Semiparametric weighting estimators for multi-period difference-in-differences designs

Anton Strezhnev[*]

August 24, 2018

Abstract

Difference-in-differences designs are a powerful tool for causal inference in observational settings where typical selection-on-observables assumptions fail to hold. When a pre-treatment period is observed for all units, the treatment effect on the treated in the second period is identified non-parametrically under a weaker "parallel trends" assumption. However, researchers lack a reliable means of generalizing this approach to designs with multiple pre- and post-treatment periods, particularly when the parallel trends assumption only holds conditional on a set of covariates. While the two-period difference-in-differences estimator is equivalent to a fully-saturated linear regression model with unit and time dummy parameters, two-way fixed effects regression estimators do not recover the average treatment effect when there are more than two treatment periods even when parallel trends holds unless the true outcome model is correctly specified. This paper clarifies the causal estimands in a multi-period difference-in-differences design and develops an estimation strategy that extends Abadie's (2005) semiparametric inverse propensity weighting method that allows researchers to incorporate covariates without necessarily making strong assumptions about the data generating process for the outcome. It evaluates this new method on the effect of United States' investment treaties on foreign direct investment.

[*] Paper prepared for the Annual Conference of the American Political Science Association August 30-September 2, 2018, Boston MA Author contact information: University of Pennsylvania Law School, Philadelphia PA Email: astrezhn@law.upenn.edu

1 Introduction

Difference-in-differences (DID) estimators are an important tool for applied researchers studying causal effects with observational data. In settings where "as-if-random" treatment assignment assumptions are unreasonable due to unobserved common causes of treatment and outcome, DID is one of the most straightforward methods for adjusting for certain forms of selection bias. When researchers have access to additional observations of the outcome from periods where all units in the sample are untreated, any observed differences between these two groups can be attributed to underlying differences in the latent characteristics of the types of units receiving treatment and control. Assuming that this selection bias is the same in both periods, subtracting this auxiliary difference from the simple difference-in-means can correct the bias due to non-random assignment – hence the name difference-in-differences. Equivalently, DID designs can also be motivated by the idea of "de-biasing" within-unit pre- and post-treatment comparisons when there are trends in the outcome over time that could account for changes between the two periods. By assuming that units that are treated would have the same underlying time trend in the absence of treatment as those units that never receive treatment, the difference-in-differences estimator subtracts the observed time-trend in untreated units from the naive pre-/post-treatment comparison.

The theory motivating the DID estimator has been primarily developed in the context of panels with only two time periods with a binary treatment assigned only in the second period. In this case, no additional modeling assumptions need to be made to estimate the effect as DID simply requires the estimation of four conditional expectations, which can be done without parametric assumptions. However, in most actual applications of the difference-in-differences framework, researchers working with panel data will observe outcomes for more than two time periods. Units in the data typically also do not all initiate treatment at the same time and some may discontinue treatment after the initiation period. Building on the well-known result that the ordinary least squares estimator with unit and time fixed effects (FE) and an indicator variable for treatment is equivalent to the non-parametric difference-in-differences estimator in the special two-period case, popular research methods texts usually suggest researchers use this same two-way fixed effects estimator in the case of multiple time periods. Angrist and Pischke (2009) term this approach "Regression DD" and Bertrand, Duflo and Mullainathan (2004) note that this is the standard estimation strategy for most econometrics research that utilizes DID in panels with more than two time periods.

However, recent work has challenged the idea that the regression DD approach produces valid estimates of causal effects in the same way that the non-parametric DID estimator does in the two-period case. Imai, Kim and Wang (2018) note that the two-way fixed effects estimator itself does not correspond to any valid matching estimator and can often impute improper counterfactuals for treated units. Abraham and Sun (2018) and de Chaisemartin and D'Haultfoeuille (2018) find that the two-way fixed effects estimator places non-uniform weights on the treated units in the sample, resulting in effect estimates that may mischaracterize the sample when effects are heterogeneous across units even when researchers attempt to model effects or pre-treatment trends over time via leads or lags of the treatment variable. Borusyak and Jaravel (2017) argues that when effects are variable over time, two-way fixed effects estimators will up-weight short-run effects and down-weight long-term effects.

In this paper, I provide a simple, unifying, explanation for why the two-way FE estimator fails to estimate unbiased treatment effects by showing that the two-way FE estimator can be written as a uniform average of all possible two-period DID estimators in the sample. Some of these estimators are unbiased DID comparisons and require no additional assumptions beyond the standard "parallel trends" assumption. However, others are unbiased *only* if an additional assumption is made such that treatment effects do not persist beyond a single period. This is because two-way FE considers time periods where two units are both exposed to treatment as a valid de-biasing second difference. However, even if two units are both under the same treatment status in a particular period, their overall exposures to treatment may differ. One unit will have been under treatment for a longer period of time than the other. Therefore, the difference in observed outcomes between those two periods cannot be attributed exclusively to "bias" as is the case for periods where both units have never been exposed to treatment. In addition to this bias resulting from the use of invalid second-differences, I show that the two-way FE estimator exhibits the same weighting problems as any multiple regression estimator used to estimate causal effects (Aronow and Samii, 2016). This is a variation of the problem found by Imai and Kim (2017) for one-way unit fixed effects estimators – units for which treatment status is well predicted by the fixed effects receive less weight when averaging over the distribution of treatment effects. The result is that for a single treated unit, treatment effects for different time periods will receive non-uniform weights. Across units, treatment effects may be weighted in a way that is unrepresentative of a typical intervention, creating problems for interpretation and generalization for the sample.

I then lay out a framework for estimating treatment effects using DID beyond the two-period setting. In doing so, I build on a recent literature that attempts to generalize the DID approach to the setting with multiple time periods, persistent treatment effects and variable treatment uptake times. The "synthetic control" method of Abadie, Diamond and Hainmueller (2010, 2015) considers estimation of the entire effect trajectory for a single unit by re-weighting the pool of units never exposed to obtain a counterfactual time trend had that unit never been exposed to treatment. The weights are constructed such that the re-weighted pool of control units matches the treated unit with respect to observed covariates and pre-treatment outcomes. Xu (2017) generalizes this approach to the case of multiple treated units by directly modelling the outcome among units not receiving treatment via an interactive fixed effects model and subsequently using this model to impute counterfactuals for treated units. While this method works well to address potential unobserved confounding, it requires strong functional form assumptions for the outcome model, which are difficult to validate. Moreover, increased model flexibility comes at the cost of higher variance and more taxing requirements on the data. I develop an alternative estimation approach that generalizes the DID framework without requiring any models for the outcome. It deviates from the "synthetic controls" approach by retaining the "parallel trends" assumption underlying the standard DID rather than conditioning on pre-treatment outcomes.[1] I first define a new quantity of interest, the Average Cohort Treatment Effect on the Treated (ACTT), that corresponds to a well-defined intervention with respect to the distribution of units initiating treatment in a given sample. I then show how this quantity can be non-parametrically identified and estimated as a weighted average of two-period difference-in-differences estimates using a generalization of the parallel trends assumption. I then show how researchers can further relax the parallel trends assumptions by incorporating covariates through a generalization of the inverse propensity score weighting method for DID developed by Abadie (2005). This new weighting method is closest in spirit to recent work by Imai, Kim and Wang (2018) and Hazlett and Xu (2018) which propose either matching or re-weighting control observations such that the covariate and pre-treatment

[1] However, this is not an essential assumption - in principle, with sufficient regularization, it should be possible to adjust for a much higher-dimensional set of covariates. I discuss this more in the conclusion with regards to possible extensions of the method

outcome histories between treated and control conditions are balanced. Crucially, both of these approaches also avoid a major pitfall associated with including covariates directly into a two-way fixed effects regression: bias due to conditioning on a post-treatment variable (Rosenbaum, 1984).

I apply this new method to answer an ongoing puzzle in the study of investment law: whether investment treaties signed by states are actually effective in promoting foreign direct investment. Over the last several decades, many governments have entered into bilateral legal agreements with other states that offer a mutual grant of formal legal protections to foreign investors. Salacuse and Sullivan (2005) describes the bilateral investment treaty (BIT) regime as a "grand bargain" between capital importers and exporters. By tying the hands of capital importing countries and raising costs of expropriation or creating a hostile climate for FDI, BITs make credible a country's commitment to protecting investors' rights. This restriction of policy autonomy is the cost a capital importing country pays in exchange for the promised benefit of making it a more attractive investment target relative to its competitors. Whether this second half of the bargain is truly upheld is a challenging empirical question. Existing work on this question with time-series cross-sectional data has shown mixed results. One reason for this may be due to the use of two-way fixed effects estimators to account for unit-fixed and temporal sources of confounding. While the DID framework is a powerful tool for addressing omitted variable bias in panel data analyses, its implementation via two-way FE may in fact be generating misleading results. Because the effects of treaty adoption are likely to manifest over a period of many years rather than instantaneously, estimates from two-way FE will be biased towards zero. Furthermore, because researchers will typically also include time-varying covariates into these models, inferences are also likely to suffer from post-treatment bias if effects on those variables also exist and persist over time.

I analyze a new dataset on the presence of United States multinational firms abroad based on statistics provided by the Bureau of Economic Analysis (BEA) from 1983 to 2013. Specifically, I look at the effect of investment treaties on the *extensive* margin of FDI: whether new firms are willing to enter a market after the entry-into-force of a bilateral investment treaty with the United States. I find that two-way fixed effects estimates substantially understate the average effect of BIT entry into force. While the fixed effect models suggest effects statistically indistinguishable from 0, the new generalized difference-in-differences approach shows that investment treaties boosted the number of U.S. firms in a market by about .2 log-points (a roughly 20% increase). Incorporating covariates into the two-way fixed effects model related to economic development, democracy, and other investment treaty commitments drives the point estimate to 0, suggesting post-treatment bias. Conversely, using the proposed weighting method to adjust for covariates does not change the point estimate substantially from the simple DID estimate. These results suggest that investment treaties do actually have an influence on firms' investment decisions over the long-run and that two-way fixed effect estimators will tend to understate these effects due to their inherent biases.

The remainder of this paper is structured as follows: Section 2 develops the theory behind causal effects in a panel data setting and the assumptions behind the difference-in-differences estimator. After summarizing the classical DID framework, I generalize these assumptions to the case with many time periods and define a new causal estimand, the Average Cohort Treatment effect on the Treated (ACTT). I propose a straightforward non-parametric approach to estimation under a generalization of the parallel-trends assumption in many time periods. Section 3 then illustrates how the two-way fixed effects estimator is biased for the ACTT. It decomposes the source of the bias into two components: a bias due to the use of improper second-difference terms, and a bias due to the way in which OLS weights treatment effects, an analogue of the well-known "regression weighting" problem (Aronow and Samii, 2016). Section 4 explains how researchers can incorporate covariates into the proposed generalized DID estimator using an inverse propensity weighting approach that builds on the method described by Abadie (2005). Crucially, this method avoids problems with post-treatment bias that arise when including covariates that might be affected by treatment directly into a time-series regression model. Section 5 compares the difference between the two-way FE approach and the new estimator in an analysis of the effect of U.S. investment treaties on the activity of U.S. multinationals. Section 6 concludes with recommendations for applied researchers working in a panel data setting.

