Difference-in-Differences with Multiple Time Periods, Journal of Econometrics, 2021

Wenzhi Wang [*]

August 20, 2024

Callaway and Sant'Anna (2021)

1. Introduction

In this article, we provide a unified framework for average treatment effects in DiD setups with multiple time periods, variation in treatment timing, and when the parallel trends assumption holds potentially only after conditioning on observed covariates. We concentrate our attention on DiD with *staggered adoption*, i.e., to DiD setups such that once units are treated, they remain treated in the following periods.

The core of our proposal relies on separating the DiD analysis into three separate steps: (i) identification of policy-relevant disagreegated causal parameters; (ii) aggregation of these parameters to form summary measures of the causal effects; and (iii) estimation and inference about these different target parameters. Our approach allows for estimation and inference on interpretable causal parameters allowing for arbitrary treatment effect heterogeneity and dynamic effects, thereby completely avoiding the issues of interpreting results of standard two-way fixed effects (TWFE) regressions as causal effects in DiD setups.

The identification step of the analysis provides a blueprint for the other steps. In this paper, we pay particular attention to the disaggregated causal parameter that we call the *group-time average treatment effect*, i.e., the average treatment effect for group g at time t, where a "group" is defined by the time period when units are first treated. In the canonical DiD setup with two periods and tow groups, these parameters reduce to the ATT which is typically the parameter of interest in that setup. *An attractive feature of the group-time average treatment effect parameters is that they do not directly restrict heterogeneity with respect to observed covariates, the period in which units are first treated, or the evolution of treatment effect heterogeneity, and/or to construct many other more aggregated causal parameters.*

[*] This note is written in my pre-doc period at the University of Chicago Booth School of Business.

We provide sufficient conditions related to treatment anticipation behavior and conditional parallel trends under which these group-time average treatment effects are nonparametrically point-identified. A unique feature of our framework is that it shows how researchers can flexibly incorporate covariates into the staggered DiD setup with multiple groups and multiple periods. This is particularly important in applications in which differences in observed characteristics create non-parallel outcome dynamics between different groups - in this case, unconditional DiD strategies are generally not appropriate to recover sensible causal parameters of interest.

We propose three different types of DiD estimands in staggered treatment adoption setups: one based on *outcome regressions*, one based on *inverse probability weighting*, and one based on *doubly-robust methods*. We provide versions of these estimands both for the case with panel data and for the case with repeated cross sections data. To the best of our knowledge, this paper is the first to show how one can allow for covariate-specific trends across groups in DiD setups with variation in treatment timing. Our results also highlight that, in practice, one can rely on different types of parallel trends assumptions and allow some types of treatment anticipation behavior; our proposed estimands explicitly reflect these assumptions.

Our framework acknowledges that in some applications there may be many group-time average treatment effects and researchers may want to aggregate them into different summary causal effect measures. This characterizes the aggregation step of the analysis. We provide ways to aggregate the potentially large number of group-time average treatment effects into a variety of intuitive summary parameters and discuss specific aggregation schemes that can be used to highlight different sources of treatment effect heterogeneity across groups and time periods.

In particular, we consider aggregation schemes that deliver a single overall treatment effect parameter with similarities to the ATT in the two period and two group case as well as partial aggregations that highlight heterogeneity along certain dimensions such as (a) how average treatment effects vary with length of exposure to the treatment (event-study-type estimands); (b) how average treatment effects vary across treatment groups; and (c) how cumulative average treatment effects evolve over calendar time. We also provide a formal discussion of the costs and benefits of balancing the sample in "event time" when analyzing dynamic treatment effects. Overall, our setup makes it clear that, in general, the "best" aggregation scheme is application-specific, as it depends on the type of question one wants to answer.

Given that our identification results are constructive, we propose easy-to-use plug-in type (parametric) estimators for the causal parameters of interest. Although the outcome regression, inverse probability weighting and doubly-robust estimands are equivalent from the identification point of view, they suggest different types of DiD estimators one can use in practice. Here, we note that using doubly-robust estimators can be particularly attractive as they rely on less stringent modeling conditions than the outcome regression and the inverse probability weighting procedures.

