SDiD
1. Baseline	50	1.00	3.63	12.58	1.43	1.00	5.44	18.40	-
	200	1.00	2.46	3.60	2.08	1.00	2.65	3.86	-
Outcome Model									
2. No corr	50	1.00	3.67	12.78	1.36	1.00	5.46	18.57	-
	200	1.00	2.47	3.58	2.11	1.00	2.67	3.85	-
3. No M	50	1.00	3.56	12.90	1.32	1.00	5.32	18.83	-
	200	1.00	2.33	3.39	1.84	1.00	2.52	3.64	-
4. No F	50	1.00	1.42	1.85	1.44	1.00	2.14	2.72	-
	200	1.00	1.58	2.18	1.57	1.00	1.71	2.35	-
Treatment Assignment									
6. Gun law	50	1.00	6.41	16.94	4.33	1.00	7.00	18.12	-
	200	1.00	8.86	13.18	5.92	1.00	8.99	13.33	-
7. Abortion	50	1.00	6.40	16.44	3.84	1.00	7.14	18.00	-
	200	1.00	8.53	13.30	5.44	1.00	8.76	13.53	-
Outcome Variable									
9. Ln Hours	50	1.00	1.05	2.01	1.03	1.00	1.58	2.95	-
	200	1.00	1.37	2.07	1.33	1.00	1.48	2.21	-
10. Ln U-rate	50	1.00	0.75	1.34	0.79	1.00	1.13	1.98	-
	200	1.00	0.99	1.51	0.99	1.00	1.07	1.62	-

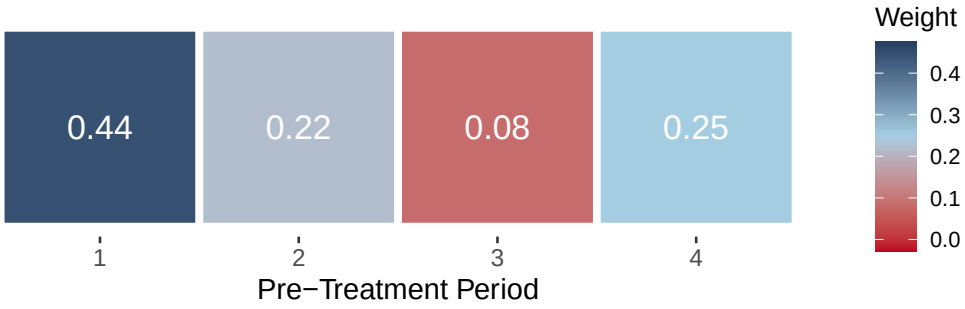
Weights: DGP 1

Figure 5: Contribution of pre-treatment periods for the Efficient DiD estimator for ES_{avg}: DGP 1 Baseline



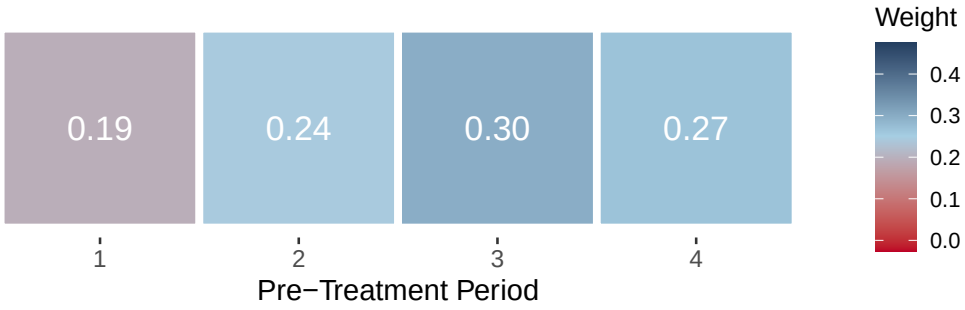
Weights: DGP 3

Figure 6: Contribution of pre-treatment periods for the Efficient DiD estimator for ES_{avg}: DGP 3 Outcome Model with no M



Weights: DGP 9

Figure 7: Contribution of pre-treatment periods for the Efficient DiD estimator for ES_{avg} - DGP 9 Ln Hours as outcome



Monte Carlo Simulations

Simulations calibrated with Compustat data with staggered treatment

Empirically-calibrated simulations using Compustat data

- We consider an empirically-calibrated DGP that builds on Baker, Larcker and Wang (2022), a setup with staggered treatment dates.
- We differ from Baker et al. (2022) in three aspects:
  1. We sample $n = 400$ firms and follow them for $T = 11$ years—consistent with traditional DiD setup with "large n" and fixed T;
  2. We vary the serial autocorrelation ρ of the error term to allow for richer outcome dynamics;
  3. We assign firms to treatment groups instead of assigning firms to states and states to treatment groups (observational equivalent from a random sample perspective).
- Apart from these differences, the DGP follows from Baker et al. (2022):
  - 3 treatment dates; all units eventually treated.
  - Dynamic treatment effects with heterogeneous trends.

Details of the DGP

Some simulations: Alternative Staggered DiD estimators

- Parameter of interest: average of the post-treatment event-study

$$ES_{\text{avg}} = \frac{\sum_{e=0}^{5} ES(e)}{6}.$$

- $ES_{\text{avg}} = 0.589$ in our simulations.
- We compare our efficient DiD (EDiD) with popular Staggered DiD estimators:
  - Callaway and Sant'Anna (2021) and Sun and Abraham (2021) (CS-SA) based on never-treated comparison group.
  - Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2020) (CS-dCDH) based on not-yet-treated comparison group.
  - Borusyak et al. (2024), Gardner (2021), and Wooldridge (2021) (BJS-G-W) imputation DiD estimator.