2 Setup and Theoretical Framework

To understand how a difference-in-differences design could work in a multi-period setting, it is important to clearly explain the quantity of interest being estimated. This section develops a theoretical framework for defining and understanding causal effects in a time-series context. Consider a sample of N units each indexed by i. Each unit is observed over a total of T time periods indexed by t. The research goal is to estimate the effect of some exposure or "treatment" on an outcome over a series of time periods. Each unit i is assigned to some treatment history denoted $\vec{A_i}$, which is a T-length vector of the unit's particular treatment status in each time period: $\vec{A_i} = \{A_{i1}, A_{i2}, \ldots, A_{iT}\}$. Denote the set of all possible treatment vectors as $\mathcal{A}$. For the purposes of this paper, I focus exclusively on the case where treatment in any given period is a binary indicator with $A_{it} = 1$ indicating that a unit is exposed to a particular treatment at time t and $A_{it} = 0$ indicating that unit is not exposed. Furthermore, denote a unit's partial treatment history as the sub-vector of treatments up to some time period t: $\vec{A_{it}} = \{A_{i1}, A_{i2}, \ldots, A_{it}\}$. A unit that has not yet initiated treatment at time t will have a partial treatment history of all zeroes $\vec{A_{it}} = \{0, 0, \ldots, 0\} = \vec{0}$ The observed outcome in time t for unit i is denoted Y_{it}. Likewise, the vector of all outcomes for unit i is $\vec{Y_i}$ and the sub-vector up to time t is $\vec{Y_{it}}$. Finally, let $\vec{X_i}$ be the K by T matrix of observed covariates for unit i with X_{it} the K-length vector of covariates observed at time t and $\vec{X_{it}}$ the K by t matrix of covariates up to time t. Assume that X_{it} is observed prior to the assignment of treatment in period t but can be affected by treatment assigned in prior periods. For the purposes of illustration, I will focus here on identification without conditioning on covariates. However, Section 4 will discuss methods for covariate adjustment when the necessary identification assumptions only hold given some set of observed pre-treatment covariates.

I define causal effects using the conventional *potential outcomes* framework, also often referred to as the "Rubin Causal Model" (Neyman, 1923; Rubin, 1974) A causal effect is the change in the outcome that would be observed if a unit had been assigned to one treatment regime versus another. Since we only ever observe units under a single treatment regime, identifying a causal effect from data requires reasoning about counterfactuals and that researchers make asumptions about the treatment assignment process. This is often referred to as the "fundamental problem of causal inference" (Holland, 1986). Formally, let $Y_{it}(\vec{a})$ denote the potential outcome that we would observe for unit i in time period t if that unit were assigned to the particular treatment history $\vec{A_i} = \vec{a}$.[2] Next, I make the assumption of "consistency" to connect the observed data to

[2] Note that by writing the potential outcome only in terms of the treatment vector for unit i, I am implicitly making what is often known as the Stable Unit Treatment Value assumption or SUTVA with respect to the units in the sample (Rubin, 1986). This assumption states that a unit's potential outcomes only depend on the assignment of their particular treatment history and not on the treatment histories of other units. Often, this assumption is stated separately, depending on the particular theoretical treatment, but is also implied by the consistency assumption.

counterfactuals.

Assumption 1 *Consistency*

$$Y_{it} = Y_{it}(\vec{a}) \text{ } if \text{ } \vec{A}_i = \vec{a} \tag{1}$$

Consistency states that the observed outcome for units with an observed treatment history equal to $\vec{a}$ is equal to that unit's potential outcome had it been assigned to treatment history $\vec{a}$.[3]

Because individual causal effects cannot be estimated due to the fundamental problem of causal inference, researchers typically focus on averages of effects. I define the "average treatment history effect" in some time period t as the difference in the expected potential outcome under assignment to two different treatment histories

Definition 1 *Average Treatment History Effect*

$$ATE_t(\vec{a}, \vec{a}^*) = E[Y_{it}(\vec{a}) - Y_{it}(\vec{a}^*)] \tag{2}$$

Treatment effects can also be defined for sub-groups within the population. In the DID context, researchers focus on estimating the average treatment effect on the treated (ATT) since under the necessary identification assumptions, counterfactuals can only be imputed for treated units and not for the controls.

Definition 2 *Average Treatment History Effect on the Treated*

$$ATT_t(\vec{a}, \vec{a}^*) = E[Y_{it}(\vec{a}) - Y_{it}(\vec{a}^*)|A_i = \vec{a}] \tag{3}$$

With a few sensible assumptions, it is possible to know, with certainty, that some treatment history effects are zero. Intuitively, if two treatment histories differ only in the treatment levels

[3] In randomized experiments and studies where treatment is directly manipulated and assigned by a researcher, this is a straightforward and trivial assumption. However, it is worth noting that tricky theoretical questions can arise in observational studies where the idea of "treatment" is used more loosely to refer to some event that arises as a result of nature. Simply put, the notion of causality in the potential outcomes framework requires some notion of manipulability – that a unit could have been assigned to a different condition than what we observe. Consistency implies then that the outcome that we happen to see in the data is equivalent to the outcome we would observe under that hypothetical manipulation if each unit were "assigned" to their observed treatment status (VanderWeele and Vansteelandt, 2009).

assigned in periods *after* some time period t, then it is impossible for there to be a causal effect in period t. For a manipulation in the future to affect an outcome in the past would violate known properties of time, which physicists understand as an asymmetric process which flows in a single direction. Cause temporally precedes effect, a property often described as "time's arrow."[4] Counterfactual reasoning also retains this time-asymmetric property (Lewis, 1979). An intervention made in the past has the potential to affect the present. But an intervention made in the present cannot change the past. I formalize this notion in a "no reverse causality" assumption, which I will show forms the basis of identification in the difference-in-differences setting.

Assumption 2 *No reverse causality*

$$Y_{it}(\vec{a}) = Y_{it}(\vec{a}^*) \text{ } \textit{if } \vec{a}_t = \vec{a}^*_t \tag{4}$$

In other words, if two treatment histories are identical up to time t, the potential outcomes in time t associated with those histories will also be the same, even if the histories differ in periods after t. Under this assumption, it is possible to find comparisons between groups assigned to two different treatment histories where we know the causal effect must be 0. Therefore, any difference in observed outcomes can be attributed not to the effect of treatment, but rather to underlying differences between the types of units assigned to one history versus the other – the omitted variable bias.

2.1 Difference-in-differences with two time periods

With the no reverse causality assumption, it is possible to use observations from one period to "de-bias" the naive difference-in-means estimates of treatment effects in other periods where treatment varies. This facilitates identification in the DID setting with repeated observations of units. In this section I outline the assumptions behind the classic difference-in-differences estimator in the simplest setting with only two time periods ($T = 2$) and two possible treatment histories. Assume that at time 1, all units are under control. In time period 2, units can either initiate treatment or

[4] While the question of reconciling the perceived asymmetry of time with theories of the physical universe remains a serious puzzle in the field of theoretical physics, discussion of these complexities is far beyond the scope of this paper. See Halliwell, Pérez-Mercader and Zurek (1996) for a review.

remain under control. Let a^1 denote the treatment history for units that are under control only in period 1 and initiate treatment in period 2, and a^2 denote always-control treatment history.

In this setting, there is only one non-zero treatment effect of interest, the ATT in period 2.

$$\text{ATT}_2(a^1, a^2) = E[Y_{i2}(a^1)|\vec{A}_i = a^1] - E[Y_{i2}(a^2)|\vec{A}_i = a^1] \tag{5}$$

The first term can be identified directly from the data under the consistency assumption, as it is simply the expected outcome in period 2 for units assigned to treatment. However, the second is a counterfactual quantity that must be imputed from those observations assigned to the control history. If treatment assignment were completely randomized, there would be no differences in expectation between units under treatment and control except for the manipulated treatment condition. Therefore, the potential outcome under control for treated units would be equal to the average observed outcome for units receiving the control. In observational studies, researchers will typically invoke a conditional version of this assumption to obtain identification – that treatment is as good as randomized given some set of covariates. Under such a "selection on observables" assumption, the treatment effect can be estimated via typical covariate adjustment methods such as regression, sub-classification, matching or inverse propensity weighting (Imbens, 2004).

When there remain unobserved confounders of treatment and outcome, adjusting on observed covariates alone will still yield biased estimates of the treatment effect. Difference-in-differences designs relax the "selection on observables" assumption by allowing for the existence of unobserved, unit-fixed confounders of treatment assignment and outcome. Formally, the DID approach makes the assumption of "parallel trends":

Assumption 3 *Parallel trends*

$$E[Y_{i2}(a^2) - Y_{i1}(a^2)|\vec{A}_i = a^1] = E[Y_{i2}(a^2) - Y_{i1}(a^2)|\vec{A}_i = a^2] \tag{6}$$

This assumption states that in the absence of treatment, units assigned to the treated history would have the same linear trend in the outcome compared to units assigned to the control history. Under no reverse causality, we know that $Y_{i1}(a^2) = Y_{i1}(a^1)$ since both treatment histories are identical up to period 1. Therefore, under parallel trends, the ATT is identified non-parametrically by:

$$\text{ATT}_2(a^1, a^2) = \big(E[Y_{i2}|A_i = a^1] - E[Y_{i2}|A_i = a^2]\big) - \big(E[Y_{i1}|A_i = a^1] - E[Y_{i1}|A_i = a^2]\big) \quad (7)$$

This estimator consists of a difference in two differences, hence the name difference-in-differences. For this paper, I will refer to the "first difference" as the naive difference-in-means estimator in period 2 and the "second difference" as the bias correction estimated from the observed difference in outcomes between the treatment histories for period 1.[5] If treatment is in fact randomly assigned, the second difference should be zero in expectation (as treated and control are in expectation the same on all pre-treatment covariates) and the expression will reduce to the typical difference-in-means estimator.

2.2 Difference-in-differences with multiple time periods

While the two-period difference-in-differences case is well-studied, extending the intuition from that case to multiple time periods is complicated by the absence of a comparable causal estimand or quantity of interest. Adding additional time periods expands the number of possible treatment histories for which an ATT can be defined. With T time periods and no restrictions on possible treatment histories, there are 2^T possible unique treatment histories. As researchers add more and more time periods, purely non-parametric estimation with no additional restrictions becomes increasingly infeasible due to the "curse of dimensionality" as some treatment histories may only be observed for a handful of units while others may be never observed at all. Moreover, researchers are rarely interested in the effect of one particular history, but rather some sort of *average* of treatment history effects for the entire sample.

