ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Design-based analysis in Difference-In-Differences settings with staggered adoption☆

Susan Athey [a,b], Guido W. Imbens [a,b,c,d,*]

[a] Graduate School of Business, Stanford University, United States of America
[b] NBER, United States of America
[c] Department of Economics, Stanford University, United States of America
[d] SIEPR, United States of America

ARTICLE INFO

Article history:
Received 22 February 2019
Received in revised form 22 February 2019
Accepted 19 October 2020
Available online xxxx

Keywords:
Staggered adoption design
Difference-In-Differences
Fixed effects
Randomization distribution

ABSTRACT

In this paper we study estimation of and inference for average treatment effects in a setting with panel data. We focus on the staggered adoption setting where units, *e.g.* individuals, firms, or states, adopt the policy or treatment of interest at a particular point in time, and then remain exposed to this treatment at all times afterwards. We take a design perspective where we investigate the properties of estimators and procedures given assumptions on the assignment process. We show that under random assignment of the adoption date the standard Difference-In-Differences (DID) estimator is an unbiased estimator of a particular weighted average causal effect. We characterize the exact finite sample properties of this estimand, and show that the standard variance estimator is conservative.

1. Introduction

In this paper we study estimation of and inference for average treatment effects in a setting with panel data. We focus on the setting where units, *e.g.*, individuals, firms, or states, adopt the policy or treatment of interest at a particular point in time, and then remain exposed to this treatment at all times afterwards. The adoption date at which units are first exposed to the policy may, but need not, vary by unit. We refer to this as a *staggered adoption design* (SAD), such designs are sometimes also referred to as event study designs. An early example is Athey and Stern (1998) where adoption of an enhanced 911 technology by counties occurs over time, with the adoption date varying by county. This setting is a special case of the general Difference-In-Differences (DID) set up (*e.g.*, Card, 1990; Meyer et al., 1995; Angrist and Pischke, 2008; Angrist and Krueger, 2000; Abadie et al., 2010; Borusyak and Jaravel, 2016; Athey and Imbens, 2006; Card and Krueger, 1994; Freyaldenhoven et al., 2019; de Chaisemartin and d'Haultfœuille, 2020; Abadie, 2005) where units can switch back and forth between being exposed or not to the treatment. In this SAD setting we are concerned with identification issues as well as estimation and inference. In contrast to most of the theoretical DID literature, *e.g.*, Bertrand et al. (2004), Shah et al. (1977), Conley and Taber (2011), Donald and Lang (2007), Stock and Watson (2008), Arellano (1987b, 2003), Sun

☆ We are grateful for comments by participations in the conference in honour of Gary Chamberlain at Harvard in May 2018, and in particular by Gary Chamberlain. Gary's insights over the years have greatly affected our thinking on these problems. We are also grateful for comments by Kosuke Imai. We also wish to thank Sylvia Klosin and Michael Pollmann for superb research assistance. This research was generously supported by ONR, United States grant N00014-17-1-2131.

* Corresponding author at: Graduate School of Business, Stanford University, United States of America.
E-mail addresses: athey@stanford.edu (S. Athey), imbens@stanford.edu (G.W. Imbens).

https://doi.org/10.1016/j.jeconom.2020.10.012
0304-4076/

and Abraham (2020), Wooldridge (2010), de Chaisemartin and d'Haultfœuille (2017), Callaway and Sant'Anna (2020), Goodman-Bacon (2018) and de Chaisemartin and d'Haultfœuille (2020), we take a *design-based* perspective where the properties of the estimators arises from the stochastic nature of the treatment assignment, rather than a *sampling-based* or *model-based* perspective where these properties arise from the random sampling of units from a large population in combination with assumptions on this population distribution. Such a design perspective, motivating randomization or permutation based inference, has been common for many years in the analysis of randomized experiments, e.g., Neyman (1923/1990) and Rosenbaum (2002), and has more recently received attention in observational study settings (Rosenbaum, 2017; Aronow and Samii, 2016; Abadie et al., 2016, 2020). This perspective is particularly attractive in the settings when the sample comprises the entire population, *e.g.*, all states of the US, or all countries of the world so that non-degenerate sampling properties would require postulating an imaginary super-population. In this design setting our critical assumptions involve restrictions on the assignment process as well as exclusion restrictions, but in contrast to other work in this area they do not include functional form assumptions. Commonly made common trend assumptions (de Chaisemartin and d'Haultfœuille, 2020; Sun and Abraham, 2020; Hull, 2018) follow from some of our assumptions, but are not the starting point.

As in Sun and Abraham (2020) we set up the problem with the adoption date, rather than the actual exposure to the intervention, as the basic treatment indexing the potential outcomes. We consider assumptions under which this discrete multi-valued treatment (the adoption date) can be reduced to a binary one, defined as the indicator whether or not the treatment has already been adopted. We then investigate the properties of the standard DID estimator under assumptions about the assignment of the adoption date and under various exclusion restrictions. We show that under a random adoption date assumption, the standard DID estimator can be interpreted as the weighted average of average causal effects of changes in the adoption date. We also consider design-based inference for this estimand. We derive the exact variance of the DID estimator in this setting. We show that under a random adoption date assumption the standard Liang–Zeger (LZ) variance estimator (Liang and Zeger, 1986; Chamberlain, 1984; Arellano, 1987a; Bertrand et al., 2004), or the clustered bootstrap, are conservative. For this case we propose an improved (but still conservative) variance estimator.

Our paper is most closely related to a set of recent papers on DID methods that explicitly focus on issues with heterogenous treatment effects (Borusyak and Jaravel, 2016; Goodman-Bacon, 2018; Sun and Abraham, 2020; de Chaisemartin and d'Haultfœuille, 2020; Han, 2020; Callaway and Sant'Anna, 2020; Hull, 2018; Strezhnev, 2018; Imai and Kim, 2019; Hazlett and Xu, 2018; Ben-Michael et al., 2018; Arkhangelsky and Imbens, 2018, and Arkhangelsky et al., 2019). Among other things these papers derive interpretations of the DID estimator as weighted averages of causal effects and bias terms under various assumptions. In many cases they find that these interpretations involve weighted averages of basic average causal effects with potentially negative weights and derive conditions, or propose alternative estimators that ensure the weights are non-negative. Optimal design issues in the staggered adoption case have been considered in Xiong et al. (2019) and Doudchenko et al. (2019).

2. Set up

Using the potential outcome framework for causal inference, we consider a setting with a finite population of N units. Each of these N units are characterized by a set of potential outcomes (*e.g.,* Rubin, 1974; Imbens and Rubin, 2015) in T periods for $T+1$ treatment levels, $Y_{it}(a)$. Here $i \in \{1, \ldots, N\}$ indexes the units, $t \in \mathbb{T} = \{1, \ldots, T\}$ indexes the time periods, and the argument of the potential outcome function $Y_{it}(\cdot)$ is the adoption date $a \in \mathbb{A}$, where $\mathbb{A} = \mathbb{T} \cup \{\infty\} = \{1, \ldots, T, \infty\}$. This argument, a, which indexes the discrete treatment, is the date that the binary policy was first adopted by a unit. Units can adopt the policy at any of the time periods $1, \ldots, T$, or not adopt the policy at all during the period of observation, in which case we code the adoption date as ∞. Once a unit adopts the treatment, it remains exposed to the treatment for all periods afterwards. This set up matches that in Sun and Abraham (2020) and Hazlett and Xu (2018), and differs from that in most of the DID literature where the binary indicator whether a unit is exposed to the treatment in the current period indexes the potential outcomes. The notion of focusing on a full treatment path rather than a binary treatment is also related to the dynamic treatment effect literature (*e.g.,* Hernan and Robins, 2020; Han, 2020. We observe for each unit in the population the adoption date $A_i \in \mathbb{A}$ and the sequence of T realized outcomes, Y_{it}, for $t \in \mathbb{T}$, where the realized outcome for unit i in period t equals

$$Y_{it} \equiv Y_{it}(A_i). \tag{2.1}$$

We may also observe pre-treatment characteristics, denoted by the K-component vector X_i, although for most of the discussion we abstract from their presence. Let $\boldsymbol{Y}$, $\boldsymbol{A}$, and $\boldsymbol{X}$ denote the $N \times T$, $N \times 1$, and $N \times K$ matrices with typical elements Y_{it}, A_i, and X_{ik} respectively. Implicitly we have already made a sutva-type assumption (Rubin, 1978; Imbens and Rubin, 2015) that units are not affected by the treatments (adoption dates) for other units. Our design-based analysis views the potential outcomes $Y_{it}(a)$ as deterministic, and only the adoption dates A_i, as well as functions thereof such as the realized outcomes, as stochastic. Distributions of estimators will be fully determined by the adoption date distribution, with the number of units N and the number of time periods T fixed, unless explicitly stated otherwise. Following the literature we refer to this as a randomization, or designed-based, distribution (Rosenbaum, 2017; Imbens and Rubin, 2015; Abadie et al., 2020), as opposed to a sampling or model-based distribution.

In many cases the units themselves are clusters of units of a lower level of aggregation. For example, the units may be states, and the outcomes could be averages of outcomes for individuals in that state, possibly of samples drawn from subpopulations from these states. In such cases N and T may be as small as 2, although in many of the cases we consider N will be at least moderately large. This distinction between cases where Y_{it} is itself an average over basic units or not, affects some, but not all, of the formal statistical analyses. It may make some of the assumptions more plausible, and it may affect the inference, especially if individual level outcomes and covariates are available.

Define $W : \mathbb{A} \times \mathbb{T} \mapsto \{0, 1\}$, with $W(a, t) = \mathbf{1}_{a \leq t}$ to be the binary indicator for the adoption date a preceding t, and define W_{it} to be the indicator for the policy having been adopted by unit i prior to, or at, time t, so that:

$$W_{it} \equiv W(A_i, t) = \mathbf{1}_{A_i \leq t}.$$

The $N \times T$ matrix $\boldsymbol{W}$ with typical element W_{it} has the form:

$$\boldsymbol{W}_{N\times T} = \begin{pmatrix} 1 & 2 & 3 & 4 & \ldots & T & \text{(time period)} \\ & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & \text{(never adopter)} \\ 0 & 0 & 0 & 0 & \ldots & 1 & \text{(late adopter)} \\ 0 & 0 & 0 & 0 & \ldots & 1 & \\ 0 & 0 & 1 & 1 & \ldots & 1 & \\ 0 & 0 & 1 & 1 & \ldots & 1 & \text{(medium adopter)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 1 & 1 & 1 & \ldots & 1 & \text{(early adopter)} \end{pmatrix}.$$

Let $N_a \equiv \sum_{i=1}^N \mathbf{1}_{A_i=a}$ be the number of units in the sample with adoption date a, define $\pi_a \equiv N_a/N$, for $a \in \mathbb{A}$, as the fraction of units with adoption date equal to a, and define $\Pi_t \equiv \sum_{s=1}^t \pi_s$, for $t \in \mathbb{T}$, as the fraction of units with an adoption date on or prior to t.