Monte Carlo results: Relative RMSE and Bias

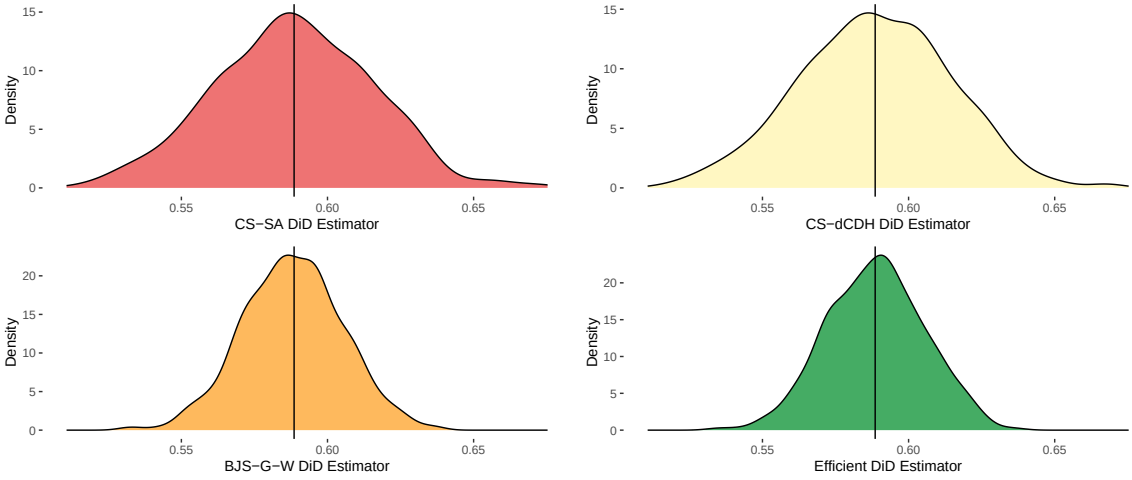
	Relative RMSE				Bias ($\times 10$)			
ρ	E-DiD	BJS-G-W	CS-SA	CS-dCDH	E-DiD	BJS-G-W	CS-SA	CS-dCDH
$\rho = 0$	1	1.02	1.62	1.56	0.00	-0.01	0.00	0.00
$\rho = 0.5$	1	1.06	1.28	1.24	0.01	0.01	0.00	0.00
$\rho = 1$	1	1.19	1.09	1.04	0.04	-0.02	-0.01	-0.01
$\rho = 1.1$	1	1.56	1.28	1.23	0.07	0.05	0.04	0.04
$\rho = -0.5$	1	1.05	2.39	2.30	0.00	0.00	-0.01	0.00
$\rho = -1$	1	1.62	3.24	3.43	-0.01	-0.01	0.00	0.00
$\rho = -1.1$	1	2.22	3.36	3.76	0.00	0.01	0.00	0.01

Monte Carlo results: Coverage and Relative Length of Confidence Interval

	Coverage (analytical)				Relative Length of CI			
ρ	EDiD	BJS-G-W	CS-SA	CS-dCDH	EDiD	BJS-G-W	CS-SA	CS-dCDH
$\rho = 0$	0.96	0.96	0.96	0.96	1	1.04	1.64	1.57
$\rho = 0.5$	0.95	0.96	0.96	0.96	1	1.08	1.31	1.26
$\rho = 1$	0.94	0.96	0.94	0.94	1	1.26	1.12	1.08
$\rho = 1.1$	0.95	0.95	0.95	0.95	1	1.56	1.29	1.24
$\rho = -0.5$	0.94	0.95	0.94	0.94	1	1.08	2.42	2.31
$\rho = -1$	0.94	0.96	0.96	0.95	1	1.67	3.38	3.52
$\rho = -1.1$	0.94	0.95	0.95	0.95	1	2.26	3.46	3.85

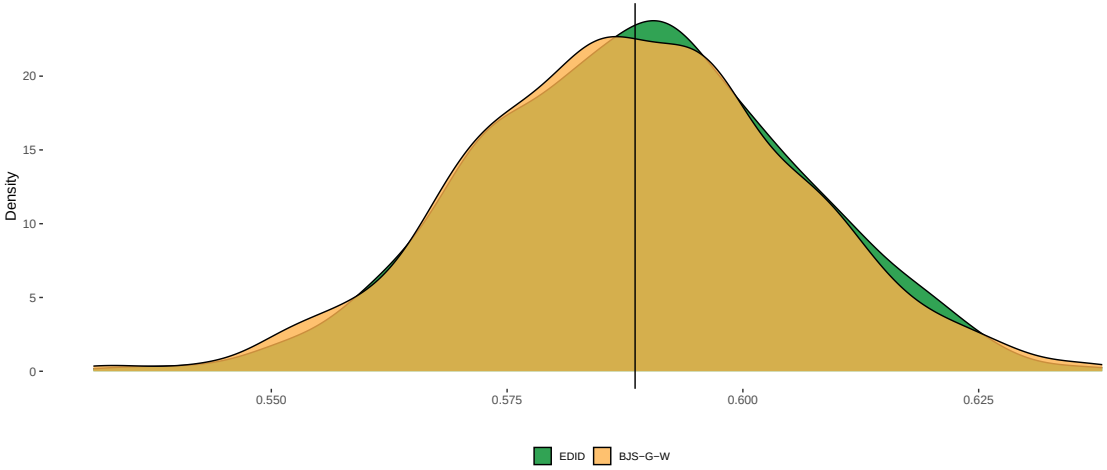
Monte Carlo results: Staggered design with $\rho = 0$

Figure 8: Monte Carlo for Baker et al. DGP: $\rho = 0$



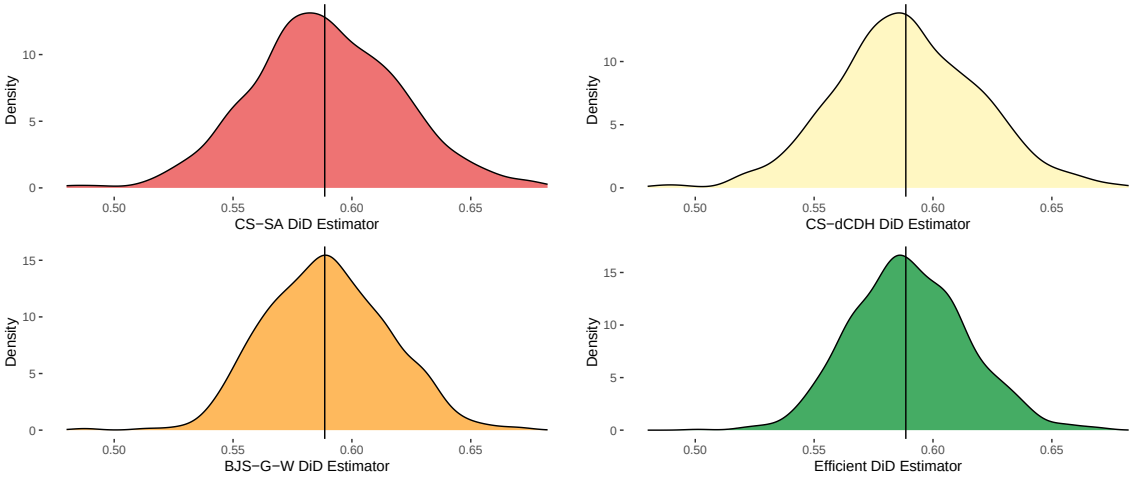
Monte Carlo results: Staggered design with $\rho = 0$