In this section, I define a general quantity of interest for difference-in-differences designs under certain limitations on the possible treatment histories. Specifically, I constrain units from reverting their treatment status once they initiate treatment. In many studies, it is reasonable to assume that treatment uptake can only go in one direction. Once a unit receives treatment, it is always

[5] An equivalent way of writing the DID estimator is to re-arrange expectations and first take the difference in expected outcomes for periods 2 and 1 for units assigned treatment, subtracting that from the difference in expected outcomes for the same time periods for those assigned to the control history. The version used here is preferrable for the purposes of this paper as it makes clear the connection to the simple difference-in-means and clarifies the role of the second difference as a bias-correction.

under treatment until the final time period T. In studies of policy adoption, this is typically the case when the time period under consideration is relatively short. Imai, Kim and Wang (2018) refer to this as the assumption of "stable policy change" in a given intervention. This restricts the number of possible treatment histories to T if it is also assumed that each unit is under control for at least one time period.[6] However, units may not necessarily receive treatment all in the same time period. For example, governments may adopt similar policies at slightly different times with some units being leaders and others lagging behind.[7] Following the terminology of (Abraham and Sun, 2018), I will refer to the groups of units initiating treatment at the same time as treatment "cohorts." Each cohort corresponds to some value C_i, which denotes the last period under which that unit i is under control. A unit's treatment history is determined entirely by C_i. For a unit with $C_i = c$, $A_{it} = 0$ for all $t \leq c$ and $A_{it} = 1$ for all $t > c$. Units with $C_i = T$ never receive treatment and are always under the control condition.

To simplify the notation of treatment histories, let a^c denote the treatment history associated with cohort $C_i = c$, $c \in \{1, \dots, T\}$. $Y_{it}(a^c)$ is the potential outcome observed for unit i in time t if it initiated treatment at time $c + 1$. Define the Cohort Average Treatment effect on the Treated (CATT) in time t as

$$CATT_t(c) = E[Y_{it}(a^c) - Y_{it}(a^T)|\vec{A}_i = a^c] \tag{8}$$

This corresponds to a natural effect of interest: the change in the expected outcome at time t if a unit that initiated treatment at time $c + 1$ were instead never exposed. While researchers could plausibly be interested in other counterfactual comparisons with histories other than the "never treated" history, it is the most intuitive starting point for defining causal contrasts.

Under the no reverse causality assumption, $CATT_t(c) = 0$ for all $c \geq t$. Additionally,

$$Y_{it}(a^c) - Y_{it}(a^T) = Y_{it}(a^c) - Y_{it}(a^j) \text{ for all } t \leq j \leq T \tag{9}$$

In other words, the CATT for a given time period depends only on a unit's treatment history up

[6] Note that it is not possible to non-parametrically estimate the ATT for units that are always under treatment as there are no control periods that can serve as part of the de-biasing term.

[7] In econometrics, these variable treatment uptake scenarios are often termed "event studies" as the units being studied experience some "event" at potentially different times and each unit may have a distinct set of pre-event and post-event observations (Abraham and Sun, 2018).

to time t. This allows treated units that initiate treatment at time periods after t to act as control units for those that intiate treatment prior to t.

Identification in the multiple-period setting relies on a generalization of the parallel trends assumption.

Assumption 4 *Generalized parallel trends*

$$E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i = c] = E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i \geq t] \tag{10}$$

for all $t > c$, $t' \leq c$

This assumption states that, in the absence of treatment uptake at time c, the expected change in outcome between time t and some past time $t' \leq c < t$ would have been the same for units that initiate treatment in period $c + 1$ and units that remain under control up until period t. With this assumption, no-reverse causality, and the assumption the CATT in time t for cohort $C_i = c$ is non-parametrically identified by

$$CATT_t(c) = \frac{1}{c}\sum_{t'=1}^{c} [E[Y_{it}|C_i = c] - E[Y_{it'}|C_i = c]] - [E[Y_{it}|C_i \geq t] - E[Y_{it'}|C_i \geq t]] \tag{11}$$

This generalized multi-period difference-in-difference estimator is essentially an average of c "two-period" difference-in-differences estimators with the "treatment" period always equal to t and the "control" period changing between all time periods prior to the cohort's initiation of treatment. Note that this definition makes no additional restrictions on which time periods can be affected by treatment so long as t is greater than c. Consistent estimates of each conditional expectation can be obtained by directly substituting in the sample analogues.

Researchers studying the effects of a particular event or policy will typically not just be interested in the effect for a specific cohort and for a single time period. Treatment may have different effects in earlier periods relative to later ones and a natural way of aggregating this combination of short and long-term effects for a given cohort is to average the individual $CATT_t(c)$ effects over all periods for which the cohort is exposed to treatment. Define the overall Cohort Average

Treatment Effect on the Treated as

$$CATT(c) = \frac{1}{T - c - 1} \sum_{t=c+1}^{T} CATT_t(c) \tag{12}$$

This aggregation also corresponds to a natural causal question of interest: what would have happened, on average, had a cohort never initiated treatment in those time periods where the treatment could have had an impact. Note, however, that the CATT for one cohort will cover a different set of time periods than a CATT for another timep eriod. How then, should multiple CATTs be aggregated into a summary quantity. One approach is to only use a set number of post-treatment periods (denoted F) for each cohort, as in Imai, Kim and Wang (2018). However, this approach requires that researchers make an additional up-front assumption about how many post-treatment periods are of interest. It also throws away data: units with more treated periods than F go partially unused, while units with fewer than F treated periods cannot be part of the analysis at all.

Here I define a causal estimand that does not require the user to specify a particular number of post-treatment periods, the Average Cohort Treatment Effect on the Treated (ACTT)

Definition 3 *Average Cohort Treatment Effect on the Treated (ACTT)*

$$ACTT = \sum_{c=1}^{T-1} CATT(c) Pr(C_i = c | C_i \neq T) \tag{13}$$

The ACTT corresponds to a weighted average of cohort treatment effects with the weights proportional to the in-sample frequencies of each cohort (among cohorts receiving treatment). It has a natural interpretation as the CATT for a unit chosen randomly from the sample. To estimate the ACTT non-parametrically, it suffices to estimate each $CATT(C_i)$ for each unit in the data and take the average across the sample. Moreover, inference is straightforward as it is possible to interpret this as a weighted average of all feasible two-period difference-in-differences estimators in the sample. Since this estimator is linear in Y, valid standard errors can easily be calculated using bootstrapping methods, namely the block bootstrap which resamples units with replacement in order to preserve the outcome correlation structure within each unit (Bertrand, Duflo and Mullainathan, 2004).

3 Bias from OLS with unit/time fixed effects

Standard practice among researchers estimating DID effects in panels with many time periods is to fit an ordinary least squares regression with fixed effect parameters for both unit and time. In the two-period case with only two treatment histories, it is well known that the ordinary least squares regression estimator with time and unit fixed effects is identical to the non-parametric difference-in-differences estimator. Texts on causal inference typically will recommend using this same two-way fixed effects model in order to estimate difference-in-differences effects more generally, an approach Angrist and Pischke (2009) term "Regression DD" (pp. 223). The underlying regression model assumes the following data-generating process for Y_{it}

$$Y_{it} = \alpha_i + \gamma_t + \beta A_{it} + \varepsilon_{it} \tag{14}$$

where α_i is a fixed effect parameter for each unit, γ_t is the fixed effect parameter for each time period, and ε_{it} is a mean-zero error term. As before A_{it} is an indicator that takes on a value of 1 if unit i is under treatment at period t and 0 if it is not. Researchers will typically report the "average treatment effect" as the estimated coefficient $\hat{\beta}$ on A_{it}.

Unfortunately, interpreting this coefficient as a meaningful treatment effect requires very strong assumptions on the way in which treatment can affect the outcome. And even when these assumptions are satisfied, the corresponding regression coefficient may not be representative of treatment effects in the sample or reflect an average over units in the sample that is substantively interesting to a researcher. I show here that the two-way fixed effects estimator can be written as a uniform average of all possible difference-in-differences comparisons in the sample.

Proposition 1 *The OLS estimate of β in the two-way fixed effects model is equivalent to*

$$\hat{\beta} = \frac{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{[Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}]\}}{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{1 - A_{it'} + A_{jt'}\}} \tag{15}$$

where $\sum_{i:A_{it}=1}$ denotes a sum over all units (indexed i) where treatment status $A_{it} = 1$.

The complete proof is given in the appendix. In this section I will discuss the main intuitions behind it. Within every time period t, the two-way FE estimator startbs by matching each treated unit to each unit under control. This corresponds to the "first-difference" in the difference-in-differences estimator. For each matched pair, it then iterates through all other time periods (denoted t') and subtracts from the first-difference the "second-difference," comprised of the outcomes in time t' for the same pair of units.

It is in this "second difference" that the main source of bias is induced. There are three possible matches that can be found in the data for each treatment/control pair. First, a pair that is under treatment/control at time t can be matched to a time period t' where both units are under control. If treatment histories are restricted such that no unit can revert from treated to control, it must be the case that $t > t'$ and that units' treatment histories are identical up to t' – they are both always under control. Therefore, this is a valid second difference with respect to the DID estimator as there is no treatment effect of one history relative to the other at time t'. Second, a treatment/control pair can be matched to a time period where the unit under treatment remains under treatment and the unit under control remains under control. In this case, the difference-in-differences will cancel out as each treatment/control pair will appear once in a first difference and once in a second difference. Finally, treatment/control pairs can be matched to a time period where both units are under treatment. Under the assumption that no unit reverts from treated to control, such periods are necessarily in the future ($t' > t$). Additionally, the two units will not have the same treatment history up to t' since we know they differ at t. Therefore, unlike the control/control case, these observations act as invalid second-differences because the treatment effect of one history versus the other is not guaranteed to be 0. Borusyak and Jaravel (2017) refer to this as a "forbidden extrapolation." For these pairs of observations to act as valid second differences, it must be the case that the effect of treatment does not persist beyond a single period. In this case, the potential outcome for a unit depends only on its treatment assignment in period t and not on past assignments. With this additional assumption, there now exists no treatment effect for two units with the same treatment level at time t and periods where units are both under treatment can in fact serve as valid second differences for treatment/control comparisons in previous time periods. Unfortunately, such assumptions are highly restrictive and implausible in most settings.[8]

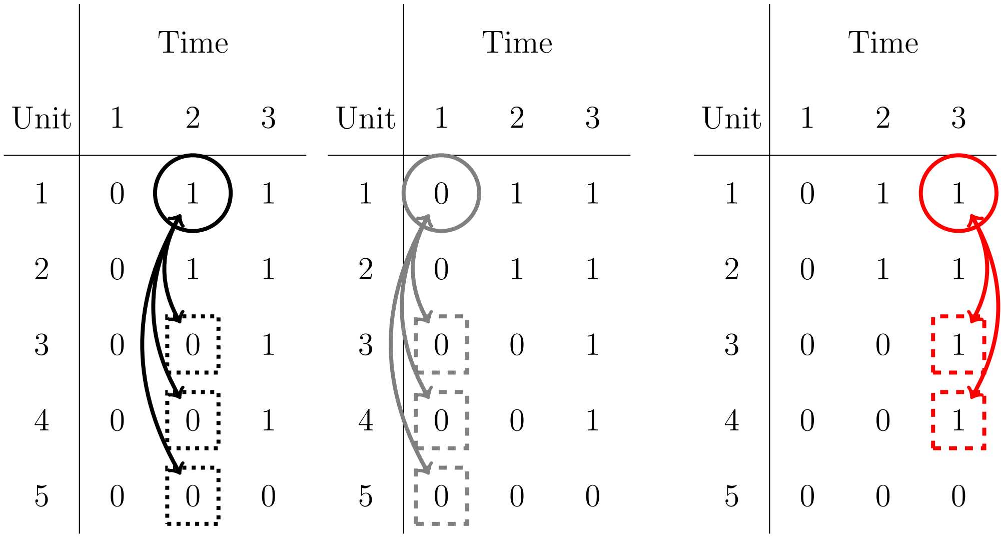
[8] When there are no restrictions on treatment histories, a fourth type of match can occur: a treated-control pair is matched to a control-treated pair. In this case, the difference-in-differences counts *twice* for the average as the model assumptions imply that the difference between these two differences is equal to 2β.