Also define $\overline{Y}_t(a)$ to be the population average of the potential outcome in period t for adoption date a:

$$\overline{Y}_t(a) \equiv \frac{1}{N}\sum_{i=1}^N Y_{it}(a), \qquad \text{for } t \in \mathbb{T}, a \in \mathbb{A}.$$

Define the unit and average causal effects of adoption date a' relative to a, on the outcome in period t, as

$$\tau_{it,aa'} \equiv Y_{it}(a') - Y_{it}(a), \qquad \tau_{t,aa'} \equiv \frac{1}{N}\sum_{i=1}^N \Big\{Y_{it}(a') - Y_{it}(a)\Big\} = \overline{Y}_t(a') - \overline{Y}_t(a).$$

Sun and Abraham (2020) focus on slightly different building blocks. In a super-population perspective they focus on the population average effect of adopting in period a relative to never adopting, on the outcome in period $t + a$, for the subpopulation of units who adopt the treatment in period a, $CATT_{a,t} = \mathbb{E}[\tau_{it+a,\infty a}|A_i = a]$. Callaway and Sant'Anna (2020) and Goodman-Bacon (2018), also in a super-population setting, focus on the average effect of adopting in period a, relative to never adopting, on the outcome in period t, again for the subpopulation who adopts the treatment in period a, ATT(a, t), equal to $\mathbb{E}[\tau_{it,\infty a}|A_i = a]$. This implies that $ATT(t, a) = CATT_{a,t-a}$. Under random assignment of the adoption date, and with N infinite, it follows that $ATT(t, a) = \tau_{t,\infty a}$ and $CATT_{t,a} = \tau_{t+a,\infty a}$.

The average causal effects $\tau_{t,aa'}$ form the components of many of the estimands we consider later. A particularly interesting average effect is

$$\tau_{t,\infty 1} = \frac{1}{N}\sum_{i=1}^N \Big(Y_{it}(1) - Y_{it}(\infty)\Big),$$

the average effect in period t of switching the entire population from never adopting the policy ($a = \infty$), to adopting the policy in the first period ($a = 1$). Formally there is nothing special about the particular average effect $\tau_{t,\infty 1}$ relative to any other $\tau_{t,aa'}$, but $\tau_{t,\infty 1}$ will be useful as a benchmark. Part of the reason is that for all t and i the comparison $Y_{it}(1) - Y_{it}(\infty)$ is between potential outcomes for adoption prior to or at time t (namely adoption date $a = 1$) and potential outcomes for adoption later than t (namely, never adopting, $a = \infty$). In contrast, any other average effect $\tau_{t,aa'}$ will for some t involve comparing potential outcomes neither of which correspond to having adopted the treatment yet, or comparing potential outcomes both of which correspond to having adopted the treatment already. Therefore, $\tau_{t,\infty 1}$ reflects more directly on the effect of having adopted the policy than any other $\tau_{t,aa'}$.

3. Assumptions

We consider three sets of assumptions. The first set, containing only a single assumption, is about the *design*, that is, the assignment of the treatment, here the adoption date, conditional on the potential outcomes and possibly pretreatment

variables. We refer to this as a design assumption because it can be guaranteed by design of the study. The second set of assumptions is about the *potential outcomes*, and rules out the presence of certain treatment effects. These exclusion restrictions are substantive assumptions, and they cannot be guaranteed by design. The third set of assumptions consists of four *auxiliary assumptions*, two about homogeneity of certain causal effects, one about sampling from a large population, and one about an outcome model in a large population. The nature of these three sets of assumptions, and their plausibility, is very different, and it is in our view useful to carefully distinguish between them. The current literature often combines various parts of these assumptions with functional form assumptions, either implicitly in the notation used, or in assumptions about the statistical models for the realized outcomes.

3.1. The design assumption

The first assumption is about the assignment process for the adoption date A_i. Our starting point is to assume that the adoption date is completely random:

Assumption 1 (*Random Adoption Date*)**.** For some set of positive integers N_a, for $a \in \mathbb{A}$,

$$\text{pr}(\boldsymbol{A} = \boldsymbol{a}) = \left(\frac{N!}{\prod_{a\in\mathbb{A}} N_a!} \right)^{-1},$$

for all N-vectors $\boldsymbol{a}$ such that for all $a \in \mathbb{A}$, $\sum_{i=1}^{N} \mathbf{1}_{a_i=a} = N_a$.

This assumption is obviously very strong. However, without additional assumptions, this assumption has no testable implications in a setting with exchangeable units, as stated formally in the following Lemma.

Lemma 1 (*No Testable Restrictions*)**.** *Suppose all units are exchangeable. Then Assumption 1 has no testable implications for the joint distribution of* $(\boldsymbol{Y}, \boldsymbol{A})$.

All proofs are given in Appendix.

Hence, if we wish to relax the assumptions, we need to bring in additional information. Such additional information can come in the form of pretreatment variables, that is, variables that are known not to be affected by the treatment. In that case we can relax the assumption by requiring only that the adoption date is completely random within subpopulations with the same values for the pre-treatment variables, using the generalized propensity score (Imbens, 2000). Additional information can also come in the form of restrictions on the potential outcomes or the treatment effects. The implications of such restrictions on the ability to relax the random adoption assumption is more complex, as discussed in more detail in Section 3.2.

Under Assumption 1 the marginal distribution of the adoption dates is fixed, and so also the fraction π_a is fixed in the repeated sampling thought experiment. This part of the set up is similar in spirit to fixing the number of treated units in the sample in a completely randomized experiment. It is convenient for obtaining finite sample results. Note that it implies that the adoption dates for units i and j are not independent. Note also that in the standard framework where the uncertainty arises solely from random sampling, this fraction does not remain constant in the repeated sampling thought experiment.

An important role in our analysis is played by what we label the *adjusted treatment*, adjusted for unit and time period averages:

$$\dot{W}_{it} \equiv W_{it} - \overline{W}_{\cdot t} - \overline{W}_{i\cdot} + \overline{W},$$

where $\overline{W}_{\cdot t} \equiv \sum_{i=1}^{N} W_{it}/N$, $\overline{W}_{i\cdot} \equiv \sum_{t=1}^{T} W_{it}/T$, and $\overline{W} \equiv \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}/(NT)$ are averages over units, time periods, and both, respectively. We can also write the adjusted treatment indicator as

$$\dot{W}_{it} = g(t, A_i),$$

where

$$g(t,a) \equiv \left(\mathbf{1}_{a\leq t} - \sum_{s\leq t} \pi_s \right) + \frac{1}{T}\left(a\mathbf{1}_{a\leq T} - \sum_{s=1}^{T} s\pi_s \right) + \frac{T+1}{T}\left(\mathbf{1}_{a=\infty} - \pi_\infty \right), \tag{3.1}$$

where, with some minor abuse of notation, we adopt the convention that $a\mathbf{1}_{a\leq T}$ is zero if $a = \infty$ Under Assumption 1, which fixes the marginal distribution of A_i, the sum $\sum_{i,t} \dot{W}_{it}^2$ is non-stochastic, even though the adjusted treatment $\dot{W}_{it}$ and its square $\dot{W}_{it}^2$ are stochastic. The fact that this sum is non-stochastic enables us to derive exact finite sample results for the standard DID estimator as discussed in Section 4. This is similar in spirit to the derivation of the exact variance for the estimator for the average treatment effect in completely randomized experiments when we fix the number of treated and controls.

3.2. Exclusion restrictions

The next two assumptions concern the potential outcomes. Their formulation does not involve the assignment mechanism, that is, the distribution of the adoption date. In that sense they are unlike the strict and weak exogeneity or no-feedback assumptions (Chamberlain, 1984; Engle et al., 1983; Chamberlain et al., 1993). In essence these are exclusion restrictions, assuming that particular causal effects are absent. Collectively these two assumptions imply that we can think of the treatment as a binary one, the only relevant component of the adoption date being whether a unit is exposed to the treatment at the time we measure the outcome. This rules out dynamic treatment effects of the type considered in Abbring and Heckman (2007) and Hernan and Robins (2020). Versions of such assumptions are also considered in the DID setting in Borusyak and Jaravel (2016), de Chaisemartin and d'Haultfœuille (2020), Sun and Abraham (2020), Hazlett and Xu (2018) and Imai and Kim (2019), where in the latter a graphical approach is taken in the spirit of the work by Pearl (2000).

The first of the two assumptions, and likely the more plausible of the two in practice, rules out effects of future adoption dates on current outcomes. More precisely, it assumes that if the policy has not been adopted yet, the exact future date of the adoption has no causal effect on potential outcomes for the current period.

Assumption 2 (*No Anticipation*)**.** For all units i, all time periods t, and for all adoption dates a, such that $a > t$,

$$Y_{it}(a) = Y_{it}(\infty).$$

We can also write this assumption as requiring that for all triples (i, t, a),

$$Y_{it}(a) = \mathbf{1}_{a\leq t} Y_{it}(a) + \mathbf{1}_{a>t} Y_{it}(\infty), \qquad \text{or} \quad \mathbf{1}_{a>t}\Big(Y_{it}(a) - Y_{it}(\infty)\Big) = 0.$$

This last representation shows most clearly how the assumption rules out certain causal effects. Note that this assumption does not involve the adoption date, and so does not restrict the distribution of the adoption dates. Violations of this assumption may arise if the policy is anticipated prior to its implementation (*e.g.,* Abbring and Van den Berg, 2003).

The next assumption is arguably much stronger. It asserts that for potential outcomes in period t it does not matter how long the unit has been exposed to the treatment, only whether the unit is exposed at time t.

Assumption 3 (*Invariance to History*)**.** For all units i, all time periods t, and for all adoption dates a, such that $a \leq t$,

$$Y_{it}(a) = Y_{it}(1).$$

This assumption can also be written as

$$Y_{it}(a) = \mathbf{1}_{a\leq t} Y_{it}(1) + \mathbf{1}_{a>t} Y_{it}(a), \qquad \text{or} \quad \mathbf{1}_{a\leq t}\Big(Y_{it}(a) - Y_{it}(1)\Big) = 0,$$

with again the last version of the assumption illustrating the exclusion restriction in this assumption. Again, the assumption does not rule out any correlation between the potential outcomes and the adoption date, only that there is no causal effect of an early adoption versus a later adoption on the outcome in period t, as long as adoption occurred before or on period t.