Figure 9: Monte Carlo for Baker et al. DGP: $\rho = 0$



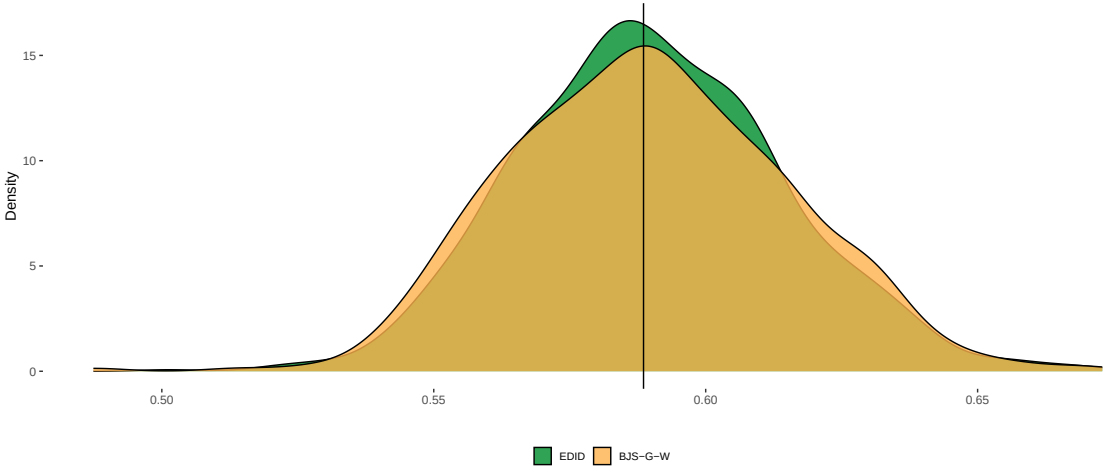
Monte Carlo results: Staggered design with $\rho = 0.5$

Figure 10: Monte Carlo for Baker et al. DGP: $\rho = 0.5$



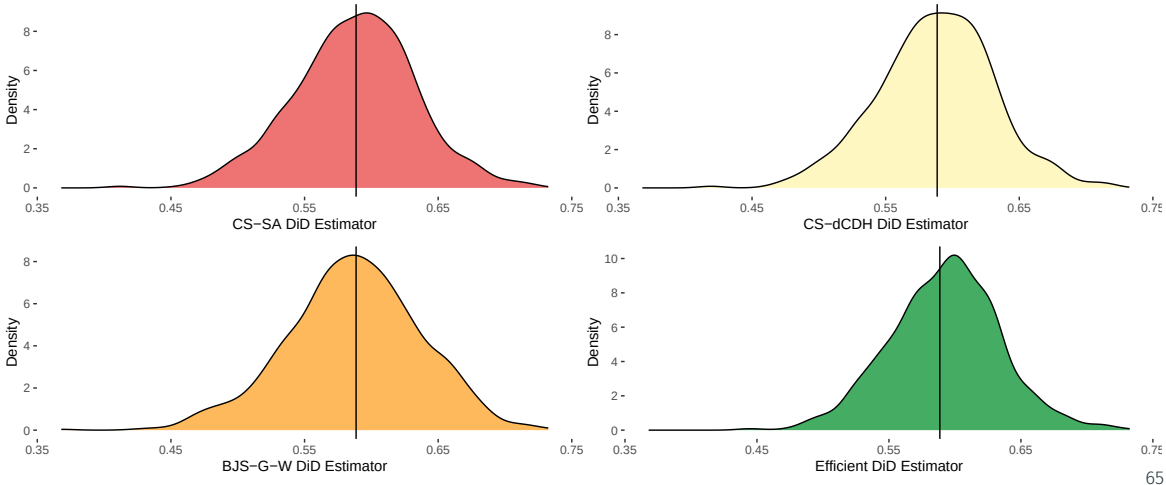
Monte Carlo results: Staggered design with $\rho = 0.5$

Figure 11: Monte Carlo for Baker et al. DGP: $\rho = 0.5$



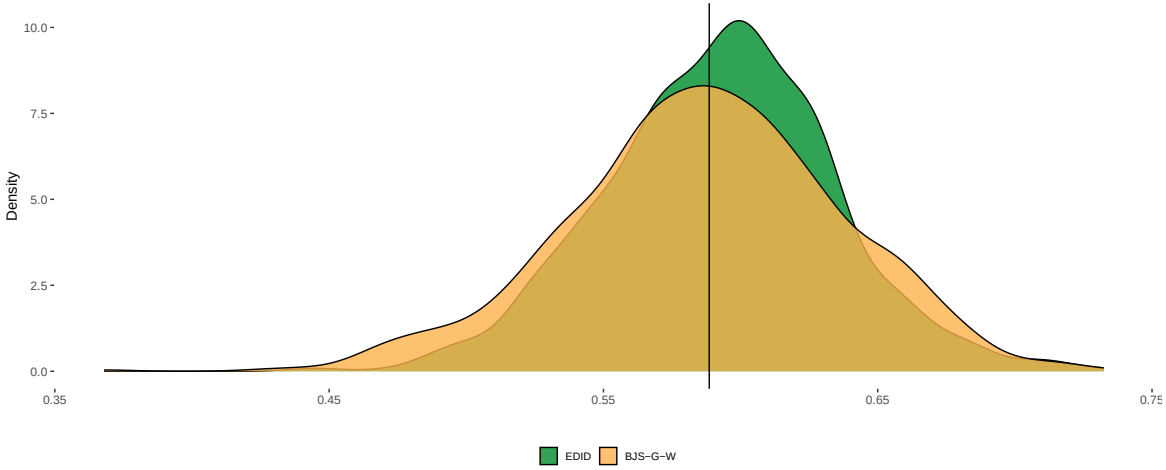
Monte Carlo results: Staggered design with $\rho = 1$

Figure 12: Monte Carlo for Baker et al. DGP: $\rho = 1$



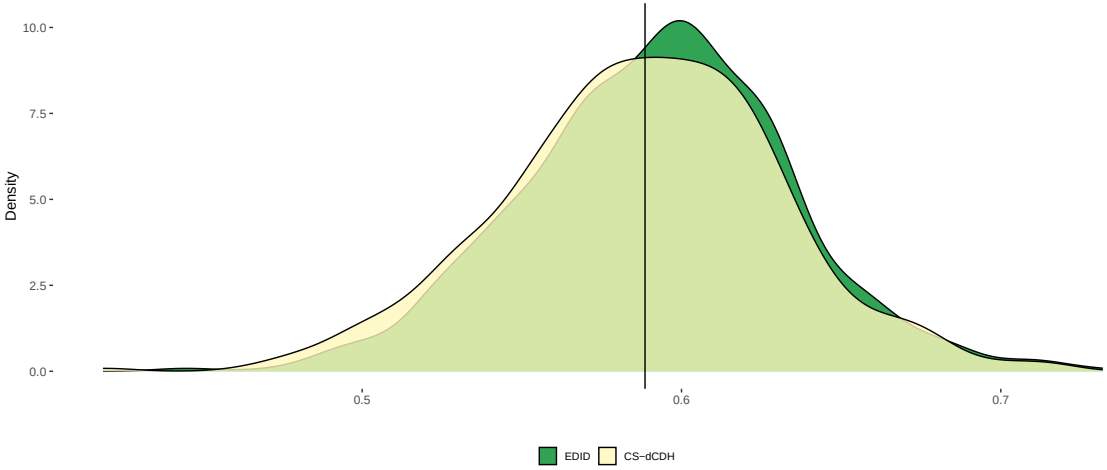
Monte Carlo results: Staggered design with $\rho = 1$