Second, it is clear from the expression in Proposition 1 that, while the average over DID comparisons is uniform, treated units receive non-uniform weights. This is because the number of units matched to a treated unit in the first difference and as part of the second-difference varies from time period to time period. Cohorts for which there are many pre-treatment periods receive greater weight than cohorts with few pre-treatment periods. Within an individual cohort, future time periods receive less weight than past ones as the number of within-period controls decreases over time. This persists even if one were to eliminate the invalid second differences, suggesting that the problem of up-weighting short-term versus long-term effects is not just due to problem of invalid second differences. As a result, this weighting of units does not correspond to the ACTT as defined in the previous section since cohorts receive in-sample weights that are not necessarily proportional to their prevalence and time periods are not weighted uniformly within each cohort. This is a particular instance of a well-known property of multiple regression coefficients, the "regression weighting problem" (Aronow and Samii, 2016). In estimating $\hat{\beta}$, units whose treatment status is well predicted by the covariates (in this case, the fixed effect parameters), are down-weighted when calculating the average while those whose treatment status is poorly predicted receive greater weights. In the presence of effect heterogeneity both over time and across units, this can give misleading inferences.

To provide a concrete illustration of both sources of bias in the two-way FE estimator, I consider a simple numerical example with $N = 5, T = 3$. Figure 1 presents a hypothetical treatment history assignment. Two of the units (1 and 2) initiate treatment in period 2 which carries over to period 3. Units 3 and 4 only initiate treatment in period 3. Unit 5 receives no treatment. All units are untreated in period 1.[9]

For a particular treated unit in period 2, that unit is matched to the three other observations under control in time 2 to construct the first difference term. However, the second-difference consists both of the valid differences from period 1, where no units are under treatment, and the invalid second differences from period 3 for units 3 and 4 which are under treatment in time 3. Because the two-way FE estimator has no concept of "ordering" when it comes to time, it treats future and past periods as equivalent.

[9] One could, of course, consider a much larger sample where the proportions of units assigned to each history remain the same. For ease of exposition, I limit the example to 5 but note that issues of bias are not simply confined to small samples. Rather, they are a function of the particular distribution of treatment histories in a given sample.



	Time		
Unit	1	2	3
1	0	1	1
2	0	1	1
3	0	0	1
4	0	0	1
5	0	0	0

Matched controls

	Time		
Unit	1	2	3
1	0	1	1
2	0	1	1
3	0	0	1
4	0	0	1
5	0	0	0

Valid second differences

	Time		
Unit	1	2	3
1	0	1	1
2	0	1	1
3	0	0	1
4	0	0	1
5	0	0	0

Invalid second differences

Figure 1: Illustration of valid and invalid second difference sets under a two-way fixed effects estimator

Figure 2 gives the implied weights on each treated unit for the two-way FE estimator. Note that for units 1 and 2 in period 3, the weights are *negative*, implying that those unit-periods are more often part of the second-difference term, acting as controls, than the first, acting as treated units. Even if we restrict the DID estimator to only those valid second differences, it does not solve the regression weighting problem. Effects in period 2 receive a greater weight when averaging than those in period 3, due to the reduced number of control observations in the third time period as illustrated in Figure 2.

Two-way fixed effects estimators will not yield a valid estimate of the treatment effect under parallel trends if treatment effects persist over time and are heterogeneous across cohort and time. Outside of the two-period setting, two-way FE relies heavily on the restrictions implied by the parametric model for Y_{it}. Two important restrictions are that A_{it} only affects Y_{it} and not future outcomes and that β is a constant. In most applied settings, however, treatment effects are rarely instantaneous and the impact of a particular intervention often takes many time periods to reveal itself. Additionally, units will typically respond differentially to treatment. Therefore, the underlying assumptions of the two-way fixed effects model are not plausible for most quantitative research. The alternative "generalized difference-in-differences" estimator outlined in this paper does not suffer from either of these drawbacks and is more appropriate in situations

	Time		
Unit	1	2	3
1	0	1	1
2	0	1	1
3	0	0	1
4	0	0	1
5	0	0	0

Treatment assignment

	Time		
Unit	1	2	3
1		5	-1
2		5	-1
3			4
4			4
5			

Implied weights on cohorts

	Time		
Unit	1	2	3
1		3	1
2		3	1
3			2
4			2
5			

Valid second differences only

Figure 2: Implied weights on treated units under a two-way fixed effects estimator

where treatment exhibits persistence after initiation.

4 Inverse propensity weighting estimators for multi-period DID effects

For many applications, even the parallel trends assumption is unlikely to hold unconditionally as factors associated with treatment history may also be associated with different pre-treatment paths. Therefore researchers will typically want to incorporate covariates in order to make the parallel trends assumption more credible. One of the reasons why two-way fixed effects models are so popular is that they facilitate easy inclusion of covariates as part of the outcome model. An alternative to regression modeling is inverse propensity weighting, which Abadie (2005) illustrates for the case of two-period difference-in-differences estimators. By up-weighting units with covariate profiles that are underrepresented among controls relative to treated units and down-weighting those that are overrepresented, weighting approaches allow for estimation when identification assumptions hold only conditionally. In this section I extend this approach to the generalized difference-in-differences estimand with more than two outcome periods and treatment histories.

I again focus on identification of the $CATT_t(c)$, the cohort ATT for a particular time period

t. Since this is identified by an average of two-period DID estimators, it is possible to apply the method in Abadie (2005) to each of these estimators in turn. Identification with covariates relies on a slightly weaker version of the parallel trends assumption

Assumption 5 *Conditional generalized parallel trends*

$$E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i = c, \vec{X_{ic}} = \vec{x}] = E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i \geq t, \vec{X_{ic}} = \vec{x}] \tag{16}$$

for all $t > c$, $t' \leq c$

In other words, the parallel trend from t' to t can vary depending on the value of $\vec{X_{ic}}$, the covariate profile of unit i up to time c.[10] When all covariates are time-constant, the time index is unnecessary. However, this formulation allows for the presence of covariates that follow patterns over time that may account for the violation of the unconditional parallel trends assumption. Note also that only the covariate history up to c is part of the conditioning set even when $t > c$. This is because values of X_{it} for periods after c are potentially affected by the treatment itself. Conditioning on variables potentially affected by the treatment risks post-treatment bias (Rosenbaum, 1984; Acharya, Blackwell and Sen, 2016; Montgomery, Nyhan and Torres, 2018).

Let $Pr(C_i = c|\vec{X_{ic}}, C_i \in \{c, t, \dots T\})$ denote the probability that a unit is in cohort c given both its covariate vector up to time c *and* knowing that C_i is either c or greater than or equal to t. Under the weaker assumption of conditional parallel trends, the $CATT_i(c)$ can be identified by

$$\begin{aligned} CATT_t(c) = \frac{1}{c}\sum_{t'=1}^{c} E\Bigg[& \frac{Y_{it} - Y_{it'}}{Pr(C_i = c|C_i \in \{c, t, \dots T\})} \times \\ & \frac{\mathcal{I}(C_i = c) - Pr(C_i = c|\vec{X_{ic}}, C_i \in \{c, t, \dots T\})}{1 - Pr(C_i = c|\vec{X_{ic}}, C_i \in \{c, t, \dots T\})}\Bigg] \end{aligned} \tag{17}$$

where $\mathcal{I}(C_i = c)$ is an indicator that takes on a value of 1 if $C_i = c$ and 0 otherwise. The reason for conditioning on $C_i \in \{c, t, \dots T\}$ is that for each estimate of $CATT_t(c)$, we are essentially subsetting the sample down to two sets of treatment histories, the one where treatment is initiated at $c + 1$ and the one where treatment is initiated after t. The former corresponds to the "treated"

[10] It is also necessary here to make a "positivity" or covariate overlap assumption, that the probability of observing a given treatment history conditional on the covariates is not perfectly zero or one.

group in a two-period DID while the no-reverse causality assumption implies that the latter are all equivalently "controls." In other words, we know that none of the matched control units will be ones that initiated treatment at t or earlier. With two periods and two treatment histories, we can apply weights in the vein of Abadie (2005). Intuitively, the treated units all receive a constant weight while the controls are re-weighted according to the ratio of the propensity that unit i would be treated and the propensity that it would be a control.

Under the assumption that no unit that initiates treatment reverts to being under control, the ratio of propensity scores $\frac{Pr(C_i=c|\vec{X_{ic}}, C_i \in \{c,t,\ldots T\})}{1-Pr(C_i=c|\vec{X_{ic}}, C_i \in \{c,t,\ldots T\})}$ can be written as a ratio of the probabilies that unit i initiates or does not initiate treatment in period c given that it has not initiated treatment at any previous point.

$$\frac{Pr(C_i = c|\vec{X_{ic}}, C_i \in \{c, t, \ldots T\})}{1 - Pr(C_i = c|\vec{X_{ic}}, C_i \in \{c, t, \ldots T\})} = \frac{Pr(A_{ic} = 1|\vec{A_{ic-1}} = \vec{0}, \vec{X_{ic}})}{Pr(A_{ic} = 0|\vec{A_{ic-1}} = \vec{0}, \vec{X_{ic}})} \tag{18}$$

with $\vec{0}$ again denoting a vector of all zeroes. With a large enough sample, it is possible to estimate the weights separately for each individual $CATT_t(c)$ by fitting a logistic regression model on each subset of the data, predicting cohort membership using $\vec{X_{ic}}$. However, in practice this will be infeasible because only a few units will be a part of each cohort. Additionally, with many time periods, the dimensionality of $\vec{X_{ic}}$ will be very large if there are time-varying covariates present. An approach to simplifying the problem would be to fit a pooled logistic regression model for the probability that a unit initiates treatment in a particular time period given the covariates. Because pooling across time periods involves pooling over covariate profiles $\vec{X_{it}}$ with varying dimensionality, this method of estimating the weights requires making additional modeling assumptions that restrict the number of lagged previous periods that can enter into the model. One simple approach is to assume that only the current values of X_{it} affect the probability of treatment initiation at time t. Then, one can estimate a pooled logistic regression to estimate $Pr(A_{it} = 1|\vec{A_{it-1}} = \vec{0}, X_{it})$ and obtain fitted values for the propensity scores. To account for varying propensities of treatment initiation over time, this model can incorporate a parametric time trend or time fixed-effect parameters. Inference can again be carried out using bootstrap methods, with the weighting model estimated repeatedly for each bootstrapped sample (Austin, 2016).