In order to conduct asymptotically valid inference, we justify the use of a computationally convenient multiplier-type bootstrap procedure. This approach can be used to obtain simultaneous confidence bands for the group-time average treatment effects. Unlike commonly used pointwise confidence bands, our simultaneous confidence bands asymptotically cover the entire path of the group-

time average treatment effects with fixed probability and take into account the dependency across different group-time average treatment effect estimators. Thus, our proposed confidence bands are arguably more suitable for visualizing the overall estimation uncertainty than more traditional pointwise confidence intervals.

2. Identification

2.1. Setup

We consider the case with $\mathcal{T}$ periods and denote a particular time period by t where $t = 1, \ldots, \mathcal{T}$. In a canonical DiD setup, $\mathcal{T} = 2$ and no one is treated in period $t = 1$. Let $D_{i,t}$ be a binary variable equal to one if unit i is treated in period t and equal to zero otherwise. We make the following assumption about the treatment process:

Assumption 1 (Irreversibility of Treatment). $D_{i,1} = 0$ almost surely (a.s.). For $t = 2, \ldots, \mathcal{T}$, $D_{i,t-1} = 1$ implies that $D_{i,t} = 1$ a.s.

Assumption 1 states that no one is treated at time $t = 1$, and that once a unit becomes treated, that unit will remain treated in the next period. This assumption is also called staggered treatment adoption in the literature. We interpret this assumption as if units do not "forget" about the treatment experience.

Define G_i as the time period when a unit first becomes treated. Under Assumption 1, for all units that eventually participate in the treatment, G_i defines which "group" they belong to. If a unit does not participate in any time period, we set $G = \infty$. We define G_g to be a binary variable that is equal to one if a unit is first treated in period g, i.e., $G_{i,g} = \mathbb{I}(G_i = g)$. Define C to be a binary variable that is equal to one for units that do not participate in the treatment in any time period, i.e., $C_i = \mathbb{I}(G_i = \infty) = 1 - D_{i,\mathcal{T}}$. Let $\overline{g} = \max_{i=1,\ldots,n} G_i$ be the maximum G in the dataset.

Next, denote the generalized propensity score as

$$p_{g,s}(X) = \mathbb{P}\left(G_g = 1 \mid X, G_g + (1 - D_s)(1 - G_s) = 1\right).$$

Note that $p_{g,s}(X)$ indicates the probability of being first treated at time g, conditional on pre-treatment covariates X and on either being a member of group g (in this case, $G_g = 1$) or a member of the "*not-yet-treated*" group by time s (in this case, being a member of group $(1 - D_s)(1 - G_s) = 1$). Many of our results use a specialized version of this generalized propensity score, and, henceforth, we define

$$p_g(X) = p_{g,\mathcal{T}}(X) = \mathbb{P}\left(G_g = 1 \mid X, G_g + C = 1\right)$$

which is the probability of being first treated in period g conditional on covariates and either being a member of group g or not participating in the treatment in any time period. Let $\mathcal{G} = \text{supp}\,(G) \setminus \{\overline{g}\} \subset$

$\{2, 3, \ldots, \mathcal{T}\}$ dentoe the support of G excluding $\overline{g}$[1]. Likewise, let $\mathcal{X} = \text{supp}\,(X) \subset \mathbb{R}^k$ denote the support of the pre-treatment covariates. Finally, for a generic $\delta \geq 0$, let $\mathcal{G}_\delta = \mathcal{G} \bigcap \{2 + \delta, 3 + \delta, \ldots, \mathcal{T}\}$.

Next, we set up the potential outcomes framework. Here, we combine the dynamic potential outcomes framework with multi-stage treatment adoption setup. Let $Y_{it}(0)$ denote unit i's untreated potential outcome at time t if they remain untreated through time period $\mathcal{T}$; i.e., if they were not to participate in the treatment across all available time periods. For $g = 2, \ldots, \mathcal{T}$, let $Y_{i,t}(g)$ denote the potential outcome that unit i would experience at time t if they were to first become treated in time period g. Note that our potential outcomes notation accounts for potential dynamic treatment selection, though it also accommodates (pre-specified) treatment regimes. The observed and potential outcomes for each unit i are related through

$$Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{\mathcal{T}} \left(Y_{i,t}(g) - Y_{i,t}(0)\right) \cdot G_{i,g}. \tag{1}$$

We also impose the following random sampling assumption.