Figure 13: Monte Carlo for Baker et al. DGP: $\rho = 1$



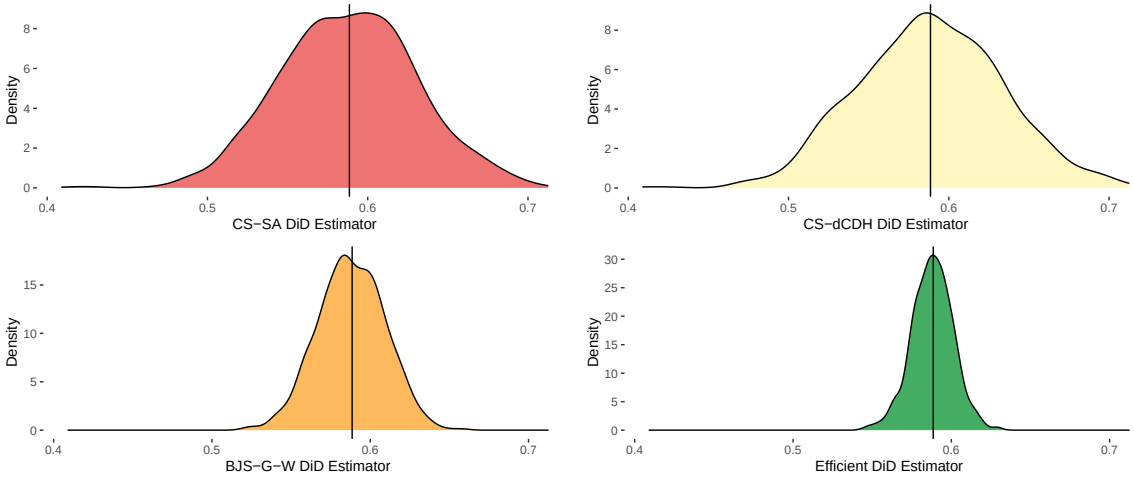
Monte Carlo results: Staggered design with $\rho = 1$

Figure 14: Monte Carlo for Baker et al. DGP: $\rho = 1$



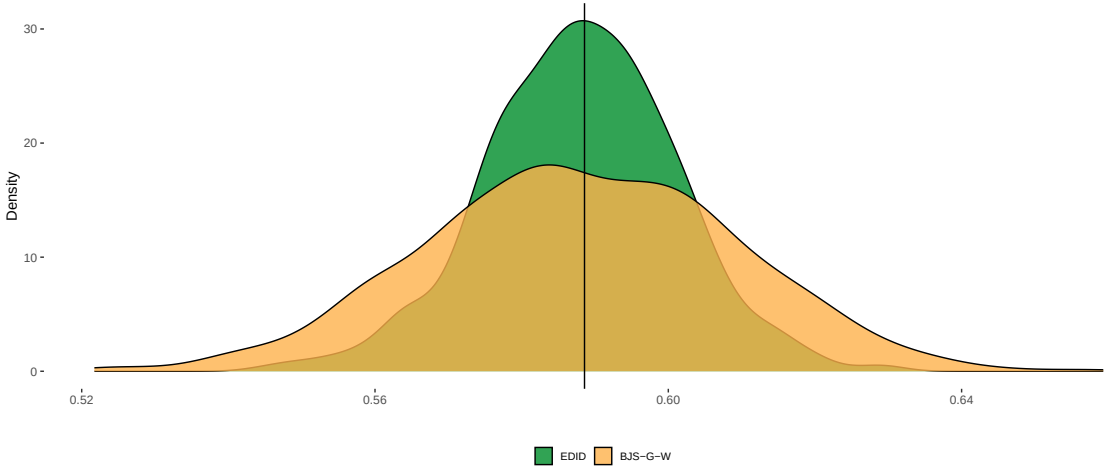
Monte Carlo results: Staggered design with $\rho = -1$

Figure 15: Monte Carlo for Baker et al. DGP: $\rho = -1$



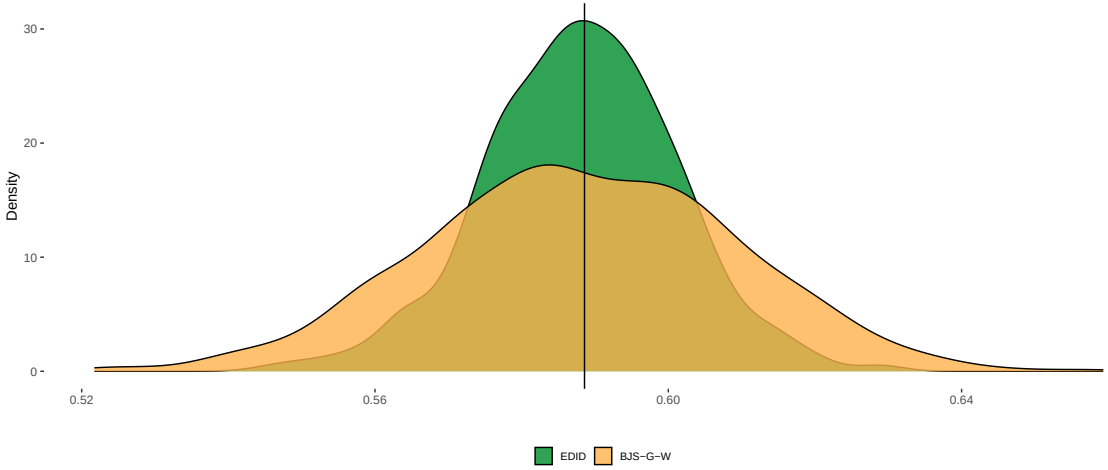
Monte Carlo results: Staggered design with $\rho = -1$

Figure 16: Monte Carlo for Baker et al. DGP: $\rho = -1$



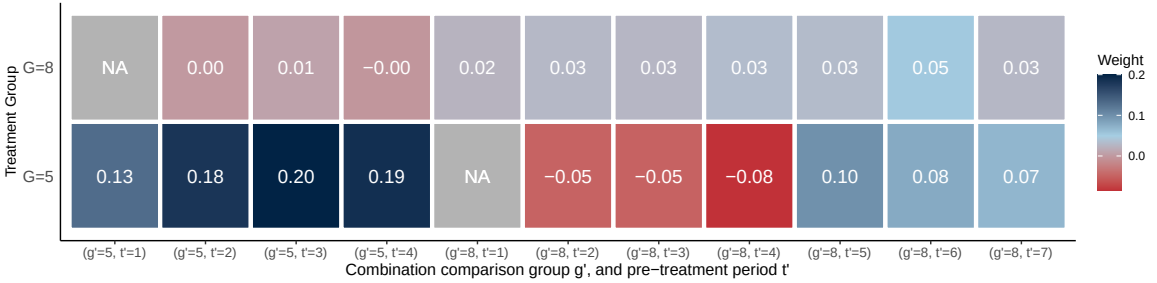
Monte Carlo results: Staggered design with $\rho = -1$