In general, this assumption is very strong. However, there are important cases where it may be more plausible. Suppose the units are clusters of individuals, where in each period we observe different sets of individuals. To be specific, suppose the units are states, the time periods are years, and outcome is the employment rate for twenty-five year olds, and the treatment is the presence or absence of some regulation, say a subsidy for college tuition. In that case it may well be reasonable to assume that the educational choices for students graduating high school in a particular state depends on what the prevailing subsidy is, but much less on the presence of subsidies in previous years.

If both the exclusion restrictions, that is, both Assumptions 2 and 3, hold, then the potential outcome $Y_{it}(a)$ can be indexed by the binary indicator $W(a, t) = \mathbf{1}_{a\leq t}$:

Lemma 2 (*Binary Treatment*)**.** *Suppose Assumptions 2 and 3 hold. Then for all units i, all time periods t and adoption dates* $a > a'$, (*i*)

$$Y_{it}(a') - Y_{it}(a) = \mathbf{1}_{a'\leq t<a}\Big(Y_{it}(1) - Y_{it}(\infty)\Big),$$

so that,

$$Y_{it}(a) = Y_{it}(\infty) + \mathbf{1}_{a\leq t}\Big(Y_{it}(1) - Y_{it}(\infty)\Big) = \begin{cases} Y_{it}(\infty) & \text{if } a \leq t \\ Y_{it}(1) & \text{otherwise,} \end{cases}$$

and, for all time periods t, and adoption dates $a > a'$, (*ii*)

$$\tau_{t,aa'} = \tau_{t,\infty 1}\mathbf{1}_{a'\leq t<a} = \begin{cases} \tau_{t,\infty 1} & \text{if } a' \leq t < a, \\ 0 & \text{otherwise.} \end{cases}$$

If these two assumptions hold, we can therefore simplify the notation for the potential outcomes and focus on the pair of potential outcomes $Y_{it}(1)$ and $Y_{it}(\infty)$.

Note that these two assumptions are substantive, and because they only involve the potential outcomes and not the adoption date, they cannot be guaranteed by design. This in contrast to the Assumption 1, which can be guaranteed by randomization of the adoption date. It is also important to note that in many empirical studies Assumptions 2 and 3 are made, often implicitly by writing a model for realized outcome Y_{it} that depends solely on the contemporaneous treatment exposure W_{it}, and not on the actual adoption date A_i or treatment exposure $W_{it'}$ in other periods t'. In the current discussion we want to be explicit about the fact that this restriction is an assumption, and that it does not automatically hold. Note that the assumption does not restrict the time series dependence between the potential outcomes.

It is trivial to see that without additional information, the exclusion restrictions in Assumptions 2 and 3 have no testable implications because they impose restrictions on pairs of potential outcomes that can never be observed together. However, in combination with random assignment (Assumption 1), the two exclusion restrictions, Assumptions 2 and 3, have testable implications as long as $T \geq 2$ and there is some variation in the adoption date.

Lemma 3 (*Testable Restrictions from the Exclusion Restrictions*)**.**
(*i*) *Assumptions 2 and 3 jointly have no testable implications for the joint distribution of* $(\boldsymbol{Y}, \boldsymbol{W})$.
(*ii*) *Suppose* $T \geq 2$, *and* $\pi_2, \pi_\infty > 0$. *Then the combination of the random adoption date and the exclusion restrictions, Assumptions 1–3, impose testable restrictions on the joint distribution of* $(\boldsymbol{Y}, \boldsymbol{W})$.

This implies that if we maintain, say, the no-anticipation assumption, we can relax the random adoption date assumption. For example, we can allow the probability of adoption at time t to depend on the outcomes prior to period t. This type of feedback or violation of strict exogeneity (citetchamberlain1993feedback) is often important in practice (*e.g.*, Chay and Greenstone, 2005). Here we focus on the random-adoption-date case as a first step towards a design-based approach in the panel data case.

3.3. Auxiliary assumptions

In this section we consider four auxiliary assumptions that are convenient for some analyses, and in particular can have implications for the variance of specific estimators, but that are not essential in many cases. These assumptions are often made in empirical analyses without researchers explicitly discussing them.

The first of these assumptions assumes that the effect of adoption date a', relative to adoption date a, on the outcome in period t, is the same for all units.

Assumption 4 (*Constant Treatment Effect Over Units*)**.** For all units i, j and for all time periods t and all adoption dates a and a'

$$Y_{it}(a') - Y_{it}(a) = Y_{jt}(a') - Y_{jt}(a).$$

The second assumption restricts the heterogeneity of the treatment effects over time.

Assumption 5 (*Constant Treatment Effect over Time*)**.** For all units i and all time periods t and t'

$$Y_{it}(1) - Y_{it}(\infty) = Y_{it'}(1) - Y_{it'}(\infty).$$

We only restrict the time variation for comparisons of the adoption dates 1 and ∞ because we typically use this assumption in combination with Assumptions 2 and 3. In that case we obtain a constant binary treatment effect set up, as summarized in the following Lemma.

Lemma 4 (*Binary Treatment and Constant Treatment Effects*)**.** *Suppose Assumptions 2–5 hold. Then for all* t *and* $a' < a$

$$Y_{it}(a') - Y_{it}(a) = \mathbf{1}_{a' \leq t < a} \tau_{1\infty}.$$

The next assumption allows us to view the potential outcomes as random by postulating a large population from which the sample is drawn.

Assumption 6 (*Random Sampling*)**.** The sample can be viewed as a random sampling from an infinitely large population, with joint distribution for $(A_i, Y_{it}(a), a \in \mathbb{A}, t \in \mathbb{T})$ denoted by $f(a, y_1(1), \ldots, y_T(\infty))$.

Under the random sampling assumption we can put additional structure on average potential outcomes.

Assumption 7 (*Additivity*)**.**

$$\mathbb{E}[Y_{it}(\infty)] = \alpha_i + \beta_t.$$

4. Difference-In-Differences estimators: Interpretation and inference

In this section we consider the standard Differences-In-Differences (DID) set up (*e.g.*, Meyer et al., 1995; Bertrand et al., 2004; Angrist and Pischke, 2008; Donald and Lang, 2007; de Chaisemartin and d'Haultfœuille, 2020). In the simplest setting with N units and T time periods, without additional covariates, the realized outcome in period t for unit i is modelled as

$$Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \varepsilon_{it}. \tag{4.1}$$

In this model there are unit effects α_i and time effects β_t, but both are additive with interactions between them ruled out. The effect of the treatment is implicitly assumed to be additive and constant across units and time periods.

We interpret the DID estimand under the randomized adoption date assumption, leading to a different setting from that considered in de Chaisemartin and d'Haultfœuille (2020), Callaway and Sant'Anna (2020), Sun and Abraham (2020) and Goodman-Bacon (2018). We also derive its variance and show that in general it is lower than the standard random-sampling based variance. Finally we propose a variance estimator that is smaller than the regular variance estimators such as the Liang–Zeger and clustered bootstrap variance estimators.

4.1. Difference-In-Differences estimators

Consider the least squares estimator for τ based on the specification in (4.1):

$$\left(\hat{\tau}_{\text{did}}, \{\hat{\alpha}_i\}_{i=2}^N, \{\hat{\beta}_t\}_{t=1}^T\right) = \arg\min_{\tau, \{\alpha_i\}_{i=2}^N, \{\beta_t\}_{t=1}^T} \sum_{i=1}^N \sum_{t=1}^T \Big(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\Big)^2.$$

It is convenient to write the least squares estimator for the treatment effect in terms of the adjusted treatment indicator $\dot{W}_{it}$ as

$$\hat{\tau}_{\text{did}} = \frac{\sum_{i,t} \dot{W}_{it} Y_{it}}{\sum_{i,t} \dot{W}_{it}^2}.$$

The primary question of interest in this section concerns the properties of the estimator $\hat{\tau}_{\text{did}}$. This includes the interpretation of its expectation under various sets of assumptions, and its variance. Mostly we focus on exact properties in finite samples. Note that under Assumption 1, the denominator of $\hat{\tau}_{\text{did}}$ is non-stochastic.

In order to interpret the expected value of $\hat{\tau}_{\text{did}}$ we consider some intermediate objects. Define, for all adoption dates $a \in \mathbb{A}$, and all time periods $t \in \mathbb{T}$ the average outcome in period t for units with adoption date a:

$$\overline{Y}_{t,a} = \begin{cases} \frac{1}{N_a} \sum_{i:A_i=a} Y_{it} & \text{if } N_a > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Under Assumption 1 the stochastic properties of these averages are well-defined because the N_a are fixed over the randomization distribution. The averages are stochastic because the realized outcomes depend on the adoption date. Define also the following two difference between outcome averages:

$$\hat{\tau}_{t,aa'} = \overline{Y}_{t,a'} - \overline{Y}_{t,a}.$$

Given the random adoption date assumption (Assumption 1) these differences estimate average causal effects.