One advantage of the proposed weighting method as opposed to simply including confounders as regressors in the two-way FE linear model is that it avoids inducing post-treatment bias for long-term effects (Montgomery, Nyhan and Torres, 2018). Estimating a model of the form

$$Y_{it} = \alpha_i + \gamma_t + \beta A_{it} + \delta X_{it} + \varepsilon_{it} \tag{19}$$

where δ is a vector of coefficients on each covariate in X_{it} is common, but the coefficient on β will only represent a meaningful causal effect if there are no effects of treatment that persist beyond the first period.[11] This is because X_{it} is potentially affected by and part of the causal effect of A_{it-1}. Controlling for X_{it} blocks this causal pathway and may therefore attenuate treatment effects towards zero. Even if X_{it} is not a causal mechanism for treatment, it is still unwise to control for consequences of treatment as bias will also be induced if there exist common causes of X_{it} and Y (Elwert and Winship, 2014).

An alternative to weighting is to simply match each treated cohort to a set of controls based on the pre-treatment covariates and possibly lagged outcomes, as is suggested by Imai, Kim and Wang (2018). Certainly, this is also a feasible approach here and the choice of covariate adjustment method will depend on how researchers choose to resolve the bias-variance trade-off. Relative to weighting, matching is typically inefficient (Abadie and Imbens, 2006) but may provide additional advantages in terms of making the covariate adjustment procedure more transparent.

5 Application: The effect of investment treaties on U.S. foreign direct investment

Foreign direct investment (FDI) is an increasingly important vehicle for cross-border economic activity. Many firms look to situate elements of their production abroad and for governments looking to attract jobs and stimulate growth, competition for capital is fierce. In addition to direct incentives to foreign firms and domestic policy reforms, many governments have also looked to international legal arrangements to improve the attractiveness of their country to global capital. Bilateral Investment Treaties (BITs) have emerged as one of the most ubiqutious legal instru-

[11] Another issue with including covariates as part of the outcome model is that any time-invariant covariates will be perfectly colinear with the unit fixed effects and drop out of the model.

ments used in the governance of cross-border investment. Competitive pressures from other states (Elkins, Guzman and Simmons, 2006) and economic downturns (Simmons, 2014) often push states to sign these agreements with major capital exporters. BITs typically commit states to refrain from expropriation or discriminatory treatment of foreign investors, among other investment protection obligations. To enforce these commitments, many BITs also contain provisions for Investor-State Dispute Settlement (ISDS) in which states pre-commit to allow foreign investors covered under the BIT to pursue binding arbitration in an international forum such as the International Centre for the Settlement of Investment Disputes (ICSID) in the event of a breach of the treaty. In addition to these separate investment treaties, many states are now party to regional trade agreements that include investor-state dispute settlement provisions, for example, Chapter 11 of the North American Free Trade Agreement (NAFTA).

By granting foreign investors secure property rights that can be enforced outside of a country's domestic courts, BITs raise the cost of opportunistic expropriations. Kerner (2009) highlights two mechanisms through which BITs may, as a consequence, increase multinationals' activities in a host country. First, they solve a time-inconsistent preference problem by committing states to maintaining property rights protections after an investment has been made, and thereby reducing investors' anticipated risks ex-ante. Second, by imposing costs to violating property rights, BITs send a credible signal to uncertain foreign firms that ratifiers will respect investors' property. This latter mechanism operates beyond just investors that are able to access arbitration under the BIT and potentially affects overall FDI flows from non-covered source countries. Ultimately, BITs and their concommitant ISDS provisions function as a constraint on a government's future policy flexibility, in theory providing predictability to foreign investors who are then expected to be more willing to undertake costly and illiquid investments.

Whether BITs are actually effective in promoting investment remains an elusive empirical question. Unfortunately for scholars, states do not enter into BITs at random. Moreover, even measuring FDI is itself a challenge, with varying definitions of MNE activity potentially generating variable empirical results. Different statistical modeling approaches, specifications, and strategies to control for omitted variable bias have yielded differing results. While early studies in the literature showed no effect on FDI flows (Hallward-Driemeier, 2003), subsequent work highlighted a positive association between BITs and FDI flows (Egger and Pfaffermayr, 2004), though other researchers have found that the effect is conditional on strong domestic institutions (Tobin and Rose-Ackerman, 2011). While more recent papers that attempt to better adjust for the temporal dynamics in panel analyses of FDI find positive support for the hypothesis that BITs increase FDI (Egger and Merlo, 2007; Busse, Königer and Nunnenkamp, 2010), other studies cast some doubt on the credible commitment story (Yackee, 2010).

One of the central challenges in estimating the causal effect of investment treaties is choosing an appropriate modelling strategy for observational data. The workhorse model for many of the existing analyses of BITs and FDI in the literature is the linear dynamic panel approach (Tobin and Rose-Ackerman, 2011) which models the outcome for a unit i at time t, as a function of covariates (X_{it}), unit fixed effects γ_i and lagged outcome values (typically by one period). The basic dynamic regression model is therefore of the form:

$$Y_{it} = \alpha Y_{it-1} + \beta X_{it} + \gamma_i + \varepsilon_{it} \tag{20}$$

where ε_{it} is a mean zero random error.

This model assumes ignorability of treatment assignment in period t conditional on covariates and the lagged outcome in $t-1$. Incorporating the lagged outcome term is treated as a means of accounting for a possible reverse-causal relationship between BIT initiation and FDI. Unfortunately, adjusting for both outcome lags and unit fixed effects poses a number of statistical problems. It does not generally address the problem of confounding unless the model assumptions hold exactly. For example, if BIT ratification is predicted not by the 1-period lag, but rather by 3-, 4-, or 5- period lags, estimates will suffer from omitted variable bias (Bellemare, Masaki and Pepinsky, 2015). The validity of inferences depend heavily on the specific linear modeling assumptions chosen which can be difficult to validate.

In many cases, lags are omitted but additional fixed effect parameters for both time and unit are included.[12] For example, the dyadic regressions in Busse, Königer and Nunnenkamp (2010), Berger et al. (2013) and Aisbett, Busse and Nunnenkamp (2016) employ additive regression models with dummy variables for both year and dyad (the unit of analysis). As discussed in this paper, while

[12] Typically, researchers choose *either* to include the lagged outcome as a regressor *or* to use fixed effects when using ordinary least squares. Incorporating both risks biased OLS estimates for the parameters of the dynamic model when the number of time periods is small (Nickell, 1981).

such two-way fixed effect estimators are often motivated by an implicit difference-in-differences design, with the intent to adjust for unobserved time-constant confounders, they fail to do so in practice and require strong restrictions on how treatment can affect the outcome over time.

Many of the aforementioned studies also employ an instrumental variables strategy to address the problem of dynamic confounding. However, in all cases, the authors highlight the difficulty in justifying the exclusion restriction underpinning the instruments. Tobin and Rose-Ackerman (2011) admit that "good instruments are elusive, and weak at best" (15).[13] Both Tobin and Rose-Ackerman (2011) and Busse, Königer and Nunnenkamp (2010) employ (among other IV approaches) a Generalized Method of Moments (GMM) strategy, instrumenting for BITs using lagged values of the independent variables. However, this strategy is only valid if lagged treatments *only* affect the outcome through their effect on current treatment – that is, if the outcome model is exactly true and there are no persistent effects over time – precisely the assumption that this paper finds is highly problematic. In general, if the instrument affects the outcome via a mechanism outside of its effect on the treatment, instrumental variables estimates will be biased. Moreover, commonly used statistical tests for instrument validity in time-series instrumental variables models often fail to detect exclusion restriction violations due to lack of power (Bazzi and Clemens, 2013). Determining whether an instrument is valid is a matter for theory, not for statistical testing.

In lieu of these strong parametric methods, I employ the generalized DID estimator outlined in this paper, adjusting for possible violations of the parallel trends assumption due to observed covariates using inverse propensity weighting. Because dyadic analyses are fraught with issues of cross-dyadic dependence, complex correlation structures and clustering, I focus on the effect of investment treaties with a single country: the United States. This is a particularly good case to consider given the United States' long history of promoting BITs as a means of protecting its firms' investments abroad and the emphasis U.S. BITs in particular placed on investor-state dispute settlement (Vandevelde, 1993). Additionally, all U.S. BITs in force contain provisions for investor-state dispute settlement and almost all of the United States' recent preferential trade agreements have some sort of ISDS mechanism as well.

Focusing on U.S. foreign investment alone also helps address problems related to the measure-

[13] Tobin and Rose-Ackerman (2011) instrument BIT ratification with the number of BITs ratified by a country's neighbors. However, this instrument would be invalid if there exist unobserved regional patterns in FDI and BIT ratification such that a country's ratification patterns and its neighbors' ratification patterns are affected by a common cause.

ment of FDI. Constructing a valid measure of multinational involvement in a host economy is a difficult task. The most common measures of MNC activity with the broadest coverage temporally and spatially are FDI flow and stock data. Unfortunately, these measures are also often the least theoretically applicable. As Kerner (2014) notes, FDI flows and stocks are derived primarily from balance of payments statistics gathered by governments and central banks. Ultimately flows reflect the cumulation of cross-border financial transactions. For many research questions, these measures are poor proxies for the extent to which MNCs are willing to make investments in production within a country. Flows can be particularly misleading as an observation of $0 flows can reflect the absence of foreign affiliates or it can indicate that firms' repatriation of profits (net negative flow) matches the increase in firms' foreign position (e.g. via reinvested earnings) (Kerner, 2014). FDI stock data perhaps provides a better measure of the value of foreign-owned capital in a particular country, but proper valuation of stocks (whether on market value or by historical value) depends on the research question of interest.