Figure 17: Monte Carlo for Baker et al. DGP: $\rho = -1$



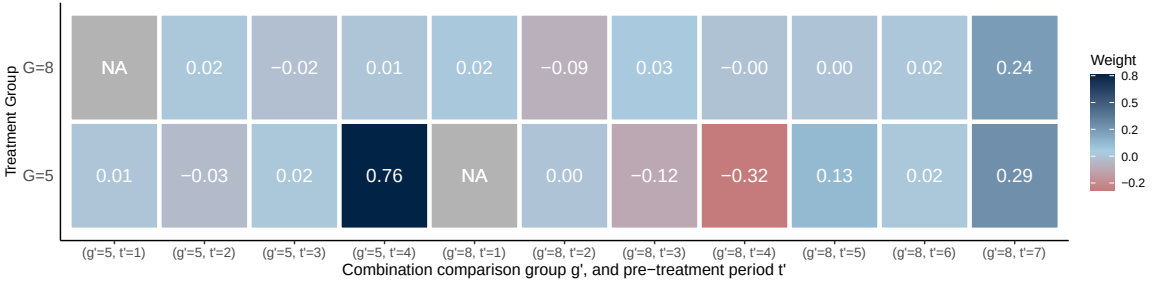
Weights: DGP with $\rho = 0$

Figure 18: Contribution of treatment and comparison groups, and pre-treatment periods for the Efficient DiD estimator for ES_{avg}: DGP with $\rho = 0$



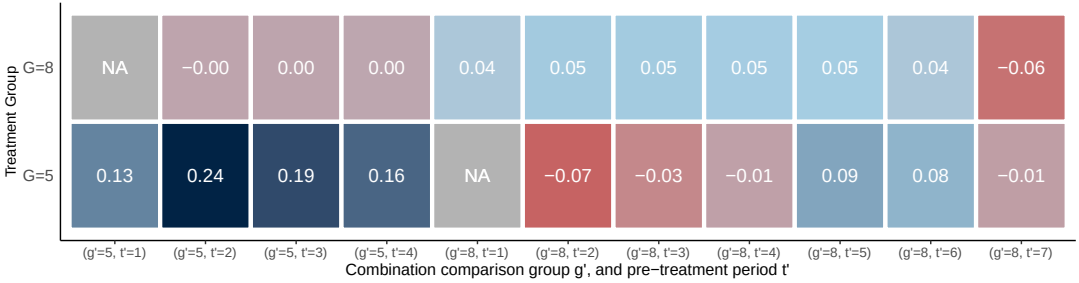
Weights: DGP with $\rho = 1$

Figure 19: Contribution of treatment and comparison groups, and pre-treatment periods for the Efficient DiD estimator for ES_{avg}: DGP with $\rho = 1$



Weights: DGP with $\rho = -1$

Figure 20: Contribution of treatment and comparison groups, and pre-treatment periods for the Efficient DiD estimator for ES_{avg}: DGP with $\rho = -1$



Estimands based on Influence Function

- Let $\pi_g := \mathbb{E}[G_g]$; $m_{g,t,t'}(X) := \mathbb{E}[Y_t - Y_{t'}|G = g, X]$.
- For $1 \leq t' \leq g-1$, let

$$\mathbb{IF}_{t'}^{att(g,t)} = \frac{1}{\pi_g}\Big(G_g\left(m_{g,t,t'}(X) - m_{\infty,t,t'}(X) - ATT(g,t)\right)\Big) + \frac{1}{\pi_g}\Big(G_g\left(Y_t - Y_{t'} - m_{g,t,t'}(X)\right) - \frac{p_g(X)G_\infty}{p_\infty(X)}\left(Y_t - Y_{t'} - m_{\infty,t,t'}(X)\right)\Big), \tag{12}$$

- $\mathbb{E}[\mathbb{IF}_{t'}^{att(g,t)}] = 0$ implies that, for any $t' \in \{1, \ldots, g-1\}$,

$$ATT(g,t) = \frac{1}{\pi_g}\,\mathbb{E}\left[\left(G_g - \frac{p_g(X)G_\infty}{p_\infty(X)}\right)\left(Y_t - Y_{t'} - m_{\infty,t,t'}(X)\right)\right]$$

- Under Assumptions M and PT-Post-g, $\mathbb{IF}_{t'=g-1}^{att(g,t)}$ is in fact an efficient influence function for $ATT(g,t)$ using $t' = g-1$ as the baseline period and t as the post-treatment period.
- However, if Assumption PT-All-g holds $\mathbb{IF}_{t'=g-1}^{att(g,t)}$ is no longer the efficient influence function for $ATT(g,t)$.

Semiparametric efficiency bound when PT holds only conditional on covariates

- Denote the $g-1$ column vector that stacks the $g-1$ influence functions $\mathbb{IF}_{t'}^{att(g,t)}$ by

$$\mathbb{IF}^{att(g,t)} = (\mathbb{IF}_1^{att(g,t)}, \mathbb{IF}_2^{att(g,t)}, \cdots, \mathbb{IF}_{g-1}^{att(g,t)})'.$$

Theorem (Efficient DiD with single treatment time)

Under Assumptions M and PT-All-g, the efficient influence function of a $ATT(g,t)$, $t \geq g$, is given by

$$EIF^{att(g,t)} = \frac{\mathbf{1}' V_{gt}(X)^{-1}}{\mathbf{1}' V_{gt}(X)^{-1}\mathbf{1}} \mathbb{IF}^{att(g,t)}, \quad V_{gt}(X) = \text{Cov}\left(\mathbb{IF}^{att(g,t)} | X\right)$$

The semiparametric efficient variance bound for any regular asymptotic linear (RAL) estimator for the $ATT(g,t)$ is given by

$$V_{eff} = \mathbb{E}\left[\left(EIF^{att(g,t)}\right)^2\right]$$

Efficient DiD estimand with single treatment date and covariates

- We can form efficient DiD estimators for the $ATT(g,t)$ by exploring the efficient influence function:

$$ATT(g,t) = \mathbb{E}\left[\frac{\mathbf{1}'V_{gt}(X)^{-1}}{\mathbf{1}'V_{gt}(X)^{-1}\mathbf{1}}\theta_{g,t}(W)\right], \tag{13}$$

where $\theta_{g,t}(W) = (\theta_{g,t,1}(W), \ldots, \theta_{g,t,g-1}(W))'$ is a $(g-1)\times 1$ column vector with

$$\theta_{g,t,t'}(W) = \frac{1}{\pi_g}\left(G_g - \frac{(1-G_g)p_g(X)}{1-p_g(X)}\right)(Y_t - Y_{t'} - m_{\infty,t,t'}(X)).$$

- Optimal weights also depend on the covariates and the covariance of outcome changes across periods and groups.
- This estimand provides guidance on constructing efficient DiD estimators: plug-in principle. We can also do DML, if we want to.