Example. To facilitate the interpretation of some of the results it is useful to consider a special case where the results from completely randomized experiments directly apply. Suppose $\mathbb{T} = \{1, 2\}$, and $\mathbb{A} = \{2, \infty\}$, with a fraction $\pi = \pi_2 = 1 - \pi_\infty$ adopting the policy in the second period. Suppose also that the first period outcome is zero for all units and all adoption dates, $Y_{i1}(a) = 0$ for all i and a, so that we can directly apply cross-section results. Then the DID estimator is

$$\hat{\tau}_{\text{did}} = \hat{\tau}_{2,2\infty} = \overline{Y}_{2,2} - \overline{Y}_{2,\infty} = \frac{1}{N_2} \sum_{i:A_i=2} Y_{i2} - \frac{1}{N_\infty} \sum_{i:A_i=\infty} Y_{i2},$$

the simple difference in means for the second period outcomes for adopters and non-adopters. Under Assumption 1, the standard results for the variance of the difference in means for a randomized experiments apply (e.g., Neyman, 1923/1990; Imbens and Rubin, 2015), and the exact variance of $\hat{\tau}_{\text{did}}$ is,

$$\mathbb{V}(\hat{\tau}_{\text{did}}) = \frac{1}{N_2(N-1)} \sum_{i=1}^N \Big\{Y_{i2}(2) - \overline{Y}_2(2)\Big\}^2 + \frac{1}{N_\infty(N-1)} \sum_{i=1}^N \Big\{Y_{i2}(\infty) - \overline{Y}_2(\infty)\Big\}^2$$

$$- \frac{1}{N(N-1)} \sum_{i=1}^N \Big\{\big(Y_{i2}(2) - \overline{Y}_2(2)\big) - \big(Y_{i2}(\infty) - \overline{Y}_2(\infty)\big)\Big\}^2.$$

The standard Neyman estimator for this variance ignores the third term, and uses unbiased estimators for the first two terms, leading to:

$$\hat{\mathbb{V}}(\hat{\tau}_{\text{did}}) = \frac{1}{N_2(N_2-1)} \sum_{i:A_i=2} \left\{Y_{i2} - \overline{Y}_{2,2}\right\}^2 + \frac{1}{N_\infty(N_\infty-1)} \sum_{i:A_i=\infty} \left\{Y_{i2} - \overline{Y}_{2,\infty}\right\}^2,$$

with $\mathbb{E}[\hat{\mathbb{V}}(\hat{\tau}_{\text{did}})] \geq \mathbb{V}(\hat{\tau}_{\text{did}})$. □

4.2. The interpretation of Difference-In-Differences estimators

The following weights play an important role in the interpretation of the DID estimand:

$$\gamma_{t,a} \equiv \frac{\pi_a g(t,a)}{\sum_{t'\in\mathbb{T}} \sum_{a'\in\mathbb{A}} \pi_{a'} g(t',a')^2}, \qquad \gamma_{t,+} \equiv \sum_{a\leq t} \gamma_{t,a}, \qquad \text{and} \quad \gamma_{t,-} \equiv \sum_{a>t} \gamma_{t,a}, \tag{4.2}$$

with $g(t,a)$ as defined in (3.1). Note that these weights are non-stochastic, that is, fixed over the randomization distribution.

Example (*Ctd*)**.** Continuing the example with two periods and adoption in the second period or never. Then we have

$$\gamma_{t,a} = \begin{cases} 0 & \text{if } (t,a) = (1,1), \\ 0 & \text{if } (t,a) = (2,1), \\ -1 & \text{if } (t,a) = (1,2), \\ 1 & \text{if } (t,a) = (2,2), \\ 1 & \text{if } (t,a) = (1,\infty), \\ -1 & \text{if } (t,a) = (2,\infty), \end{cases} \qquad \gamma_{t,+} = \begin{cases} 0 & \text{if } t=1, \\ 1 & \text{if } t=2, \end{cases} \qquad \text{and} \quad \gamma_{t,-} = \begin{cases} 0 & \text{if } t=1, \\ -1 & \text{if } t=2. \end{cases} \quad \square$$

In general the weights $\gamma_{t,a}$ have some important properties,

$$\sum_{t\in\mathbb{T}} \gamma_{t,+} = 1 \qquad \sum_{t\in\mathbb{T}} \gamma_{t,-} = -1, \qquad \text{and} \quad \sum_{t=1}^{T} \sum_{a\in\mathbb{A}} \gamma_{t,a} = \sum_{t\in\mathbb{T}} \gamma_{t,+} + \sum_{t\in\mathbb{T}} \gamma_{t,-} = 0.$$

Now we can state the first main result of the paper.

Lemma 5. *We can write* $\hat{\tau}_{\text{did}}$ *as*

$$\hat{\tau}_{\text{did}} = \sum_{t\in\mathbb{T}} \sum_{a\in\mathbb{A}} \gamma_{t,a} \overline{Y}_{t,a} = \sum_{t\in\mathbb{T}} \gamma_{t,+} \hat{\tau}_{t,\infty 1} + \sum_{t\in\mathbb{T}} \sum_{a>t} \gamma_{t,a} \hat{\tau}_{t,\infty a} - \sum_{t\in\mathbb{T}} \sum_{a\leq t} \gamma_{t,a} \hat{\tau}_{t,a1}. \tag{4.3}$$

Comment 1. Alternative characterizations of the DID estimator or estimand as a weighted average of potentially causal comparisons are presented in Sun and Abraham (2020), de Chaisemartin and d'Haultfœuille (2020), Han (2020), Goodman-Bacon (2018), Imai and Kim (2019), Borusyak and Jaravel (2016), and Kim et al. (2019). The characterizations differ in terms of the building blocks that are used in the representation and the assumptions made. Like our representation, the representation in Sun and Abraham (2020) is in terms of average causal effects of different adoption dates, but it imposes no-anticipation. Goodman-Bacon (2018) presents the DID estimator in terms of basic two-group DID estimators. Just like our representation, the Goodman-Bacon (2018) representation does not rely on any assumptions. To endow the building blocks and the representation itself with a causal interpretation requires some assumption on, for example, the assignment mechanism. □

Comment 2. The lemma implies that the DID estimator has an interpretation as a weighted average of simple estimators for the causal effect of changes in adoption dates, the $\hat{\tau}_{t,aa'}$. Moreover, the estimator can be written as the sum of three averages of these $\hat{\tau}_{t,aa'}$. The first is a weighted average of the $\hat{\tau}_{t,\infty 1}$, which are all averages of switching from never adopting to adopting in the first period, meaning that these are averages of changes in adoption dates that involve switching from not being treated at time t to being treated at time t. The sum of the weights for these averages is one, although some of the weights may be negative. The second sum is a weighted sum of $\hat{\tau}_{t,\infty a}$, for $a > t$, so that the causal effect always involves changing the adoption date from never adopting to adopting some time after t, meaning that the comparison is between potential outcomes neither of which involves being treated at the time. The sum of the weights for these averages is again one. The third sum is a weighted sum of $\hat{\tau}_{t,a1}$, for $a \leq t$, so that the causal effect always involves changing the adoption date from adopting prior to, or at time, t relative to adopting at the initial time, meaning that the comparison is between potential outcomes both of which involves being treated at the time. These weights sum to minus one. □

If we are willing to make the random adoption date assumption we can give this representation a causal interpretation:

Theorem 1. *Suppose Assumption 1 holds. Then (i):*

$$\mathbb{E}\left[\hat{\tau}_{t,aa'}\right] = \tau_{t,aa'},$$

and (ii)

$$\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] = \sum_{t\in\mathbb{T}} \gamma_{t,+}\tau_{t,\infty 1} + \sum_{t\in\mathbb{T}}\sum_{a>t} \gamma_{t,a}\tau_{t,\infty a} - \sum_{t\in\mathbb{T}}\sum_{a\leq t} \gamma_{t,a}\tau_{t,a1}.$$

Suppose both Assumptions 1 and 2 hold. Then (iii):

$$\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] = \sum_{t\in\mathbb{T}}\sum_{a\leq t} \gamma_{t,a}\tau_{t,\infty a} = \sum_{t\in\mathbb{T}} \gamma_{t,+}\tau_{t,\infty 1} - \sum_{t\in\mathbb{T}}\sum_{a\leq t} \gamma_{t,a}\tau_{t,a1}.$$

Suppose Assumptions 1–3 hold. Then (iv):

$$\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] = \sum_{t=1}^{T} \gamma_{t,+}\tau_{t,\infty 1}.$$

Suppose Assumptions 1–5 hold. Then (v):

$$\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] = \tau_{\infty 1}.$$

Part (*iii*) of the theorem where we make the no-anticipation assumption is closely related to one of the results in Sun and Abraham (2020), who make a super-population common trend assumption that, in the super-population context, weakens our random adoption date assumption. Part (*iv*) of the theorem, where we assume both the exclusion restrictions so that the treatment is effectively a binary one, is related to the results in de Chaisemartin and d'Haultfœuille (2020).

Without either Assumptions 2 or 3, the estimand τ_{did} has a causal interpretation, but it is not clear it is a very interesting one concerning the receipt of the treatment. With the no-anticipation assumption (Assumption 2), the interpretation, as given in part (*iii*) of the theorem, is substantially more interesting. Now the estimand is a weighted average of $\tau_{t,\infty a}$ for $a \leq t$, with weights summing to one. These $\tau_{t,\infty a}$ are the average causal effect of changing the adoption date from never adopting to some adoption date prior to, or equal to, time t, so that the average always involves switching from not being exposed to the treatment to being exposed to the treatment.

4.3. The randomization variance of the Difference-In-Differences estimators

In this section we derive the randomization variance for $\hat{\tau}_{\text{did}}$ under the randomized adoption date assumption. We do not rely on other assumptions here, although such assumptions may be required for making the estimand a substantively interesting one. The starting point is the representation $\hat{\tau}_{\text{did}} = \sum_{t,a} \gamma_{t,a}\overline{Y}_{t,a}$. Because under Assumption 1 the weights $\gamma_{t,a}$ are fixed, the variance is

$$\mathbb{V}(\hat{\tau}_{\text{did}}) = \sum_{t,a} \gamma_{t,a}^2 \mathbb{V}(\overline{Y}_{t,a}) + \sum_{(t,a)\neq(t',a')} \gamma_{t,a}\gamma_{t',a'}\mathbb{C}(\overline{Y}_{t,a}, \overline{Y}_{t',a'}).$$

Note that the $\gamma_{t,a}$ are known, so the sole challenge is to find estimators for the $\mathbb{V}(\overline{Y}_{t,a})$ and $\mathbb{C}(\overline{Y}_{t,a}, \overline{Y}_{t',a'})$. Working out the variance $\mathbb{V}(\overline{Y}_{t,a})$, and finding an unbiased estimator for it, is straightforward. It is more challenging to infer the covariance terms $\mathbb{C}(\overline{Y}_{t,a}, \overline{Y}_{t',a'})$, and even more difficult to estimate them without bias. In fact, in general that is not possible. Note that for a sampling-based variance the $\gamma_{t,a}$ are not fixed, because in different samples the fractions with a particular adoption date will be stochastic. This in general leads to a larger variance, as we verify in simulations.