As an alternative, Kerner and Lawrence (2014) points to fine-grained measures of firm expenditures collected at a national level. The United States Bureau of Economic Analysis (BEA) releases data from annual surveys of the worldwide activities of U.S. multinationals.[14] These surveys are conducted annually on each foreign affiliate of a U.S. parent company and are required of all affiliates that exceed a certain size threshold. While the surveys themselves are confidential, aggregate information is made public on assets held by foreign affiliates of U.S. multinational firms on an annual level. Unfortunately, for many United States partner countries, even the aggregate data is suppressed for reasons of data privacy. This results in high levels of potentially non-random missingness in the data. However, the BEA survey data does not suppress information on the number of U.S. affiliates and this data is available for every country in which U.S. investment is reported. An advantage of focusing on raw counts of multinationals is that it allows me to directly assess the effect of bilateral investment treaties on the *extensive* margin of foreign direct investment. That is, it specifically measures whether new firms are entering a market. It is difficult to distinguish from balance of payments or even aggregate asset data between existing firms increasing their investment position versus new entrants choosing to enter a market.

The dataset I assemble consists of observations of 157 countries over a period from 1983 to

[14] See `https://www.bea.gov/surveys/diasurv.htm`.

2013. The outcome variable is the number of U.S. multinational affiliates with assets, sales or net income over $25 million in the host country in a given year as reported by the BEA's US Direct Investment Abroad (USDIA) survey. Since the surveys only report data for countries in which any foreign direct investment was reported, not all countries are covered. Some country years are missing when no firms that are surveyed report any investment. When a country previously included in the survey is missing in a subsequent year, I impute the count of multinationals as 0 for that year. Of the 157 countries in the dataset, 46 entered into a bilateral investment treaty or a regional trade agreement (RTA) with investor-state dispute settlement provisions with the United States at some time during the period of observation.[15]. I omit all countries for which an investment treaty is in force for all time periods under observation.[16] Data on BIT and FTA entry into force is obtained from the UN Conference on Trade and Development's (UNCTAD) "International Investment Agreements Navigator."[17] For countries that enter into an investment agreement with the United States, I code the year of "treatment" initiation as the entry-into-force year if the entry-into-force month is prior to July. Otherwise, I code treatment initiation in the following year. Only one U.S. investment agreement has been formally terminated after initiation. Bolivia terminated its 2001 bilateral investment treaty in 2012. I therefore omit observations for Bolivia in 2012 and 2013 from the dataset. I obtain covariate information on the number of other BITs in force for a given country-year, using the same rules for determining starting year and again relying on the UNCTAD dataset. Data on Gross Domestic Product and real GDP per-capita is taken from the World Bank's World Development Indicators (WDI) database. I also obtain a measure of distance between each country's capital and Washington D.C. using latitude and longitude data from the WDI database API. Finally, I include two indicators of democratic governance from the Varieties of Democracy (V-Dem) project measuring electoral democracy and liberal democracy respectively. Additionally, because the counts of firms differ across states by orders of magnitude, it is implausible that the parallel trends assumption will hold on an additive scale. Some countries have thousands of U.S. affiliates while others have zero. Because the parallel trends assumption is sensitive to the scale of the outcome (Athey and Imbens, 2006), to make the assumption more plausible, I transform the outcome to a logarithmic scale by taking the natural

[15] Only one additional U.S. BIT partner is omitted from the data due to missing covariate data: Grenada

[16] These countries are Armenia, Bulgaria, the Czech Republic, Kyrgyzstan, Moldova, Mongolia and Slovakia.

[17] See `http://investmentpolicyhub.unctad.org/IIA`.

log of the raw counts plus 1 (to avoid issues with taking the log of 0).

Figure 3 displays the distribution of treatment uptake times for units in the dataset. Of note is the fact that there are few treatment cohorts with more than a single unit, making it infeasable to simply estimate the treatment effect for each unique cohort. Figure 4 plots the estimated treatment effects from both the two-way fixed effects model and the generalized difference-in-differences estimator with and without covariate adjustment. In the unadjusted results, the two-way fixed effects estimate is not statistically significant and is roughly half the size of the generalized DID estimate which is positive and statistically distinguishable from zero at $\alpha = 0.05$. This is consistent with the intuition that two-way FE estimators downweight longer term effects and upweight short-term effects. Since investment treaties are unlikely to exhibit their full effects immediately and operate through longer-term channels, the two-way FE estimator may be biased towards zero. Using the new generalized DID approach, I find that, for the typical country that adopts an investment treaty with the U.S., the treaty increases the count of U.S. MNE affiliates operating in that country by about .2 log-points, or approximately a 20% increase relative to that country's baseline.

To adjust for time-varying covariates in the outcome regression, I include in the two-way fixed effects OLS regression the linear combination of a country's logged real GDP per capita measured at time t, logged real GDP also measured at t, whether that country is a member of the GATT/WTO at time t, the two V-Dem democracy indices (electoral and liberal demoracy), and the log of the number of bilateral investment treaties that country has in force with non-U.S. partners at time t plus 1. Including these covariates drives the estimated effect from the two-way fixed effects model even closer to 0.

I use a similar additive model to estimate the propensity of treatment initiation. I estimate a logistic regression and add to the set of covariates above the log of the distance between the country's capital and Washington D.C. along with a linear time trend. Using this model, I predict the probability that each unit would initiate treatment at time t and use these weights to estimate the ACTT via the method described in section 4. Surprisingly, even with the weighting adjustment, the estimated treatment effect does not change substantially. The point estimate itself shifts by only about 0.02 log-points, with a slight increase in the standard error due to the weights. While this increase in variance does raise the corresponding p-value, the estimate is still statistically

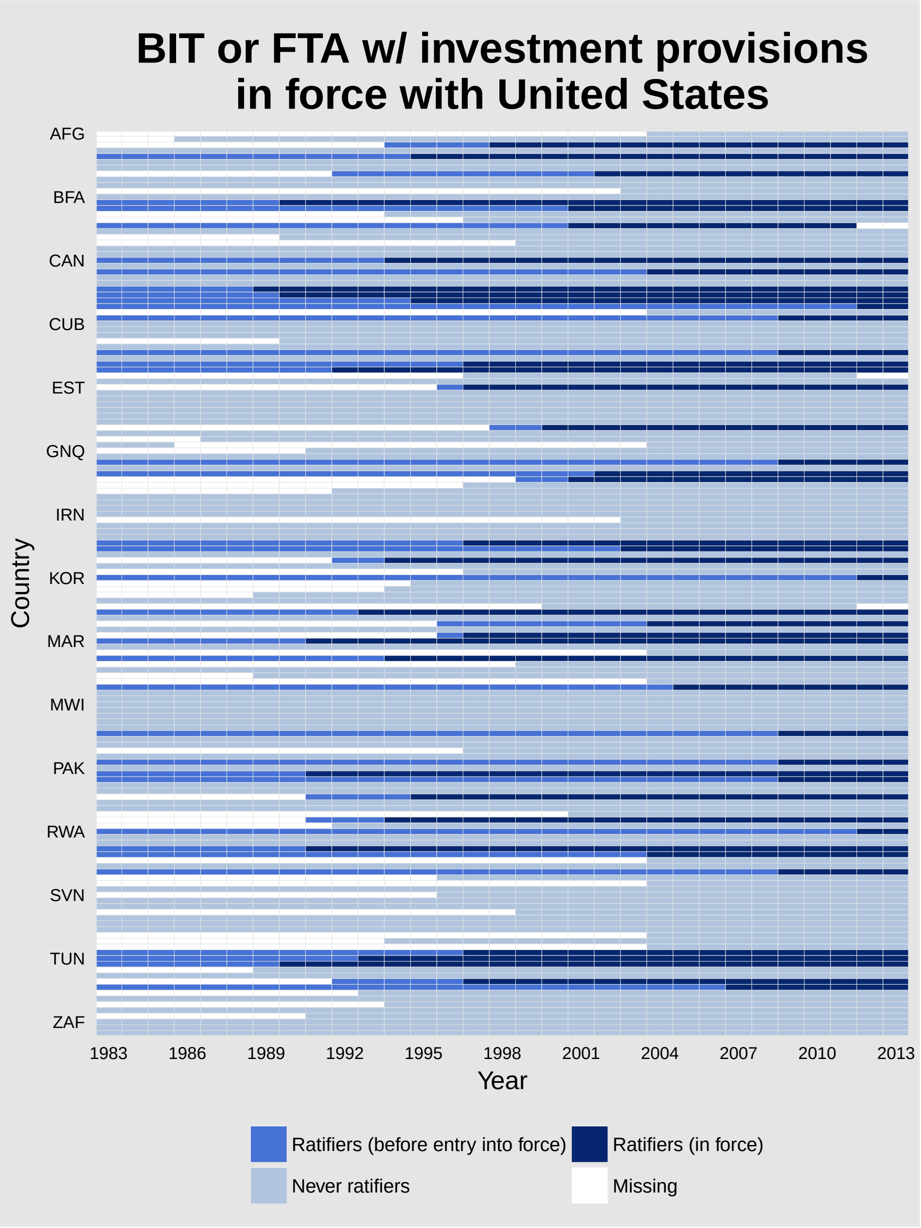
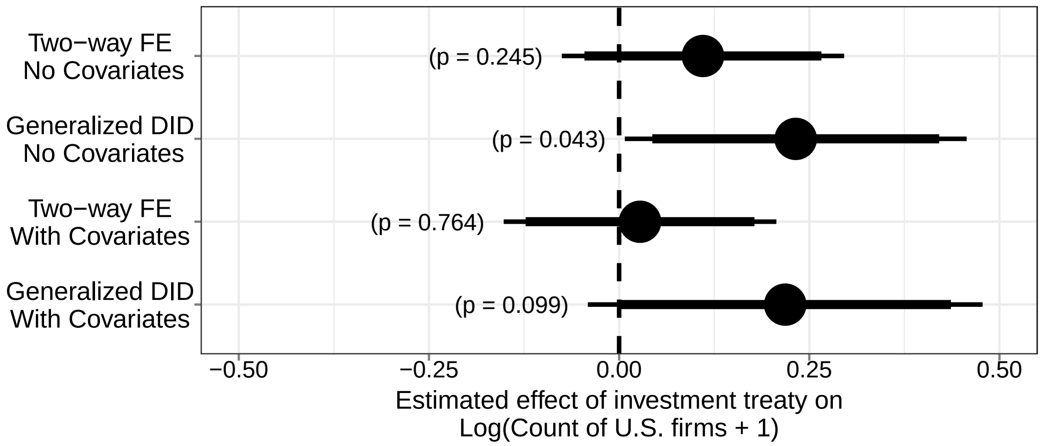


Figure 3: Distribution of investment treaties with the U.S. in force over time



Notes: Thick lines denote 90% asymptotic confidence intervals. Thin lines denote 95% confidence intervals. Standard errors estimated by block bootstrapping with 500 iterations. Confidence intervals based on normal approximation to the sampling distribution.

Figure 4: Estimated average effects of investment treaty entry-into-force on U.S. affiliates

significant at $\alpha = .1$. Additionally, the sizeable gap between the generalized DID estimate and the two-way FE regression estimate illustrates the pitfalls of conditioning on time-varying covariates that are potentially post-treatment. When the time-varying covariates are included with a method that avoids this issue, the positive effect of BIT initiation remains.