Potential outcomes in the single-treatment date case

- $Y_{i,t}(g)$: Potential outcome of unit i in time period t if it was first-treated in period g:

$$\begin{aligned} Y_{i,t}(\infty) &= \gamma_i v_t' + \varepsilon_{i,t} = \alpha_t + \eta_i + M_{i,t} + \varepsilon_{i,t}, \\ Y_{i,t}(2009) &= Y_{i,t}(\infty) + \tau_i 1\{t \geq 2009\} \end{aligned}$$

- γ_i and v_t' are 4-dimensional vectors of latent unit and time factors
- $\varepsilon_i = (\varepsilon_{i,t=1}, \ldots, \varepsilon_{i,t=T})' \sim N(0, \Sigma)$ with Σ implied by AR(2) process.
- $\tau_i \sim N(0,1)$.
- Treatment status: $D_{i,t} = 1\{t \geq 2009\}1\{G_i = 2009\}$ with

$$1\{G_i = 2009\} \sim \text{Bernoulli } (\pi_i),$$

$$\pi_i = P(G_i = 2009 | \eta_i, M_i) = \frac{\exp(\phi_\eta \eta_i + \phi_M M_i)}{1 + \exp(\phi_\eta \eta_i + \phi_M M_i)}.$$

- Use CPS data to estimate γ_i, v_t', Σ, and ϕ.

Note that we do not impose PT in the design of the simulations.

Go Back

Efficient Estimation and Inference

Estimating the $ATT(g,t)$'s

- We have already derived a "blue-print" for getting efficient DiD estimators for $ATT(g,t)$ using the efficient influence function:

$$ATT(g,t) = \mathbb{E}\left[\frac{\mathbf{1}'\Omega_{gt}(X)^{-1}}{\mathbf{1}'\Omega_{gt}(X)^{-1}\mathbf{1}}\theta_{\text{stg}}^{\text{att(g,t)}}(W)\right],$$

- To estimate it, we need to construct estimators of some nuisance functions:
  - Regression functions of outcome changes for a given group: $\mathbb{E}[Y_{t''} - Y_{t'}|G = g', X]$;
  - Ratios of generalized propensity score: $r_{g,g'}(X) = \dfrac{p_g(X)}{p_{g'}(X)}$;
  - Covariance terms of outcome changes for a given group: $\text{Cov}(Y_t - Y_{t'}, Y_t - Y_{t''}|G = g', X)$.
- Our estimator is doubly robust if we adopt parametric models.
- But we follow a nonparametric route to avoid concerns about misspecification.

Nonparametric estimators for regression and pscore ratio

- Without covariates, estimation is extremely easy:
  A matter of computing sample means and sample covariances.
- With covariates, the estimator is still simple but needs more steps now.
- We propose simple-to-use sieve-based estimators for $\mathbb{E}[Y_{t''} - Y_{t'}|G = g', X]$ and $r_{g,g'}(X)$.
- Sieve-based regression estimators for $\mathbb{E}[Y_{t''} - Y_{t'}|G = g', X]$ are standard (Chen, 2007).
- For the propensity score ratio $r_{g,g'}(X)$, we consider sieve-based estimators of the type

$$\widehat{r}_{g,g'}(X) = \psi^K(X)'\widehat{\beta}_K,$$

  - $\psi^K(x)$ is a K-dimensional vector of flexible transformations of the X such as (tensor products of) cubic B-splines;
  - Sieve-based estimator $\widehat{\beta}_K$ is given by

$$\widehat{\beta}_K := \underset{\beta_K}{\arg\min}\, \mathbb{E}_n \left[G_{g'} \left(\psi^K(X)'\beta_K \right)^2 - 2G_g \left(\psi^K(X)'\beta_K \right) \right].$$

Nonparametric Estimator for Covariance terms

- For the covariance terms used in constructing efficient weights, we propose using a Nadaraya-Watson-type estimator based on kernel smoothing.
  - Let K be a kernel function on the covariates space and $h > 0$ a bandwidth.
  - Denote $K_h(\cdot) = K(\cdot/h)/h$.
- For $x = X_i$, we use the following estimator for each the covariance terms $\text{Cov}(Y_t - Y_{t'}, Y_t - Y_{t''} | G = g', X = x)$:

$$\left(\sum_{i':G_{i'}=g'} K_h(X_{i'} - x)(Y_{i',t} - Y_{i',t'} - \hat{m}_{g',t,t'}(X_{i'}))(Y_{i',t} - Y_{i',t''} - \hat{m}_{g',t,t''}(X_{i'})) \right) \left(\sum_{i':G_{i'}=g'} K_h(X_{i'} - x) \right)^{-1}.$$

- To establish semiparametric efficiency, we only require these covariance terms to be consistently estimated: no rate requirements.
- Once you estimate all these nuisance functions, apply plug-in and get $\widehat{ATT}_{\text{stg}}(g,t)$ and $\widehat{ES}(e)$.

Efficient estimation and inference

Theorem (Semiparametric efficient estimation and inference)

Let Assumptions M, PT-All and the regularity conditions listed in the Appendix hold. Then, our proposed nonparametric estimators $\widehat{ATT}_{stg}(g,t)$ *and* $\widehat{ES}(e)$ *are asymptotically normal, i.e., as* $n \to \infty$,

$$\begin{aligned}\sqrt{n}(\widehat{ATT}_{stg}(g,t) - ATT(g,t)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w^{att(g,t)}(X)' \mathbb{IF}_{stg}^{att(g,t)}(W_i) + o_p(1) \\ &\xrightarrow{d} N(0, \mathrm{Var}(w^{att(g,t)}(X)' \mathbb{IF}_{stg}^{att(g,t)}(W))).\end{aligned}$$

In particular, when $\widehat{w}^{att(g,t)}(X)' = \mathbf{1}'\widehat{\Omega}^{*}_{gt}(X)^{-1} / \mathbf{1}'\widehat{\Omega}^{*}_{gt}(X)^{-1}\mathbf{1}$, *the estimators* $\widehat{ATT}_{stg}(g,t)$ *achieves the semiparametric efficiency bound as in Theorem 4. Consequently,* $\widehat{ES}(e)$ *is also asymptotically normal and semiparametrically efficient.*

Selecting the sieve index

- Use Information Criterion

$$\hat{K} = \arg\min_{K} 2\mathbb{E}_n \left[G_{g'} \left(\psi^K(X)' \hat{\beta}_K \right)^2 - 2G_g \left(\psi^K(X)' \hat{\beta}_K \right) \right] + \frac{C_n K}{n},$$

- AIC: $C_n = 2$; BIC: $C_n = \log(n)$

- Achieves consistent estimation (Chen and Liao, 2014)

Lemma (Consistency of the estimator selected by IC)

Under the regularity conditions specified in the paper, for any fixed K^*, we have $\hat{K} > K^*$ with probability approaching one.