Define

$$Y_i(a) = \sum_{t=1}^{T} \gamma_{t,a} Y_{it}(a), \qquad \overline{Y}(a) = \sum_{t=1}^{T} \gamma_{t,a}\overline{Y}_t(a) \qquad \text{and} \quad \overline{Y}_a = \sum_{t=1}^{T} \gamma_{t,a}\overline{Y}_{t,a}.$$

Now we can write $\hat{\tau}_{\text{did}}$ as

$$\hat{\tau}_{\text{did}} = \sum_{a\in\mathbb{A}}\sum_{t\in\mathbb{T}} \gamma_{t,a}\overline{Y}_{t,a} = \sum_{a\in\mathbb{A}} \overline{Y}_a.$$

Define also

$$S_a^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(Y_i(a) - \overline{Y}(a)\right)^2,$$

and

$$V_{a,a'}^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left\{\left(Y_i(a) - \overline{Y}(a)\right) + \left(Y_i(a') - \overline{Y}(a')\right)\right\}^2.$$

Theorem 2. *Suppose Assumption 1 holds. Then the exact variance of* $\hat{\tau}_{\text{did}}$ *over the randomization distribution is*

$$\mathbb{V}\left(\hat{\tau}_{\text{did}}\right) = \sum_{a\in\mathbb{A}} S_a^2 \left(\frac{1}{N_a} + \frac{T-1}{N}\right) - \sum_{a\in\mathbb{A}} \sum_{a'\in\mathbb{A},a'>a} \frac{V_{a,a'}^2}{N},$$

with

$$\mathbb{V}\left(\hat{\tau}_{\text{did}}\right) \leq \sum_{a\in\mathbb{A}} S_a^2/N_a.$$

Comment (*Ctd*)**.** In our two period example with some units adopting in the second period and the others not at all, and the first period outcome is zero, $Y_{i1}(a) = 0$, we have

$$\gamma_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \gamma_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \text{and} \quad \gamma_\infty = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

$$S_1^2 = 0,$$

$$S_{,2}^2 = \frac{1}{N-1}\sum_{i=1}^N \left(Y_{i2}(2) - \frac{1}{N}\sum_{j=1}^N Y_{j2}(2)\right)^2,$$

$$S_\infty^2 = \frac{1}{N-1}\sum_{i=1}^N \left(Y_{i2}(\infty) - \frac{1}{N}\sum_{j=1}^N Y_{j2}(\infty)\right)^2,$$

$$V_{1,2}^2 = 0, \qquad S_{1,\infty}^2 = 0,$$

$$V_{2,\infty}^2 = \frac{1}{N-1}\sum_{i=1}^N \left(\left(Y_{i2}(2) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(2)\right) - \left(Y_{i2}(\infty) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(\infty)\right)\right)^2,$$

so that in this special

$$\begin{aligned}\mathbb{V}(\hat{\tau}_{\text{did}}) = {} & \frac{1}{N_2(N-1)}\sum_{i=1}^N \left(Y_{i2}(2) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(2)\right)^2 \\ & + \frac{1}{N_\infty(N-1)}\sum_{i=1}^N \left(Y_{i2}(\infty) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(\infty)\right)^2 \\ & - \frac{1}{N(N-1)}\sum_{i=1}^N \left(\left(Y_{i2}(2) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(2)\right) - \left(Y_{i2}(\infty) - \frac{1}{N}\sum_{j=1}^N Y_{i2}(\infty)\right)\right)^2,\end{aligned}$$

which agrees with the Neyman variance for a completely randomized experiment. □

4.4. Estimating the randomization variance of the Difference-In-Differences estimators

In this section we discuss estimating the variance of the DID estimator. In general there is no unbiased estimator for $\mathbb{V}\left(\hat{\tau}_{\text{did}}\right)$. This is not surprising, because there is no such estimator for the simple difference in means estimator in a completely randomized experiment, and this corresponds to the special case with $T = 1$. However, it turns out that just like in the simple randomized experiment case, there is a conservative variance estimator. In the current case it is based on using unbiased estimators for the terms involving S_a^2, and ignoring the terms involving $V_{a,a'}^2$. Because the latter are non-negative, and always enter with a minus sign, ignoring them leads to an upwardly biased variance estimator. One difference with the simple randomized experiment case is that there is no simple case with constant treatment effects such that the variance estimator is unbiased.

Next, define the estimated variance of this by adoption date:

$$s_a^2 \equiv \frac{1}{N_a - 1}\sum_{i:A_i=a} \left(Y_i - \overline{Y}_a\right)^2.$$

Now we can characterize the proposed variance estimator as

$$\widehat{\mathbb{V}}_{\text{did}} \equiv \sum_{a\in\mathbb{A}} \frac{s_a^2}{N_a}.$$

Theorem 3. *Suppose Assumption 1 holds. Then*

$$\mathbb{E}\left[\widehat{\mathbb{V}}_{\text{did}}\right] \geq \mathbb{V}(\hat{\tau}_{\text{did}}),$$

so that $\widehat{\mathbb{V}}_{\text{did}}$ *is a conservative variance estimator for* $\hat{\tau}_{\text{did}}$.

There are two important issues regarding this variance estimator. The first is its relation to the standard variance estimator for DID estimators. The second is whether one can improve on this variance estimator given that in general it is conservative.

The relevant variance estimators are the Liang–Zeger clustered variance estimator and the clustered bootstrap (Liang and Zeger, 1986; Chamberlain, 1984; Arellano, 1987a; Bertrand et al., 2004). Both have large sample justifications under random sampling from a large population, so they are in general not equal to the variance estimator here. In large samples both the Liang–Zeger and bootstrap variance will be more conservative than $\hat{\mathbb{V}}_{\text{did}}$ because they also take into account variation in the weights $\gamma_{t,a}$. These weights are kept fixed under the randomization scheme, because that keeps fixed the marginal distribution of the adoption dates. In contrast, under the Liang–Zeger calculations and the clustered bootstrap, the fraction of units with a particular adoption date varies, and that introduces additional uncertainty.

The second issue is whether we can improve on the conservative variance estimator $\hat{\mathbb{V}}_{\text{did}}$. In general there is only a limited ability to do so. Note, for example, that in the two period example this variance reduces to the Neyman variance in randomized experiments. In that case we know we can improve on this variance a little bit exploiting heteroskedasticity, e.g., Aronow et al. (2014), but in general those gains are modest.

5. Some simulations

The goal is to compare the exact variance, and the corresponding estimator in the paper to the two leading alternatives, the Liang–Zeger (stata) clustered standard errors and the clustered bootstrap. We want to demonstrate settings where the proposed variance estimator differs from the Liang–Zeger clustered variance, and settings where it is the same. We have N units, observed for T time periods. We focus on the case with $T = 3$. The adoption date is randomly assigned, with two distributions for the adoption date, $\pi_I = (\pi_1, \pi_2, \pi_3, \pi_\infty) = (0, 0.67, 0, 0.33)$, and $\pi_{II} = (\pi_1, \pi_2, \pi_3, \pi_\infty) = (0, 0.5, 0.4, 0.1)$.

We consider four designs for the potential outcome distributions in the population, the $Y_i(a)$ for $a \in \{1, 2, 3, \infty\}$. In Design A the potential outcomes, are generated as

$$\begin{pmatrix} Y_{i1}(2) \\ Y_{i1}(3) \\ Y_{i1}(\infty) \\ Y_{i2}(2) \\ Y_{i2}(3) \\ Y_{i2}(\infty) \\ Y_{i3}(2) \\ Y_{i3}(3) \\ Y_{i3}(\infty) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 3 \\ 3 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

In this design Assumptions 1–5 hold: the treatment effect is constant, and depends only on whether the adoption date precedes the potential outcome date, or

$$Y_{it}(a) = \mathbf{1}_{a\leq t} + \varepsilon_{it},$$

where the ε_{it} are correlated over time.

In Design B the potential outcomes are generated as

$$\begin{pmatrix} Y_{i1}(2) \\ Y_{i1}(3) \\ Y_{i1}(\infty) \\ Y_{i2}(2) \\ Y_{i2}(3) \\ Y_{i2}(\infty) \\ Y_{i3}(2) \\ Y_{i3}(3) \\ Y_{i3}(\infty) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 1 \\ 1 \\ 2 \\ 11 \\ 1 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

Here the presence of treatment effects requires the treatment having been adopted, but the effects vary by the adoption date, so that Assumption 3 does not hold.

In Design C the potential outcomes are generated with positive correlations between the potential outcomes as

$$\begin{pmatrix} Y_{i1}(2) \\ Y_{i1}(3) \\ Y_{i1}(\infty) \\ Y_{i2}(2) \\ Y_{i2}(3) \\ Y_{i2}(\infty) \\ Y_{i3}(2) \\ Y_{i3}(3) \\ Y_{i3}(\infty) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 1 \\ 1 \\ 2 \\ 11 \\ 1 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0.9 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.9 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 1 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0.9 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 1 \end{pmatrix} \right).$$

In Design D the potential outcomes are generated with negative correlations between the potential outcomes as

$$\begin{pmatrix} Y_{i1}(2) \\ Y_{i1}(3) \\ Y_{i1}(\infty) \\ Y_{i2}(2) \\ Y_{i2}(3) \\ Y_{i2}(\infty) \\ Y_{i3}(2) \\ Y_{i3}(3) \\ Y_{i3}(\infty) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ 1 \\ 1 \\ 2 \\ 11 \\ 1 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & -0.4 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.4 & 1 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.4 & -0.4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -0.4 & -0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.4 & 1 & -0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.4 & -0.4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -0.4 & -0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.4 & 1 & -0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.4 & -0.4 & 1 \end{pmatrix} \right).$$

For a particular design, eg (A, II) draw the four sets of three-component vectors of potential outcomes for each unit (the three components corresponding to the three time periods), one set for each of the values of $a \in \{1, 2, 3, \infty\}$. We keep these sets of potential outcomes fixed across all simulations for a given design. Then for each simulation draw the adoption date according to the distribution for that design, keeping the fraction of units with a particular adoption date fixed.

We want to look at variances and the corresponding confidence intervals based on four methods for estimating the variance for the DID estimator. The confidence intervals are Normal-distribution based, simply equal to the point estimates plus and minus 1.96 times the square root of the variances. We can write $\hat{\tau}_{\text{did}}$ as a regression estimator with NT observations, and $N + T$ regressors. Let with $j = 1, \ldots, NT$. For observation j, $T_j \in \{1, \ldots, T\}$ denotes the time period the observation is from, and $N_j \in \{1, \ldots, N\}$ denotes the unit is corresponds to. Now let $Y_j = Y_{N_j,T_j}$ and $W_j = W_{N_j,T_j}$, so that the regression function can be written as

$$Y_j = \mu + \sum_{n=1}^{N-1} \alpha_n \mathbf{1}_{N_j=n} + \sum_{t=1}^{T-1} \beta_t \mathbf{1}_{T_j=t} + \tau W_j + \varepsilon_j = Y_j = X_j^\top \theta + \varepsilon_j,$$

where $X_j = (1, \mathbf{1}_{N_j=1}, \ldots, \mathbf{1}_{N_j=N-1}, \mathbf{1}_{T_j=1}, \ldots, \mathbf{1}_{T_j=T-1}, W_j)$, and $\theta = (\mu, \alpha_1, \ldots, \alpha_{N-1}, \beta_1, \ldots, \beta_{T-1}, \tau)$.