Assumption 2 (Random Sampling)**.** $\{Y_{i,1}, \ldots, Y_{i,\mathcal{T}}, X_i, D_{i,1}, \ldots, D_{i,\mathcal{T}}\}_{i=1}^n$ is independent and identically distributed (iid).

Assumption 2 implies that we have access to panel data; our results extend essentially immediately to the case with repeated cross sections data. Here, we note that Assumption 2 allows us to view all potential outcomes as random. Furthermore, it does not impose restrictions between potential outcomes and treatment allocation, nor does it restrict the time series dependence of the observed random variables. On the other hand, Assumption 2 imposes that each unit i is randomly drawn from a large population of interest.

2.2. The Group-Time Average Treatment Effect Parameter

Given that different potential outcomes cannot be observed for the same unit at the same time, researchers often focus on identifying and estimating some average causal effects. For instance, in the canonical DiD setup with two time periods, the most popular treatment effect parameter of interest is the average treatment effect on the treated, denoted by

$$ATT = \mathbb{E}\left[Y_{i,2}(2) - Y_{i,2}(0) \mid G_{i,2} = 1\right]$$

In this paper, we consider a natural generalization of the ATT that is suitable to setups with multiple treatment groups and multiple time periods. More precisely, we use the average treatment effect for units who are members of a particular group g at a particular time period t, denoted by

$$ATT(g, t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(0) \mid G_{i,g} = 1\right]$$

[1]When there is a "never treated" set of units with $G = \infty$, $\mathcal{G}$ only excludes this group. When such "never-treated" group is not available, we exclude the latest-treated group as there will be no valid untreated comparison group for them.

as the main building block of our framework. We call this causal parameter the *group-time average treatment effect*.

Note that $ATT(g,t)$ does not impose any restriction on treatment effect heterogeneity across groups or across time. Thus, focusing on the family of $ATT(g,t)$'s allow us to analyze how average treatment effects vary across different dimensions in a unified manner. For instance, by fixing a group g and varying time t, one is able to highlight how average treatment effects evolve over time for that specific group. By doing this for different groups, we can have a better understanding about how treatment effect dynamics vary across groups. In addition, one can build on the $ATT(g,t)$'s to form more aggregated causal parameters that are constructed to answer specific questions like (a) What was the average effect of participating in the treatment across all groups that participated in the treatment by time period $\mathcal{T}$? (b) Are average treatment effects heterogeneous across groups? (c) How do average treatment effects vary by length of exposure to the treatment? (d) How do cumulative average treatment effects evolve over calendar time?

2.3. Identifying Assumptions

In order the identify the $ATT(g,t)$ and their functionals, we impose the following assumptions.

Assumption 3 (Limited Treatment Anticipation)**.** There is known $\delta \geq 0$ such that

$$\mathbb{E}\left[Y_t(g) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) \mid X, G_g = 1\right]$$

a.s. for all $g \in \mathcal{G}$, $t \in \{1, \ldots, \mathcal{T}\}$ such that $t < g - \delta$.

Assumption 2 restricts anticipation of the treatment for all "eventually treated" groups. When $\delta = 0$, it imposes a "no-anticipation." This is likely to be the case when the treatment path is not a priori known and/or units ae not the ones who "choose" treatment status. However, Assumption 3 also allows for anticipation behavior, as long as we have a good understanding about the anticipation horizon δ. For instance, if units anticipate treatment by one period, Assumption 3 would hold with $\delta = 1$. Note that, under Assumption 3, $ATT(g,t) = 0$ for all pre-treatment periods such that $t < g - \delta$.

Next, we consider two alternative assumptions that impose restrictions on the evolution of untreated potential outcomes.