Overall I find, contrary to recent survey evidence suggesting BITs are irrelvant to firms' decisions to invest (Poulsen, 2010), that BITs increase the number of foreign affiliates from the BIT partner that operate in that country. While this estimate is limited to only U.S. BITs, it is not implausible that the effect generalizes to treaties with other capital exporting countries as investment treaties exhibit remarkable homogeneity across countries. It may well be that the effect of the BIT operates through subsequent policy changes in a host country that are ancillary to the treaty commitment itself. Therefore, even if firms do not directly respond to the presence or absence of an investment treaty when planning their investments, the treaty affects other variables that do factor into that decision. It is also possible that countries entering into an investment treaty with the United States are implementing a broader regime of capital-friendly policies that are difficult to disentangle from the BIT alone. Therefore, it is entirely within the realm of possibility that firms do not take BITs into account when investing, as Poulsen (2010) argues, but that effect of implementing a BIT does boost foreign direct investment overall.

6 Conclusion

This paper addresses a major flaw in the way researchers typically implement difference-in-differences estimators in panel data settings. Ordinary least squares with two-way fixed effects, while valid when there are two time periods and only two possible treatment assignment histories, is biased in the more general case of more than two time periods and treatment histories. Inference in this setting requires much stronger modeling assumptions in order to remain valid. I relax the most stringent of these assumptions, that treatment effects do not persist over time, to develop a non-parametric estimator under the constraint that units receiving treatment do not revert to control in subsequent time periods. I define a new quantity of interest, the Average Cohort Treatment effect on the Treated (ACTT), which corresponds to the average effect of initiating treatment for a unit randomly selected from the sample. I show that the ACTT is identified non-parametrically as a weighted average of two-period difference-in-differences estimates and provide a straightforward weighting method for relaxing the parallel trends assumption by conditioning on observed covariates.

One limitation of this analysis is that it considers only the case of the "static" two-way fixed effects model which does not include as regressors additional leads or lags of the treatment variable. It might then be argued that including such parameters would resolve any issues regarding treatment effects over multiple time periods as the model would include parameters for the "initial" effect of treatment and the treatment effect for subsequent periods as well. Unfortunately, the choice of the number of leads and lags remains up to the researcher, requiring an additional assumption about effect persistence. Moreover, including leads or lags of the treatment variable induces post-treatment bias in some of the coefficients in the model and not all parameters in the model will be causally interpretable in terms of counterfactual comparisons. This is particularly true if covariates are also included in the model since some time-varying covariates will be affected by model lags (Blackwell and Glynn, 2018). Additionally, as Abraham and Sun (2018) note, inclusion of leads and lags does not address issues of improper weighting of heterogeneous effects. Unfortunately, adding more parameters to the two-way FE model is not an adequate panacea for the problem identified in this paper.

The method described here is also limited to the case where units do not revert to control after initiating treatment. While this reflects many situations encountered by social scientists, particularly in studies of policy implementation, some types of treatments exhibit reversion over time, especially when the number of time periods under consideration is large. For example, researchers studying democratization have to consider the possibility of democratic backsliding. Whether a country that has always been an autocracy should be considered as having the same "treatment" condition as one that democratized and subsequently reverted to being a non-democracy is a substantive question for which the answer will depend on the particular research question being asked. Subsequent work should consider other approaches to reducing the dimensionality of this treatment history space without necessarily requiring persistence of treatment uptake, but also not restricting treatment effects to single periods as in the two-way FE estimator.

It is worth noting also the connection between the method outlined here and two other recent papers that address inference in a time-series setting. First, Imai, Kim and Wang (2018) outline a matching approach that shares many similarities with the method described in this paper. They propose inference on the ATT by matching treated units with control units that have identical treatment histories upt until the point of treatment (along with similar covariate profiles as measured by Mahalanobis distance). An advantage of this method is that it permits researchers to apply the method to settings where units have any arbitrary treatment pattern. However, this comes at the cost of requiring that researchers also pre-specify the number of past treatment periods on which to match, permitting units that had previously been under treatment to still act as control if they have reverted to control for a sufficient period of time. In the context of the situation studied here, the matching sets for any treatment history/time period combination are straightforwardly defined and in principle, the matching method of Imai, Kim and Wang (2018) could be used instead of weighting to obtain estimates for the effect of each treatment history. What this paper potentially adds to the method Imai, Kim and Wang (2018) is in defining an *aggregate* quantity of interest that combines both short and long-term effects – the ACTT – potentially obviating the need for researchers to specify in advance the number of future periods for which they want to estimate the effect.

Conversely, Hazlett and Xu (2018) focus on the two-treatment history setting where all units that initiate treatment do so at a single time period. In the context considered here, the goal is to obtain balance on pre-treatment outcomes and covariates via mean-balancing (Hainmueller, 2012), either on the raw outcome space or in a transformation to a higher-dimensional space via kernel methods (Hazlett, 2016). In principle, the same exact-balancing approach utilized by Hazlett and Xu (2018) could be used to estimate the weights in Section 4 by simply estimating weights separately for each treatment-history/time-period combination. In fact, this may help avoid some of the pitfalls that often occur with misspecification in IPTW models (Kang, Schafer et al., 2007; Imai and Ratkovic, 2014). However, when there are few observations taking on any given treatment history, these weights may become unstable in the absence of additional regularization or pooling as is used for the IPTW method described in this paper. Nevertheless, given the good performance of balancing methods in cases where adequate weights can be found, this is likely the next logical extension of the weighting method described in this paper.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The Review of Economic Studies* 72(1):1–19.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program." *Journal of the American statistical Association* 105(490):493–505.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2015. "Comparative politics and the synthetic control method." *American Journal of Political Science* 59(2):495–510.

Abadie, Alberto and Guido W Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *econometrica* 74(1):235–267.

Abraham, Sarah and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Working Paper* .

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3):512–529.

Aisbett, Emma, Matthias Busse and Peter Nunnenkamp. 2016. "Bilateral investment treaties do work: Until they don't." *Kiel Working Paper* .

Angrist, Joshua D and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

Athey, Susan and Guido W Imbens. 2006. "Identification and inference in nonlinear difference-in-differences models." *Econometrica* 74(2):431–497.

Austin, Peter C. 2016. "Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis." *Statistics in medicine* 35(30):5642–5655.

Bazzi, Samuel and Michael A Clemens. 2013. "Blunt instruments: avoiding common pitfalls in identifying the causes of economic growth." *American Economic Journal: Macroeconomics* 5(2):152–186.

Bellemare, Marc F, Takaaki Masaki and Thomas B Pepinsky. 2015. "Lagged explanatory variables and the estimation of causal effects." *Working Paper* .

Berger, Axel, Matthias Busse, Peter Nunnenkamp and Martin Roy. 2013. "Do trade and investment agreements lead to more FDI? Accounting for key provisions inside the black box." *International Economics and Economic Policy* 10(2):247–275.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics* 119(1):249–275.

Blackwell, Matthew and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *Working Paper* .

Borusyak, Kirill and Xavier Jaravel. 2017. "Revisiting Event Study Designs with an Application to the Estimation of the Marginal Propensity to Consume." *Working Paper* .

Busse, Matthias, Jens Königer and Peter Nunnenkamp. 2010. "FDI promotion through bilateral investment treaties: more than a bit?" *Review of World Economics* 146(1):147–177.

de Chaisemartin, Clement and Xavier D'Haultfoeuille. 2018. "Double fixed effects estimators with heterogeneous treatment effects." *arXiv preprint arXiv:1803.08807* .

Egger, Peter and Michael Pfaffermayr. 2004. "The impact of bilateral investment treaties on foreign direct investment." *Journal of comparative economics* 32(4):788–804.

Egger, Peter and Valeria Merlo. 2007. "The impact of bilateral investment treaties on FDI dynamics." *The world economy* 30(10):1536–1549.

Elkins, Zachary, Andrew T Guzman and Beth A Simmons. 2006. "Competing for capital: The diffusion of bilateral investment treaties, 1960–2000." *International Organization* 60(04):811–846.

Elwert, Felix and Christopher Winship. 2014. "Endogenous selection bias: The problem of conditioning on a collider variable." *Annual Review of Sociology* 40:31–53.

Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* 20(1):25–46.

Halliwell, Jonathan J, Juan Pérez-Mercader and Wojciech Hubert Zurek. 1996. *Physical origins of time asymmetry*. Cambridge University Press.

Hallward-Driemeier, Mary. 2003. "Do bilateral investment treaties attract foreign direct investment? only a bit and they could bite." *World Bank Policy Research Working Paper No. 3121* .

Hazlett, Chad. 2016. "Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects." *Working Paper* .

Hazlett, Chad and Yiqing Xu. 2018. "Trajectory Balancing: A General Reweighting Approach to Causal Inference with Time-Series Cross-Sectional Data." *Working Paper* .

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396):945–960.

Imai, Kosuke and In Song Kim. 2017. "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *Working Paper* .

Imai, Kosuke, In Song Kim and Erik Wang. 2018. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." *Working Paper* .

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics* 86(1):4–29.

Kang, Joseph DY, Joseph L Schafer et al. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science* 22(4):523–539.

Kerner, Andrew. 2009. “Why should I believe you? The costs and consequences of bilateral investment treaties.” *International Studies Quarterly* 53(1):73–102.

Kerner, Andrew. 2014. “What we talk about when we talk about foreign direct investment.” *International Studies Quarterly* 58(4):804–815.

Kerner, Andrew and Jane Lawrence. 2014. “What’s the risk? Bilateral investment treaties, political risk and fixed capital accumulation.” *British Journal of Political Science* 44(01):107–121.

Lewis, David. 1979. “Counterfactual dependence and time’s arrow.” *Noûs* pp. 455–476.

Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 2018. “How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* .

Neyman, JS. 1923. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Translated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480).” *Annals of Agricultural Sciences* 10:1–51.

Nickell, Stephen. 1981. “Biases in dynamic models with fixed effects.” *Econometrica: Journal of the Econometric Society* pp. 1417–1426.

Poulsen, Lauge. 2010. The importance of BITs for foreign direct investment and political risk insurance: Revisiting the evidence. In *Yearbook on International Investment Law and Policy 2009-2010*. Oxford University Press pp. 539–574.

Rosenbaum, Paul R. 1984. “The consequences of adjustment for a concomitant variable that has been affected by the treatment.” *Journal of the Royal Statistical Society. Series A (General)* pp. 656–666.

Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of educational Psychology* 66(5):688.

Rubin, Donald B. 1986. “Comment: Which ifs have causal answers.” *Journal of the American Statistical Association* 81(396):961–962.

Salacuse, Jeswald W and Nicholas P Sullivan. 2005. "Do BITs really work: An evaluation of bilateral investment treaties and their grand bargain." *Harv. Int'l LJ* 46:67.

Simmons, Beth A. 2014. "Bargaining over BITs, arbitrating awards: The regime for protection and promotion of international investment." *World Politics* 66(01):12–46.