We compare five variances. The first is the (infeasible) exact randomization-based variance,

$$\mathbb{V}_{\text{did}} = \mathbb{V}\left(\hat{\tau}\right) = \sum_{a \in \mathbb{A}} \frac{S^2_{\gamma a, a}}{N_a} - \sum_{a \in \mathbb{A}} \sum_{a' \in \mathbb{A}, a' > a} \frac{V^2_{\gamma_a, a, \gamma_{a'}, a'}}{N}.$$

The other four are estimators of the variance. First, the conservative variance estimator $\widehat{\mathbb{V}}_{\text{did}}$,

$$\widehat{\mathbb{V}}_{\text{did}} \equiv \sum_{a \in \mathbb{A}} \frac{1}{N_a(N_a - 1)} \sum_{i: A_i = a} \left(Y_i - \overline{Y}_a\right)^2.$$

Second, the standard Liang–Zeger clustered variance. Start with the regression representation of the estimator, based on the regression $Y_j = X_j^\top \theta + \varepsilon_j$ with the covariates including the unit and time fixed effects and the treatment indicator. Let $\hat{\varepsilon}_j = Y_j - X_j^\top \hat{\theta}$ be the residual from this regression. Calculate the variance as

$$\widehat{\mathbb{V}}_{\text{LZ}} = \left(\sum_{j=1}^{J} X_j X_j^\top \right)^{-1} \left(\sum_{n=1}^{N} \left(\sum_{j: N_j = n} X_j \hat{\varepsilon}_j \right) \left(\sum_{j: N_j = n} X_j \hat{\varepsilon}_j \right)^\top \right) \left(\sum_{j=1}^{J} X_j X_j^\top \right)^{-1}.$$

We then use the component of this variance/covariance matrix corresponding to the estimator for the treatment effect, $\hat{\tau}_{\text{did}}$.

The last two variance estimators are two versions of the clustered bootstrap. First, the standard clustered bootstrap, $\widehat{\mathbb{V}}_{B1}$. Draw bootstrap samples based on drawing units, with all time periods for each unit drawn. Note that this explicitly

Table 1
Simulations: Variances and coverage rates for 95% confidence intervals.

Design	π	N	$\mathbb{V}_{\text{did}}$	Cov	$\hat{\mathbb{V}}_{\text{did}}$	Cov	$\hat{\mathbb{V}}_{\text{LZ}}$	Cov	$\hat{\mathbb{V}}_{\text{B1}}$	Cov	$\hat{\mathbb{V}}_{\text{B2}}$	Cov
A	I	30	0.144	0.951	0.239	0.979	0.214	0.974	0.232	0.975	0.219	0.973
B	I	30	0.111	0.947	0.187	0.986	0.163	0.978	0.182	0.982	0.172	0.978
C	I	30	0.201	0.953	0.217	0.947	0.181	0.925	0.211	0.942	0.200	0.932
D	I	30	0.064	0.949	0.265	1.000	0.230	0.999	0.257	1.000	0.244	0.999
A	II	30	0.112	0.946	0.165	0.972	0.146	0.966	0.158	0.969	0.142	0.956
B	II	30	0.085	0.947	0.139	0.973	0.268	0.999	0.269	0.999	0.119	0.962
C	II	30	0.184	0.949	0.191	0.939	0.279	0.983	0.285	0.981	0.162	0.920
D	II	30	0.081	0.950	0.164	0.992	0.285	1.000	0.280	0.999	0.142	0.987
A	I	150	0.027	0.953	0.047	0.991	0.045	0.989	0.047	0.989	0.046	0.989
B	I	150	0.022	0.955	0.041	0.994	0.039	0.992	0.041	0.992	0.041	0.992
C	I	150	0.035	0.956	0.038	0.960	0.036	0.956	0.037	0.955	0.037	0.954
D	I	150	0.019	0.950	0.044	0.997	0.044	0.997	0.044	0.996	0.043	0.995
A	II	150	0.020	0.952	0.033	0.989	0.033	0.989	0.033	0.987	0.032	0.987
B	II	150	0.021	0.945	0.036	0.985	0.053	0.997	0.052	0.997	0.035	0.984
C	II	150	0.034	0.952	0.035	0.953	0.051	0.985	0.052	0.983	0.034	0.947
D	II	150	0.016	0.950	0.028	0.990	0.044	0.998	0.044	0.998	0.028	0.987

changes from bootstrap sample to bootstrap sample the fraction of units with a particular adoption date. Second, a modification (improvement) of the standard clustered bootstrap, $\widehat{\mathbb{V}}_{B2}$, where we fix the fraction of units with each value for the adoption date.

In Table 1 we report the results. For each of the five variances we report the average of variance, and the coverage rate for the 95% confidence interval.

We see that the standard Liang–Zeger and the clustered bootstrap ($\widehat{\mathbb{V}}_{B1}$) substantially over-estimate the variance in Design B. The fixed adoption date bootstrap ($\widehat{\mathbb{V}}_{B2}$) and the proposed variance estimator ($\widehat{\mathbb{V}}_{\text{did}}$) have the appropriate coverage.

6. Conclusion

We discuss a design-based approach to Difference-In-Differences estimation in a setting with staggered adoption and argue that this clarifies the properties of DID estimators. We characterize what the standard DID estimator is estimating under a random adoption date assumption, and what the exact finite sample variance of the standard estimator is. We show that the standard DID estimand is a weighted average of different types of causal effects, for example, the effect of changing from never adopting to adopting in the first period, or changing from never adopting to adopting later. In this approach the standard Liang–Zeger and clustered bootstrap variance estimators are conservative, similar to standard variance estimators in randomized experiments. In contrast to simple randomized experiments, however, one can improve systematically on the standard variance estimator, and we do so with an improved variance estimator, $\widehat{\mathbb{V}}_{B2}$, the bootstrap where we keep the distribution of the adoption dates fixed.

Appendix

Proof of Lemma 1. Let $\boldsymbol{Y}^p$ denote the $N \times (T \cdot (T+1))$ dimensional matrix with all the potential outcomes. Because the units are exchangeable we can write the joint distribution of the potential outcomes and $\boldsymbol{A}$ as

$$f(\boldsymbol{Y}^p, \boldsymbol{A}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_i^p, A_i).$$

Now we shall construct a distribution $f(\boldsymbol{Y}_i^p, A_i)$ that satisfies two conditions. First, A_i is independent of all the potential outcomes and second, the implied distribution for the adoption date and the realized outcome is consistent with the actual distribution. To do so we assume independence of the sets potential outcomes $Y_{i1}(a), \ldots, f_{iT}(a)$ for different a, and assume that

$$f(Y_{i1}(a), \ldots, f_{iT}(a)) = f(Y_{i1}(a), \ldots, f_{iT}(a)|A_i = a) = f(Y_{i1}, \ldots, f_{iT}|A_i = a). \quad \square$$

Proof of Lemma 2. By Assumption 2 we have

$$Y_{it}(a) = \mathbf{1}_{a \le t} Y_{it}(a) + \mathbf{1}_{a > t} Y_{it}(\infty),$$

and by Assumption 3 we have

$$Y_{it}(a) = \mathbf{1}_{a \le t} Y_{it}(1) + \mathbf{1}_{a > t} Y_{it}(a).$$

Combining the two assumptions implies

$$Y_{it}(a) = \mathbf{1}_{a\leq t} Y_{it}(1) + \mathbf{1}_{a>t} Y_{it}(\infty).$$

Hence

$$\begin{aligned} Y_{it}(a') - Y_{it}(a) &= \mathbf{1}_{a'\leq t} Y_{it}(1) + \mathbf{1}_{a'>t} Y_{it}(\infty) - \left(\mathbf{1}_{a\leq t} Y_{it}(1) + \mathbf{1}_{a>t} Y_{it}(\infty)\right) \\ &= \mathbf{1}_{a'\leq t<t} \left(Y_{it}(1) - Y_{it}(\infty)\right), \end{aligned}$$

which proves part (*i*).

For part (*ii*)

$$\begin{aligned} \tau_{t,aa'} &= \frac{1}{N}\sum_{i=1}^{N}\Big(Y_{it}(a') - Y_{it}(a)\Big) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{a'\leq t<t} \left(Y_{it}(1) - Y_{it}(\infty)\right) \\ &= \mathbf{1}_{a'\leq t<t} \frac{1}{N}\sum_{i=1}^{N} \left(Y_{it}(1) - Y_{it}(\infty)\right) = \mathbf{1}_{a\leq t<a'} \tau_{t,\infty 1}. \quad \square \end{aligned}$$

Proof of Lemma 3. Part (*i*) follows directly from the fact that the exclusion restrictions place restrictions only on potential outcomes that cannot be observed together.

Let us turn to part (*ii*). By assumption

$$Y_{it}(a) \perp\!\!\!\perp A_i,$$

which as a special case includes

$$Y_{i1}(\infty) \perp\!\!\!\perp A_i.$$

Hence

$$Y_{i1}(\infty) \perp\!\!\!\perp A_i \,\Big|\, A_i \geq 2$$

which implies

$$Y_{i1} \perp\!\!\!\perp A_i \,\Big|\, A_i \geq 2$$

and thus

$$Y_{i1} \perp\!\!\!\perp A_i \,\Big|\, A_i \in \{2, \infty\},$$

which is a testable restriction. $\square$

Proof of Lemma 4. By Assumptions 2 and 3 we have

$$Y_{it}(a) - Y_{it}(\infty) = \mathbf{1}_{a\leq t}\Big(Y_{it}(1) - Y_{it}(\infty)\Big).$$

By Assumptions 4 and 5, $Y_{it}(1) - Y_{it}(\infty) = \tau_{1\infty}$, so that

$$Y_{it}(a) - Y_{it}(\infty) = \mathbf{1}_{a\leq t}\tau_{1\infty}. \quad \square$$