Assumption 4 (Conditional Parallel Trends Based on a "Never-Treated" Group)**.** Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and $t \in \{2, \ldots, \mathcal{T}\}$ such that $t \geq g - \delta$,

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

Assumption 5 (Conditional Parallel Trends Based on a "Not-Yet-Treated" Group)**.** Let δ be as defined in Assumption 3. For each $g \in \mathcal{G}$ and $t \in \{2, \ldots, \mathcal{T}\}$ such that $t \geq g - \delta$ and $t + \delta \leq s < \overline{g}$,

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

Assumptions 4 and 5 are two different conditional parallel trends assumptions that generalize the two-period parallel trends assumption to the case where there are multiple periods and multiple treatment groups. Both assumptions hold after conditioning on covariates X. This can be important in many applications in economics particularly in cases where there are covariate specific trends in outcomes over time and when the distribution of covariates is different across groups.

Assumptions 4 and 5 differ from each other depending on the comparison group one is willing to use in a given application. More specifically, Assumption 4 states that, conditional on covariates, the average outcomes for the group first treated in period g and for the "never-treated" group would have followed parallel paths in the absence of treatment. Assumption **??** imposes conditional parallel trends between group g and groups that are "not-yet-treated" by time $t + \delta$.

Importantly, both of these assumptions allow for covariate-specific trends and do not restrict the relationship between treatment timing and the potential outcomes, $Y_g(g)$'s. Thus, they are weaker than the randomization-based assumption made by Athey and Imbens (2022).

In our view, *practitioners may favor Assumption 4 with respect to Assumption 5 when there is a sizeable group of units that do not participate in the treatment in any period, and, at the same time, these units are similar enough to the "eventually treated" units*. When a "never-treated" group of units is not available or "too small", researchers may favor Assumption 5 as it allows one to use more groups as valid comparison units, which potentially leads to more informative inference procedures.

However, it is important to stress that favoring Assumption 5 with respect to Assumption 4 also involves potential drawbacks. For instance, in the absence of treatment anticipation ($\delta = 0$), Assumption 4 does not restrict observed pre-treatment trends across groups, whereas Assumption 5 does. Not restricting pre-treatment trends may be particularly meaningful in applications where the economic environment during the "early-periods" was potentially different from the "later-periods". In these cases, the outcomes of different groups may evolve in a non-parallel manner during "early-periods", perhaps because the groups were exposed to different shocks, while trends become parallel in the "later-periods".

Assumption 6 (Overlap)**.** For each $t \in \{2, \ldots, \mathcal{T}\}$, $g \in \mathcal{G}$, there exists some $\varepsilon > 0$ such that $\mathbb{P}(G_g = 1) > \varepsilon$ and $p_{g,t}(X) < 1 - \varepsilon$ a.s.

It states that a positive fraction of the population starts treatment in period g, and that, for all g and t, the generalized propensity score is uniformly bounded away from one.

Remark 1. Note that Assumptions 3 and 4 (Assumption 5) are intrinsically connected. For instance, when one imposes the "no-anticipation" condition (so that $\delta = 0$), Assumption 4 would then impose conditional parallel trends only for post-treatment periods $t \geq g$. If one allows for anticipation behavior (so that $\delta > 0$), Assumption 4 would then impose conditional parallel trends in some pre-treatment periods, too. In fact, the parallel trends assumptions become stronger as one increases δ.

Remark 2. In some applications, practitioners may not be comfortable with using "never-treated" units as part of the comparison group because they behave very differently from the other "eventually treated" units. In these cases, practitioners could drop all "never-treated" units from the analysis and proceed with Assumption 5.

2.4. Nonparametric Identification of the Group-Time Average Treatment Effects

In this section, we show that the family of group-time average effects are nonparametrically point-identified under the aforementioned assumptions. Furthermore, we show that one can use outcome regression (OR), inverse probability weighting (IPW), or doubly robust (DR) estimands to recover the $ATT(g,t)$'s.

References

Athey, Susan and Guido W. Imbens (2022) "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 226 (1), 62–79, 10.1016/j.jeconom.2020.10.012.

Callaway, Brantly and Pedro H.C. Sant'Anna (2021) "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225 (2), 200–230, 10.1016/j.jeconom.2020.12.001.