Staggered without covariates

Special case: Efficient Staggered DiD without covariates

- When covariates are not important for identification, the results follow directly from the case with covariates, taking $X = 1$ a.s.
- Heuristically, the efficient staggered DiD estimand involves considering different DiD estimands that use different baseline periods and comparison groups, and aggregating these many DiDs using efficient weights.
- Of course, the notation gets simpler and some terms in the IF disappear.
- Plus, when it comes to estimation, everything becomes very easy as you just need to compute sample means and sample covariances.
- For completeness, let's get to some of the details.

Efficient Staggered DiD without covariates

- For $g' \in \mathcal{G}_{\text{trt}}$ and $1 \leq t'' \leq g' - 1$, let

$$\mathbb{IF}^{att(g,t)}_{g',t''} = \frac{G_g}{\pi_g}(Y_t - Y_1 - \mathbb{E}[Y_t - Y_1 | G = g]) - \left(\frac{G_\infty}{\pi_\infty}(Y_t - Y_{t''} - \mathbb{E}[Y_t - Y_{t''} | G = \infty]) + \frac{G_{g'}}{\pi_{g'}}(Y_{t''} - Y_1 - \mathbb{E}[Y_{t''} - Y_1 | G = g']) \right).$$

- Next, we collect all noncollinear $\mathbb{IF}^{att(g,t)}_{g',t''}$.
  - First, for $g = g'$, we collect all $\mathbb{IF}^{att(g,t)}_{g',t''}, 1 \leq t'' \leq g' - 1$, into the vector

$$\mathbb{IF}^{att(g,t)}_{g} = (\mathbb{IF}^{att(g,t)}_{g,1}, \mathbb{IF}^{att(g,t)}_{g,2}, \cdots, \mathbb{IF}^{att(g,t)}_{g,g-1})'.$$

  - Next, for each $g' \neq g$, we collect $\mathbb{IF}^{att(g,t)}_{g',t''}, 2 \leq t'' \leq g' - 1$, into a vector

$$\mathbb{IF}^{att(g,t)}_{g'} = (\mathbb{IF}^{att(g,t)}_{g',2}, \mathbb{IF}^{att(g,t)}_{g',3}, \cdots, \mathbb{IF}^{att(g,t)}_{g',g'-1})'.$$

Efficient Staggered DiD without covariates

- Let $\mathbb{IF}_{\text{stg}}^{att(g,t)}$ denote the vector stacking all these vectors together

$$\mathbb{IF}_{\text{stg}}^{att(g,t)} = (\mathbb{IF}_{g'}^{att(g,t),'}, g' \in \mathcal{G}_{\text{trt}})'. \tag{14}$$

- Let $\Omega_{gt,\text{UC}}$ be covariance matrix of $\mathbb{IF}_{\text{stg}}^{att(g,t)}$ with (j,k)-th element given by

$$\begin{aligned}
&\frac{1}{\pi_g}\text{Cov}(Y_t - Y_1, Y_t - Y_1 | G = g) + \frac{1}{\pi_\infty}\text{Cov}(Y_t - Y_{t_j''}, Y_t - Y_{t_k''} | G = \infty) \\
&-\frac{\mathbf{1}\{g = g_j'\}}{\pi_g}\text{Cov}(Y_t - Y_1, Y_{t_j''} - Y_1 | G = g) - \frac{\mathbf{1}\{g = g_k'\}}{\pi_g}\text{Cov}(Y_t - Y_1, Y_{t_k''} - Y_1 | G = g) \\
&+\frac{\mathbf{1}\{g_j = g_k'\}}{\pi_{g_j'}}\text{Cov}(Y_{t_j''} - Y_1, Y_{t_k''} - Y_1 | G = g_j'),
\end{aligned} \tag{15}$$

where (g_s', t_s'') is the value that g' and t'' takes in the s-th entry of $\mathbb{IF}_{\text{stg}}^{att(g,t)}$

Efficient Staggered DiD without covariates

- We can form the following simple but efficient $ATT(g,t)$ estimand,

$$ATT(g,t) = \frac{\mathbf{1}'\Omega_{gt,\text{UC}}^{-1}}{\mathbf{1}'\Omega_{gt,\text{UC}}^{-1}\mathbf{1}}\theta_{\text{stg,UC}}^{\text{att(g,t)}}, \tag{16}$$

where $\theta_{\text{stg,UC}}^{\text{att(g,t)}}$ is the column vector

$$\theta_{\text{stg,UC}}^{\text{att(g,t)}} = (\theta_{g',\text{UC}}^{\text{att(g,t)}}, g' \in \mathcal{G}_{\text{trt}})',$$

such that, for $g' = g$,

$$\theta_{g',\text{UC}}^{\text{att(g,t)}} = (\theta_{g,1,\text{UC}}^{\text{att(g,t)}}, \ldots, \theta_{g,g-1,\text{UC}}^{\text{att(g,t)}})',$$

and, for $g' \neq g$,

$$\theta_{g',\text{UC}}^{\text{att(g,t)}} = (\theta_{g',2,\text{UC}}^{\text{att(g,t)}}, \ldots, \theta_{g',g'-1,\text{UC}}^{\text{att(g,t)}})',$$

with

$$\theta_{g',t'',\text{UC}}^{\text{att(g,t)}} = \mathbb{E}[Y_t - Y_1 | G = g] - \left(\mathbb{E}[Y_t - Y_{t''} | G = \infty] + \mathbb{E}[Y_{t''} - Y_1 | G = g']\right).$$

Semiparametric efficiency result

Corollary (Efficient unconditional DiD with staggered adoption)

Let Assumptions M and PT-All hold unconditionally. Then, plug-in estimator based on $ATT(g,t)$ *as defined in* (16) *is semiparametrically efficient for the* $ATT(g,t)$*, with the efficiency bound being* $(\mathbf{1}'\Omega_{gt,UC}^{-1}\mathbf{1})^{-1}$.