Proof of Lemma 5. Using the definition for $g(t, a)$, we can write $\hat{\tau}_{\text{did}}$ as

$$\begin{aligned} \hat{\tau}_{\text{did}} &= \frac{\sum_{i,t} \dot{W}_{it} Y_{it}}{\sum_{i,t} \dot{W}_{it}^2} = \frac{\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}}\sum_{i:A_i=a} \dot{W}_{it} Y_{it}}{N\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}} \pi_a g(a,t)^2} = \frac{\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}}\sum_{i:A_i=a} g(a,t) Y_{it}}{N\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}} \pi_a g(a,t)^2} \\ &= \frac{\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}}\sum_{i:A_i=a} g(a,t) N_a \overline{Y}_{t,a}}{N\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}} \pi_a g(a,t)^2} \\ &= \frac{\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}} g(a,t)\pi_a \overline{Y}_{t,a}}{\sum_{t\in\mathbb{T}}\sum_{a\in\mathbb{A}} \pi_a g(a,t)^2} = \sum_{t,a} \gamma_{t,a} \overline{Y}_{t,a}, \end{aligned}$$

where $\gamma_{t,a}$ is as given in (4.2). $\square$

Proof of Theorem 1. First consider part (*i*). We will show that

$$\mathbb{E}[\overline{Y}_{ta}] = \overline{Y}_t(a),$$

which in turn implies the result in (*i*). We can write

$$\mathbb{E}[\overline{Y}_{ta}] = \mathbb{E}\left[\frac{1}{N_a}\sum_{i=1}^{N}\mathbf{1}_{A_i=a}Y_{it}\right] = \mathbb{E}\left[\frac{1}{N_a}\sum_{i=1}^{N}\mathbf{1}_{A_i=a}Y_{it}(a)\right].$$

By Assumption 1 this is equal to

$$\frac{1}{N_a}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{1}_{A_i=a}\right]Y_{it}(a) = \frac{1}{N_a}\sum_{i=1}^{N}\frac{N_a}{N}Y_{it}(a) = \frac{1}{N}\sum_{i=1}^{N}Y_{it}(a) = \overline{Y}_t(a),$$

which is the desired result.
Next consider part (*ii*). By Lemma 5,

$$\hat{\tau}_{\text{did}} = \sum_{t\in\mathbb{T}}\gamma_{t,+}\hat{\tau}_{t,\infty 1} + \sum_{t\in\mathbb{T}}\sum_{a>t}\gamma_{t,a}\hat{\tau}_{t,\infty a} - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\hat{\tau}_{t,a1},$$

so that

$$\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] = \mathbb{E}\left[\sum_{t\in\mathbb{T}}\gamma_{t,+}\hat{\tau}_{t,\infty 1} + \sum_{t\in\mathbb{T}}\sum_{a>t}\gamma_{t,a}\hat{\tau}_{t,\infty a} - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\hat{\tau}_{t,a1}\right],$$

which by Assumption 1 is equal to

$$\sum_{t\in\mathbb{T}}\gamma_{t,+}\mathbb{E}\left[\hat{\tau}_{t,\infty 1}\right] + \sum_{t\in\mathbb{T}}\sum_{a>t}\gamma_{t,a}\mathbb{E}\left[\hat{\tau}_{t,\infty a}\right] - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\mathbb{E}\left[\hat{\tau}_{t,a1}\right].$$

This in turn, by part (*i*), is equal to

$$\sum_{t\in\mathbb{T}}\gamma_{t,+}\tau_{t,\infty 1} + \sum_{t\in\mathbb{T}}\sum_{a>t}\gamma_{t,a}\tau_{t,\infty a} - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\tau_{t,a1},$$

which finishes the proof of part (*ii*).
Next consider part (*iii*). If Assumption 2 holds, then for all $a > t$, $\tau_{t,\infty a} = 0$, so that

$$\begin{aligned}\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] &= \sum_{t\in\mathbb{T}}\gamma_{t,+}\tau_{t,\infty 1} - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\tau_{t,a1}\\ &= \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\tau_{t,\infty a}.\end{aligned}$$

Next consider part (*iv*). If also Assumption 3 holds, then also for all $a \leq t$, $\tau_{t,a1} = 0$, so that

$$\begin{aligned}\mathbb{E}\left[\hat{\tau}_{\text{did}}\right] &= \sum_{t\in\mathbb{T}}\gamma_{t,+}\tau_{t,\infty 1} + \sum_{t\in\mathbb{T}}\sum_{a>t}\gamma_{t,a}\tau_{t,\infty a} - \sum_{t\in\mathbb{T}}\sum_{a\leq t}\gamma_{t,a}\tau_{t,a1}\\ &= \sum_{t\in\mathbb{T}}\gamma_{t,+}\tau_{t,\infty 1}.\end{aligned}$$

Finally, consider part (*v*). This follows directly from part (*iv*) in combination with the constant treatment effect assumption (Assumption 5). □

Next we give a preliminary result.

Lemma A.1. *Suppose that Assumption 1 holds. Then* (*i*) *the variance of* $\overline{Y}_a$ *is*

$$\mathbb{V}(\overline{Y}_a) = \frac{S_a^2}{N_a}\left(1 - \frac{N_a}{N}\right),$$

(ii), the covariance of $\overline{Y}_a$ *and* $\overline{Y}_{a'}$ *is*

$$\mathbb{C}(\overline{Y}_a, \overline{Y}_{a'}) = -\frac{1}{2N}\left(S_a^2 + S_{a'}^2 - S_{aa'}^2\right) = \frac{1}{2N}\left(S_a^2 + S_{a'}^2 - V_{aa'}^2\right),$$

(*iii*), *the variance of the sum of the* $\overline{Y}_a$ *is*

$$\mathbb{V}\left(\sum_{a\in\mathbb{A}}\overline{Y}_a\right) = \sum_{a\in\mathbb{A}}S_a^2\left(\frac{1}{N_a} + \frac{T-1}{N}\right) - \frac{1}{2N}\sum_{a,a':a\neq a'}V_{aa'}^2,$$

and (iv),

$$\mathbb{V}\left(\sum_{a\in\mathbb{A}}\overline{Y}_a\right)\leq\sum_{a\in\mathbb{A}}\frac{S_a^2}{N_a}.$$

Proof of Lemma A.1. Part (i) follows directly from the variance of a sample average with random sampling from a finite population.
Next consider part (ii). Define

$$S_{aa'}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left\{\left(Y_i(a')-\overline{Y}(a')\right)-\left(Y_i(a)-\overline{Y}(a)\right)\right\}.$$

Recall that the variance of the difference between $\overline{Y}_{a'}$ and $\overline{Y}_a$ is

$$\mathbb{V}(\overline{Y}_{a'}-\overline{Y}_a)=\frac{S_a^2}{N_a}+\frac{S_{a'}^2}{N_{a'}}-\frac{S_{aa'}^2}{N},$$

from the results in Neyman (1923/1990) and Imbens and Rubin (2015) for completely randomized experiments with a binary treatment. In general it is also true that

$$\mathbb{V}(\overline{Y}_{a'}-\overline{Y}_a)=\mathbb{V}(\overline{Y}_a)+\mathbb{V}(\overline{Y}_{a'})-2\mathbb{C}(\overline{Y}_a,\overline{Y}_{a'}).$$

Combining these two characterizations of the variance of the standard estimator for the average treatment effect, it follows that the covariance is equal to

$$\begin{aligned}\mathbb{C}(\overline{Y}_a,\overline{Y}_{a'})&=\frac{1}{2}\left\{\mathbb{V}(\overline{Y}_a)+\mathbb{V}(\overline{Y}_{a'})-\mathbb{V}(\overline{Y}_{a'}-\overline{Y}_a)\right\}\\&=\frac{1}{2}\left\{\frac{S_a^2}{N_a}\left(1-\frac{N_a}{N}\right)+\frac{S_{a'}^2}{N_{a'}}\left(1-\frac{N_{a'}}{N}\right)-\left\{\frac{S_a^2}{N_a}+\frac{S_{a'}^2}{N_{a'}}-\frac{S_{aa'}^2}{N}\right\}\right\}\\&=-\frac{1}{2N}\left\{S_a^2+S_{a'}^2-S_{aa'}^2\right\}\\&=-\frac{1}{2N}\left\{S_a^2+S_{a'}^2+V_{aa'}^2-2S_a^2-2S_{a'}^2\right\}\\&=\frac{1}{2N}\left\{S_a^2+S_{a'}^2-V_{aa'}^2\right\}.\end{aligned}$$

Next, consider part (iii). Using the result in part (ii),

$$\begin{aligned}\mathbb{V}\left(\sum_{a\in\mathbb{A}}\overline{Y}_a\right)&=\sum_{a\in\mathbb{A}}\mathbb{V}(\overline{Y}_a)+\sum_{a,a':a\neq a'}\mathbb{C}(\overline{Y}_a,\overline{Y}_{a'})\\&=\sum_{a\in\mathbb{A}}\frac{S_a^2}{N_a}\left(1-\frac{N_a}{N}\right)+\frac{1}{2N}\sum_{a,a':a\neq a'}\left\{S_a^2+S_{a'}^2-V_{aa'}^2\right\}\\&=\sum_{a\in\mathbb{A}}S_a^2\left(\frac{1}{N_a}-\frac{1}{N}+\frac{T}{N}\right)-\frac{1}{2N}\sum_{a,a':a\neq a'}V_{aa'}^2\\&=\sum_{a\in\mathbb{A}}S_a^2\left(\frac{1}{N_a}+\frac{T-1}{N}\right)-\frac{1}{2N}\sum_{a,a':a\neq a'}V_{aa'}^2.\end{aligned}$$

Finally, consider part (iv). The third term, the sum of $V_{aa'}^2$ terms is not directly estimable. Because it has a negative sign, we need to find a lower bound on this sum. A trivial lower bound is zero, but we can do better. We will show that

$$\frac{1}{2N}\sum_{a,a':a\neq a'}V_{aa'}^2\geq\sum_{a\in\mathbb{A}}S_a^2\frac{T-1}{N}.\tag{A.1}$$