Tobin, Jennifer L and Susan Rose-Ackerman. 2011. "When BITs have some bite: The political-economic environment for bilateral investment treaties." *The Review of International Organizations* 6(1):1–32.

VanderWeele, Tyler and Stijn Vansteelandt. 2009. "Conceptual issues concerning mediation, interventions and composition." *Statistics and its Interface* 2:457–468.

Vandevelde, Kenneth J. 1993. "Of Politics and Markets: The Shifting Ideology of the BITs." *Int'l Tax & Bus. Law.* 11:159.

Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25(1):57–76.

Yackee, Jason Webb. 2010. "Do Bilateral Investment Treaties Promote Foreign Direct Investment-Some Hints from Alternative Evidence." *Va. J. Int'l L.* 51:397.

7 Appendix

Proof for Proposition 1

The two-way fixed effects estimator assumes the following data-generating process

$$E[Y_{it}|A_{it}] = \alpha_i + \gamma_t + \beta A_{it}$$

where α_i denotes unit fixed effects, γ_t denotes time fixed effects and β is the quantity of interest.

The ordinary least squares estimates of the parameters: $\hat{\beta}$, $\hat{\gamma}$, $\hat{\alpha}$ solve the least-squares optimization problem

$$\hat{\beta}, \hat{\gamma}, \hat{\alpha} = \underset{\beta,\gamma,\alpha}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \gamma_t - \beta A_{it})^2$$

The first-order conditions for $\hat{\beta}$

$$\begin{aligned}
0 &= \sum_{i=1}^{N} \sum_{t=1}^{T} -2A_{it} \left(Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta} A_{it} \right) \\
0 &= \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left(Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t \right) - \hat{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2 \\
\hat{\beta} &= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left(Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left(Y_{it} - (\hat{\alpha}_i + \hat{\gamma}_t) \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}
\end{aligned}$$

For $\hat{\alpha}_i$, the FOCs are:

$$\begin{aligned}
0 &= \sum_{t=1}^{T} -2(Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta} A_{it}) \\
\hat{\alpha}_i &= \frac{1}{T} \sum_{t=1}^{T} Y_{it} - \frac{1}{T} \sum_{t=1}^{T} \hat{\gamma}_t - \hat{\beta} \frac{1}{T} \sum_{t=1}^{T} A_{it}
\end{aligned}$$

And for $\hat{\gamma}_t$

$$\begin{aligned}
0 &= \sum_{n=1}^{N} -2(Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta} A_{it}) \\
\hat{\gamma}_t &= \frac{1}{N} \sum_{n=1}^{N} Y_{it} - \frac{1}{N} \sum_{n=1}^{N} \hat{\alpha}_i - \hat{\beta} \frac{1}{N} \sum_{i=1}^{N} A_{it}
\end{aligned}$$

Let $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$, $\bar{Y}_t = \frac{1}{N} \sum_{i=1}^{N} Y_{it}$, $\bar{\bar{Y}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}$, $\bar{A}_i = \frac{1}{T} \sum_{t=1}^{T} A_{it}$, $\bar{A}_t = \frac{1}{N} \sum_{i=1}^{N} A_{it}$, $\bar{\bar{A}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}$.

Then, re-write the first-order conditions

$$\hat{\gamma_t} = \bar{Y_t} - \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha_i} - \hat{\beta}\bar{A_t}$$

$$\hat{\alpha_i} = \bar{Y_i} - \frac{1}{T}\sum_{t=1}^{T}\hat{\gamma_t} - \hat{\beta}\bar{A_i}$$

Substituting one into the other

$$\begin{aligned}\hat{\alpha_i} &= \bar{Y_i} - \frac{1}{T}\sum_{t=1}^{T}\left[\bar{Y_t} - \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha_i} - \hat{\beta}\bar{A_t}\right] - \hat{\beta}\bar{A_i} \\ &= \bar{Y_i} - \bar{\bar{Y}} + \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha_i} + \hat{\beta}\bar{\bar{A}} - \hat{\beta}\bar{A_i}\end{aligned}$$

Then substituting into the expression for $\hat{\alpha_i} + \hat{\gamma_t}$

$$\begin{aligned}\hat{\alpha_i} + \hat{\gamma_t} &= \bar{Y_i} - \bar{\bar{Y}} + \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha_i} + \hat{\beta}\bar{\bar{A}} - \hat{\beta}\bar{A_i} + \bar{Y_t} - \frac{1}{N}\sum_{i=1}^{N}\hat{\alpha_i} - \hat{\beta}\bar{A_t} \\ &= \bar{Y_t} + \bar{Y_i} - \bar{\bar{Y}} + \hat{\beta}\bar{\bar{A}} - \hat{\beta}\bar{A_i} - \hat{\beta}\bar{A_t} \\ &= \bar{Y_t} + \bar{Y_i} - \bar{\bar{Y}} - \hat{\beta}(\bar{A_i} + \bar{A_t} - \bar{\bar{A}})\end{aligned}$$

Substituting back into $\hat{\beta}$

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \left(\bar{Y}_t + \bar{Y}_i - \bar{\bar{Y}} - \hat{\beta}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})\right)\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}} + \hat{\beta}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}\right) + \hat{\beta}\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} + \hat{\beta}\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta} - \hat{\beta}\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta}\left[1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}(\bar{A}_i + \bar{A}_t - \bar{\bar{A}})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2}\right] &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(A_{it})^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(A_{it} - \bar{A}_i - \bar{A}_t + \bar{\bar{A}}\right)}
\end{aligned}$$

It is possible to expand the expression in the numerator to:

$$Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{\bar{Y}}$$

$$Y_{it} - \frac{1}{N}\sum_{i'=1}^{N} Y_{i't} - \frac{1}{T}\sum_{t'=1}^{T} Y_{it'} + \frac{1}{NT}\sum_{i'=1}^{N}\sum_{t'=1}^{T} Y_{i't'}$$

$$\left[1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT}\right] Y_{it} - \frac{1}{N}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T}Y_{i't}\right] - \frac{1}{T}\sum_{t'\neq t}\left[Y_{it'} - \frac{1}{N}Y_{it'}\right] + \frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t} Y_{i't'}$$

$$\frac{NT - N - T + 1}{NT} Y_{it} - \frac{1}{N}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T}Y_{i't}\right] - \frac{1}{T}\sum_{t'\neq t}\left[Y_{it'} - \frac{1}{N}Y_{it'}\right] + \frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t} Y_{i't'}$$

$$\frac{(N-1)(T-1)}{NT} Y_{it} - \frac{T-1}{NT}\sum_{i'\neq i} Y_{i't} - \frac{N-1}{NT}\sum_{t'\neq t} Y_{it'} + \frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t} Y_{i't'}$$

$$\frac{(N-1)(T-1)}{NT}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \frac{1}{NT}\sum_{i'\neq i}\left[(T-1)Y_{i't} - \sum_{t'\neq t} Y_{i't'}\right]$$

$$\frac{(N-1)(T-1)}{NT}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \frac{T-1}{NT}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T-1}\sum_{t'\neq t} Y_{i't'}\right]$$

$$\frac{(T-1)}{NT}\left\{(N-1)\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T-1}\sum_{t'\neq t} Y_{i't'}\right]\right\}$$

$$\frac{(N-1)(T-1)}{NT}\left\{\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \frac{1}{N-1}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T-1}\sum_{t'\neq t} Y_{i't'}\right]\right\}$$

Denote the normalizing constant C with

$$C = \frac{\frac{(N-1)(T-1)}{NT}}{\sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left(A_{it} - \bar{A}_i - \bar{A}_t + \bar{\bar{A}}\right)}$$

Then write $\hat{\beta}$ as

$$\begin{aligned}\hat{\beta} &= C \times \sum_{i=1}^{N}\sum_{t=1}^{T} A_{it}\left\{\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \frac{1}{N-1}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T-1}\sum_{t'\neq t} Y_{i't'}\right]\right\} \\ &= C \times \sum_{t=1}^{T}\left\{\sum_{i=1}^{N} A_{it}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - \sum_{i=1}^{N} A_{it}\frac{1}{N-1}\sum_{i'\neq i}\left[Y_{i't} - \frac{1}{T-1}\sum_{t'\neq t} Y_{i't'}\right]\right\}\end{aligned}$$

Let $N_t^{(a)}$ denote the number of units that are under treatment a in period t. Let $T_i^{(a)}$ denote the number of periods for which unit i receives treatment a.

Note that in the second difference term, every unit under control in period t appears here $N_t^{(1)}$ times. Likewise, every unit under treatment appears $N_t^{(1)} - 1$ times, excluding the period when it is in the first difference. Therefore, re-write the expression as

$$\begin{aligned}\hat{\beta} = C \times \sum_{t=1}^{T}\sum_{i=1}^{N} A_{it}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'} - \frac{N_t^{(1)} - 1}{N-1}\left(Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right)\right] - \\ (1 - A_{it})\frac{N_t^{(1)}}{N-1}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right]\end{aligned}$$

$$\hat{\beta} = C \times \sum_{t=1}^{T}\left\{\sum_{i=1}^{N} A_{it}\frac{N_t^{(0)}}{N-1}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right] - (1 - A_{it})\frac{N_t^{(1)}}{N-1}\left[Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right]\right\}$$

Re-arranging terms yields

$$\begin{aligned}\hat{\beta} = C \times \sum_{t=1}^{T}\left\{\frac{N_t^{(0)}}{N-1}\sum_{i=1}^{N} Y_{it}A_{it} - \frac{N_t^{(1)}}{N-1}\sum_{i=1}^{N} Y_{it}(1 - A_{it})\right\} - \\ C \times \sum_{t=1}^{T}\left\{\frac{N_t^{(0)}}{(N-1)(T-1)}\sum_{i=1}^{N} A_{it}\sum_{t'\neq t} Y_{it'} - \frac{N_t^{(1)}}{(N-1)(T-1)}\sum_{i=1}^{N}(1 - A_{it})\sum_{t'\neq t} Y_{it'}\right\}\end{aligned}$$

$$\hat{\beta} = C \times \frac{1}{(N-1)(T-1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} Y_{it} - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} Y_{it} \right\} -$$

$$C \times \frac{1}{(T-1)(N-1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} Y_{it'} - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} Y_{it'} \right\}$$

$$\hat{\beta} \; = \; C \times \frac{1}{(N-1)(T-1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} [Y_{it} - Y_{it'}] \right\}$$

$$\hat{\beta} \; = \; C \times \frac{1}{(N-1)(T-1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} [Y_{it} - Y_{it'}] \right\}$$

$$\hat{\beta} = C \times \frac{1}{(N-1)(T-1)} \sum_{t=1}^{T} \left\{ \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}] \right\}$$

Incorporating the normalizing constant, we have a uniform average over all differences-in-differences.

$$\hat{\beta} = \frac{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{[Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}]\}}{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{1 - A_{it'} + A_{jt'}\}}$$