Simulation Details

Potential outcomes in the staggered treatment setup

- $Y_{i,t}(g)$: Potential outcome (Returns on Asset) of firm i in year t if it was first-treated in period g:

$$\begin{aligned} Y_{i,t}(\infty) &= \alpha_i + \eta_t + \varepsilon_{i,t}, \\ Y_{i,t}(5) &= Y_{i,t}(\infty) + (0.5 \times \sigma_{\text{ROA}})\, 1\{t \geq 5\}\, (t-4) \\ Y_{i,t}(8) &= Y_{i,t}(\infty) + (0.3 \times \sigma_{\text{ROA}})\, 1\{t \geq 8\}\, (t-7) \\ Y_{i,t}(11) &= Y_{i,t}(\infty) + (0.1 \times \sigma_{\text{ROA}})\, 1\{t \geq 11\}\, (t-10) \end{aligned}$$

- $\varepsilon_{i,t} = \rho \varepsilon_{i,t-1} + u_{i,t}$: Error terms follow an $AR(1)$ process with common ρ and $u_{i,t} \sim iid$.
- Treatment Groups: G_i is randomly allocated to one of the three possible treatment dates, $\{5, 8, 11\}$, with equal probability.
- Use Compustat data from 1980 to 2015 from 11,978 firms to estimate α_i, η_t, $u_{i,t}$ and σ_{ROA}.

Go back

Assessing the plausibility of PT assumptions

Incremental Hausman test

- The over-identifying restrictions in Assumption PT-All can be tested using Hausman-type tests.
- Let $\widetilde{ES} = (\widetilde{ES}(e), e \in \mathcal{E})$ denote the set of ES estimators constructed based on the just-identifying assumption (post-treatment parallel trends), i.e., using $t' = g - 1$ as baseline and the never-treated units as comparison group.
- We can construct the Hausman-type test statistic as

$$\widehat{H} = n \left(\widehat{ES} - \widetilde{ES}\right)' \left(\widehat{\text{aCov}}(\widetilde{ES} - \widehat{ES})\right)^{-1} \left(\widehat{ES} - \widetilde{ES}\right),$$

  where $\widehat{\text{aCov}}$ denotes the corresponding asymptotic covariance estimator for the difference of ES.
- The test rejects the parallel trends assumption for all periods if $\hat{H}$ exceeds the corresponding critical value of a $\chi^2(|\mathcal{E}|)$ distribution.

Incremental Hausman test

Theorem (Hausman Test)

Suppose that the estimator $\widehat{ES}$ is constructed under the conditions of Theorem 6, and $\widetilde{ES}$ also satisfies the corresponding conditions. Assume that the estimated covariance matrix $\widehat{\text{aCov}}(\widetilde{ES} - \widehat{ES})$ is consistent. Then the test statistic $\hat{H}$ converges in distribution to a $\chi^2(|\mathcal{E}|)$ distribution, where $|\mathcal{E}|$ denotes the number of elements in $\mathcal{E}$. Also, this Hausman test has nontrivial power against all local alternatives.

- We are also working on visualization tools to assess the magnitude of the PT violations.

Holm-Bonferroni Method for Selecting Moment Restrictions

- Assumptions PT-All and PT-Post define the largest and smallest sets of moment restrictions.
- Let $\mathcal{M}$ denote the set of conditional moment restrictions specified by PT-Post.
- For each $g' > t''$, let $\mathcal{M}_{g',t''}$ extend $\mathcal{M}$ by adding one more moment.
- Perform a Hausman-type test for each $\mathcal{M}_{g',t''}$, compute p-values $\mathrm{p}_{g',t''}$, and order them.
- Apply the Holm-Bonferroni procedure:
  - For each ℓ, compare $\mathrm{p}_{(\ell)}$ with $\frac{\alpha}{L+1-\ell}$.
  - Reject the corresponding conditional moment if $\mathrm{p}_{(\ell)} < \frac{\alpha}{L+1-\ell}$.
- This is more powerful than the simple Bonferroni correction.

References

Ai, Chunrong and Xiaohong Chen, "Efficient estimation of models with conditional moment restrictions containin unknown functions," *Econometrica*, 2003, *71* (6), 1795–1843.

_ and _, "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 2012, *170* (2), 442–457.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference-in-Differences," *American Economic Review*, 2021, *111* (12), 4088–4118.

Armstrong, Timothy B., Patrick Kline, and Liyang Sun, "Adapting to Misspecification," *Econometrica*, 2025, *93* (6), 1981–2005.

Athey, Susan and Guido Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

Baker, Andrew, David F. Larcker, and Charles C. Y. Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social Media and Mental Health," *American Economic Review*, 2022, *112* (11), 3660–93.

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

__ **and Tong Li**, "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 2019, *10* (4), 1579–1618.

Chen, Xiaohong, "Large Sample Sieve Estimation of Semi-Nonparametric Models," in James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Amsterdam: Elsevier, 2007, chapter 76, pp. 5549–5632.

__ **and Andres Santos**, "Overidentification in Regular Models," *Econometrica*, 2018, *86* (5).

__ **and Zhipeng Liao**, "Sieve inference on irregular parameters," *Journal of Econometrics*, 2014, *182* (1), 70–86.

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

__ **and** __ , "Difference-in-Differences Estimators of Intertemporal Treatment Effects," *The Review of Economics and Statistics*, 2024, *Forthcoming*, 1–45.

_ , Xavier D'Haultfoeuille, Félix Pasquier, Doulo Sow, and Gonzalo Vazquez-Bare, "Difference-in-Differences for Continuous Treatments and Instruments with Stayers," *arXiv:2201.06898v4*, 2024.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo, "The economic consequences of hospital admissions," *American Economic Review*, 2018, *108* (2), 308–352.

Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M. Taylor, "A Local Projections Approach to Difference-in-Differences," Working Paper 31184, National Bureau of Economic Research April 2023.

Gardner, John, "Two-stage differences in differences," *Working Paper*, 2021.

Ghanem, Dalia, Pedro H. C. Sant'Anna, and Kaspar Wüthrich, "Selection and parallel trends," *arXiv:2203.09001*, 2023.

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

Imbens, Guido, Nathan Kallus, and Xiaojie Mao, "Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models," *arXiv:2108.03849*, 2021.

Imbens, Guido W. and Davide Viviano, "Identification and Inference for Synthetic Controls with Confounding," *arXiv:2312.00955*, 2023.

Lee, Soo Jeong and Jeffrey M. Wooldridge, "A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data," *Working Paper*, 2023. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4516518.

Marx, Philip, Elie Tamer, and Xun Tang, "Parallel Trends and Dynamic Choices," *Journal of Political Economy: Microeconomics*, 2024, *2* (1), 129––171.

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

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *American Economic Review: Insights*, 2022, *4* (3), 305–322.

_ and Pedro H. C. Sant'Anna, "Efficient Estimation for Staggered Rollout Designs," *Journal of Political Economy: Microeconomics*, 2023, *1* (4), 669–709.

_ and _, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2023, *91* (2), 737–747.

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

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Tchetgen, Eric J. Tchetgen, Chana Park, and David B. Richardson, "Universal Difference-in-Differences for Causal Inference in Epidemiology," *Epidemiology*, 2024, *35* (1), 16–22.

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

Wooldridge, Jeffrey M., "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data," *Econometrics Journal*, 2023, *Forthcoming*.