This in turn implies

$$-\frac{1}{2N}\sum_{a,a':a\neq a'}V_{aa'}^2\leq-\sum_{a\in\mathbb{A}}S_a^2\frac{T-1}{N},$$

and thus

$$\mathbb{V}\left(\sum_{a\in\mathbb{A}}\overline{Y}_a\right) = \sum_{a\in\mathbb{A}} S_a^2\left(\frac{1}{N_a}+\frac{T-1}{N}\right) - \frac{1}{2N}\sum_{a\neq a'} V_{aa'}^2$$

$$\leq \sum_{a\in\mathbb{A}} S_a^2\left(\frac{1}{N_a}+\frac{T-1}{N}\right) - \sum_{a\in\mathbb{A}} S_a^2\frac{T-1}{N}$$

$$= \sum_{a\in\mathbb{A}}\frac{S_a^2}{N_a}.$$

The last inequality to prove is (A.1). First,

$$V_{aa'}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\left(Y_i(a')-\overline{Y}(a')\right)+\left(Y_i(a)-\overline{Y}(a)\right)\right)^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left\{\left(Y_i(a')-\overline{Y}(a')\right)^2+\left(Y_i(a)-\overline{Y}(a)\right)^2+2\left(Y_i(a')-\overline{Y}(a')\right)\left(Y_i(a)-\overline{Y}(a)\right)^2\right\}$$

$$= \frac{1}{N}\left(S_a^2+S_{a'}^2+2\mathbb{C}(Y_i(a),Y_i(a'))\right).$$

Hence

$$\frac{1}{2N}\sum_{a\neq a'} V_{aa'}^2 = \frac{1}{2N}\sum_{a,a':a\neq a'}\left\{S_a^2+S_{a'}^2+2\mathbb{C}(Y_i(a),Y_i(a'))\right\}$$

$$= \sum_{a\in\mathbb{A}} S_a^2\frac{T}{N}+\frac{1}{N}\sum_{a\neq a'}\mathbb{C}(Y_i(a),Y_i(a')). \tag{A.2}$$

Next,

$$0\leq\mathbb{V}\left(\sum_{a\in\mathbb{A}}Y_i(a)\right) = \sum_{a\in\mathbb{A}}\mathbb{V}(Y_i(a))+\sum_{a,a':a\neq a'}\mathbb{C}(Y_i(a),Y_i(a')).$$

Therefore

$$\sum_{a,a':a\neq a'}\mathbb{C}(Y_i(a),Y_i(a'))\geq-\sum_{a\in\mathbb{A}}\mathbb{V}(Y_i(a)) = -\sum_{a\in\mathbb{A}}S_a^2. \tag{A.3}$$

Combining (A.2) and (A.3) we get the bound

$$\frac{1}{2N}\sum_{a,a':a\neq a'} V_{aa'}^2 = \sum_{a\in\mathbb{A}} S_a^2\frac{T}{N}+\frac{1}{N}\sum_{a,a':a\neq a'}\mathbb{C}(Y_i(a),Y_i(a'))$$

$$\geq \sum_{a\in\mathbb{A}} S_a^2\frac{T}{N}-\sum_{a\in\mathbb{A}} S_a^2 = \sum_{a\in\mathbb{A}} S_a^2\frac{T-1}{N},$$

which proves (A.1). □

Proof of Theorem 2. This follows directly from the results in Lemma A.1. □

Proof of Theorem 3. By Assumption 1 it follows that

$$\mathbb{E}\left[s_{\gamma a,a}^2\right] = S_{\gamma a,a}^2.$$

This implies that

$$\mathbb{E}\left[\hat{\mathbb{V}}_{\text{did}}\right] = \mathbb{E}\left[\sum_{a\in\mathbb{A}} s_{\gamma a,a}^2/N_a\right] = \sum_{a\in\mathbb{A}} S_{\gamma a,a}^2/N_a \geq \mathbb{V}(\hat{\tau}_{\text{did}}),$$

where the inequality is by Theorem 2. □

References

Abadie, Alberto, 2005. Semiparametric difference-in-differences estimators. Rev. Econom. Stud. 72 (1), 1–19.
Abadie, Alberto, Athey, Susan, Imbens, Guido W., Wooldridge, Jeffrey M., 2020. Sampling-based vs design-based uncertainty in regression analysis. Econometrica 265–296.
Abadie, Alberto, Athey, Susan, Imbens, Guido, Wooldrige, Jeffrey, 2016. When should you adjust standard errors for clustering?.
Abadie, Alberto, Diamond, Alexis, Hainmueller, Jens, 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J. Amer. Statist. Assoc. 105 (490), 493–505.
Abbring, Jaap H., Van den Berg, Gerard J., 2003. The nonparametric identification of treatment effects in duration models. Econometrica 71 (5), 1491–1517.
Abbring, Jaap H., Heckman, James J., 2007. Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Handbook of Econometrics, vol. 6, pp. 5145–5303.
Angrist, Joshua, Krueger, Alan, 2000. Empirical strategies in labor economics. In: Handbook of Labor Economics, vol. 3.
Angrist, Joshua, Pischke, Steve, 2008. Mostly Harmless Econometrics: An Empiricists' Companion. Princeton University Press.
Arellano, Manuel, 1987a. Practitioners corner: Computing robust standard errors for within-groups estimators. Oxf. Bull. Econ. Stat. 49 (4), 431–434.
Arellano, Manuel, 1987b. Computing robust standard errors for within group estimators. Oxf. Bull. Econ. Stat. 49 (4), 431–434.
Arellano, Manuel, 2003. Panel Data Econometrics. Oxford university press.
Arkhangelsky, Dmitry, Athey, Susan, Hirshberg, David A., Imbens, Guido W., Wager, Stefan, 2019. Synthetic Difference in Differences. Technical Report, National Bureau of Economic Research.
Arkhangelsky, Dmitry, Imbens, Guido, 2018. The Role of the Propensity Score in Fixed Effect Models. Technical Report, National Bureau of Economic Research.
Aronow, P., Green, D., Lee, D., 2014. Sharp bounds on the variance in randomized experiments. Ann. Statist. 42 (3), 850–871.
Aronow, Peter M., Samii, Cyrus, 2016. Does regression produce representative estimates of causal effects? Am. J. Political Sci. 60 (1), 250–267.
Athey, Susan, Imbens, Guido, 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74 (2), 431–497.
Athey, Susan, Stern, Scott, 1998. An Empirical Framework for Testing Theories About Complimentarity in Organizational Design. Technical Report, National Bureau of Economic Research.
Ben-Michael, Eli, Feller, Avi, Rothstein, Jesse, 2018. The augmented synthetic control method. arXiv preprint arXiv:1811.04170.
Bertrand, Marianne, Duflo, Esther, Mullainathan, Sendhil, 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119 (1), 249–275.
Borusyak, Kirill, Jaravel, Xavier, 2016. Revisiting Event Study Designs.
Callaway, Brantly, Sant'Anna, Pedro HC., 2020. Difference-in-differences with multiple time periods. J. Econometrics.
Card, David, 1990. The impact of the mariel boatlift on the miami labor market. Ind. Labor Relat. 43 (2), 245–257.
Card, David, Krueger, Alan, 1994. Minimum wages and employment: Case study of the fast-food industry in new Jersey and Pennsylvania. Am. Econ. Rev. 84 (4), 772–793.
de Chaisemartin, Clément, d'Haultfœuille, Xavier, 2017. Fuzzy differences-in-differences. Rev. Econom. Stud. 85 (2), 999–1028.
de Chaisemartin, Clément, d'Haultfœuille, Xavier, 2020. Two-way fixed effects estimators with heterogeneous treatment effects. Amer. Econ. Rev. 110 (9), 2964–2996.
Chamberlain, Gary, 1984. Panel data. In: Handbook of Econometrics, vol. 2, pp. 1247–1318.
Chamberlain, Gary, et al., 1993. Feedback in Panel Data Medels. Technical Report, Harvard-Institute of Economic Research.
Chay, Kenneth Y., Greenstone, Michael, 2005. Does air quality matter? Evidence from the housing market. J. Political Econ. 113 (2), 376–424.
Conley, Timothy G., Taber, Christopher R., 2011. Inference with difference in differences with a small number of policy changes. Rev. Econ. Stat. 93 (1), 113–125.
Donald, Stephen G., Lang, Kevin, 2007. Inference with difference-in-differences and other panel data. Rev. Econ. Stat. 89 (2), 221–233.
Doudchenko, Nikolay, Gilinson, David, Taylor, Sean, Wernerfelt, Nils, 2019. Designing Experiments with Synthetic Controls. Technical Report.
Engle, Robert F., Hendry, David F., Richard, Jean-Francois, 1983. Exogeneity. Econometrica 277–304.
Freyaldenhoven, Simon, Hansen, Christian, Shapiro, Jesse M., 2019. Pre-event trends in the panel event-study design. Am. Econ. Rev. 109 (9), 3307–3338.
Goodman-Bacon, Andrew, 2018. Difference-in-Differences with Variation in Treatment Timing. Technical Report, National Bureau of Economic Research.
Han, Sukjin, 2020. Identification in nonparametric models for dynamic treatment effects. J. Econometrics.
Hazlett, Chad, Xu, Yiqing, 2018. Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data.
Hernan, Miguel A., Robins, James M., 2020. Causal Inference. CRC Boca Raton, FL.
Hull, Peter, 2018. Estimating treatment effects in mover designs. arXiv preprint arXiv:1804.06721.
Imai, Kosuke, Kim, In Song, 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? Am. J. Political Sci. 63 (2), 467–490.
Imbens, Guido, 2000. The role of the propensity score in estimating dose–response functions. Biometrika 87, 706–710.
Imbens, Guido W., Rubin, Donald B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
Kim, In Song, Imai, Kosuke, Wang, Erik, 2019. Matching Methods for Causal Inference with Time-Series Cross-Sectional Data.
Liang, Kung-Yee, Zeger, Scott L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.
Meyer, Bruce D., Viscusi, W.Kip, Durbin, David L., 1995. Workers' compensation and injury duration: Evidence from a natural experiment. Am. Econ. Rev. 322–340.
Neyman, Jerzey, 1923/1990. On the application of probability theory to agricultural experiments. Essay on principles. section 9. Statist. Sci. 5 (4), 465–472.
Pearl, Judea, 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA, ISBN: 0-521-77362-8.
Rosenbaum, Paul R., 2002. Observational Studies. Springer.
Rosenbaum, Paul R., 2017. Observation and Experiment: An Introduction to Causal Inference. Harvard University Press.
Rubin, Donald B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 (5), 688.
Rubin, Donald B., 1978. Bayesian inference for causal effects: The role of randomization. Ann. Statist. 34–58.
Shah, Bbabubhai V., Holt, Mary Margaret, Folsom, Ralph E., 1977. Inference about regression models from sample survey data. Bull. Int. Stat. Inst. 47 (3), 43–57.
Stock, James H., Watson, Mark W., 2008. Heteroskedasticity-robust standard errors for fixed effects panel data regression. Econometrica 76 (1), 155–174.
Strezhnev, Anton, 2018. Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs.
Sun, Liyang, Abraham, Sarah, 2020. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econometrics.
Wooldridge, J.M., 0000. Econometric Analysis of Cross Section and Panel Data. The MIT Press. ISBN 9780262232586.
Xiong, Ruoxuan, Athey, Susan, Bayati, Mohsen, Imbens, Guido W., 2019. Optimal experimental design for staggered rollouts. Available at SSRN.