On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data[*]

Kosuke Imai[†] In Song Kim[‡]

March 2, 2019

Abstract

The two-way linear fixed effects regression has become a default model for estimating causal effects from panel data. It is well known that in a simple case of two time periods and two groups this estimator is equivalent to the difference-in-difference estimator. Unfortunately, this equivalence relationship does not hold in more typical settings where units may switch in and out of the treatment group at different points in time. This implies that the two-way fixed effects estimators are generally biased under the standard identification assumptions of difference-in-differences design. Using the matching framework of causal inference, we propose a multi-period difference-in-differences estimator, which eliminates this bias. We show that this estimator is equivalent to the *weighted* two-way fixed effects regression estimator and develop a specification test for the standard two-way fixed effects regression model. We apply our methodology to the controversy about whether GATT/WTO increases trade volume. An open-source software package is available for implementing the proposed methodology.

Key Words: difference-in-differences, longitudinal data, matching, unobserved confounding, weighted least squares

[*] The methods described in this paper can be implemented via the open-source statistical software, `wfe: Weighted Linear Fixed Effects Estimators for Causal Inference`, available through the Comprehensive R Archive Network (https://cran.r-project.org/package=wfe). Earlier versions of this paper were entitled, "Understanding and Improving Linear Fixed Effects Regression Models for Causal Inference," and "On the Use of Linear Fixed Effects Regression Estimators for Causal Inference" (Imai and Kim, 2011). The results concerning the one-way fixed effects estimator in these papers are published as Imai and Kim (2019).

[†] Professor, Department of Government and Department of Statistics, Harvard University, Institute for Quantitative Social Science, Cambridge MA 02138. Phone: 617–384–6778, Email: Imai@Harvard.Edu, URL: https://imai.fas.harvard.edu

[‡] Associate Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge MA 02142. Phone: 617–253–3138, Email: insong@mit.edu, URL: http://web.mit.edu/insong/www/

1 Introduction

Many social scientists use linear regression models with time and unit fixed effects (hereafter two-way fixed effects models) as the default methodology for estimating causal effects from panel data. This practice is often motivated by the fact that the two-way fixed effects regression estimator is equivalent to the difference-in-differences estimator under the simple setting with two time periods and two groups (e.g., Bertrand *et al.*, 2004; Angrist and Pischke, 2009). Unfortunately, it is well known that this equivalence relationship does not hold in more general settings, in which units can go in and out of the treatment condition at different points in time (see e.g., Imai and Kim, 2011; Borusyak and Jaravel, 2017; Abraham and Sun, 2018; Athey and Imbens, 2018; Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018). This implies that the two-way fixed effects estimator is generally biased under the parallel trend assumption of the standard difference-in-differences design (Section 2).

In this paper, using the matching framework of Imai and Kim (2019), we propose a multi-period difference-in-differences estimator that eliminates this bias (Section 3). We show that this estimator is equivalent to the *weighted* two-way fixed effects estimator where some of the weights can be negative. This equivalence result facilitates the efficient computation of the estimator and yields a model-based cluster robust standard error for the proposed estimator. In addition, the difference between the weighted and unweighted two-way fixed effects estimators can be used as a specification test for the standard two-way fixed effects model.

Our result contributes to the fast growing literature on the causal inference methods for panel data. Abraham and Sun (2018), Athey and Imbens (2018), and Goodman-Bacon (2018) study the two-way fixed effects and related estimators under the staggered adoption setting, in which units may switch from the control condition to the treatment condition at different points in time but never switch back to the control condition. In contrast, we consider a more general setting where units can go back and forth beween the treatment and control conditions. Chaisemartin and D'Haultfœuille (2018) investigate a slightly more general setting than staggered adoption but under the assumption of treatment monotonicity where the treatment status is stochastically increasing within a group. Our result is also related to that of Arkhangelsky *et al.* (2018) who shows that the weighted two-way fixed effects estimator can improve the synthetic control method of Abadie *et al.* (2010) in the settings where a single unit receives the treatment at the final time period. We, on the other hand, connect the weighted two-way fixed effects estimator to the multi-period difference-in-difference estimator under a much more general setting. The weighted regression formulation clarifies the extent to which each observation contributes to the resulting causal inference (see e.g., Humphreys, 2009; Aronow and Samii, 2015; Solon *et al.*, 2015, for the use of weighted regressions for causal inference with cross-sectional data).

In Section 4, we illustrate the proposed methodology by revisiting the controversy about whether GATT (General Agreement on Tariffs and Trade) membership increases international trade. We show that the empirical results based on the weighted two-way fixed effects estimator substantially differ from those based on the standard two-way fixed effects estimator. Specifically, we find little evidence that joint participation in the international legal agreement to reduce global trade barriers increases bilateral trade. The final section provides concluding remarks and suggestions for applied researchers. An open-source software package, `wfe: Weighted Linear Fixed Effects Estimators for Causal Inference`, is available from the Comprehensive R Archive Network (CRAN; `https://cran.r-project.org/package=wfe`) for implementing the proposed methods.

2 The Two-way Fixed Effects Regression Estimator

In this section, we introduce the two-way fixed effects regression estimator and examine the set of assumptions it requires for causal inference. We show that the two-way fixed effects regression estimator can be written as an weighted average of three different causal effects estimators, making its causal interpretation difficult. We also point out the absence of dynamic causal relationships between the treatment and outcome variables as the fundamental restriction of the two-way fixed effects regression estimator.

2.1 The Estimator

Suppose that we have a panel data set of N units and T time periods. Although our methodology readily extends to the case of unbalanced panel (as demonstrated in our empirical analysis of Section 4), for the sake of notational simplicity, we assume a balanced panel data set when describing the methodological development. Let X_{it} and Y_{it} represent the binary treatment indicator and observed outcome variables for unit i at time t, respectively. We consider the following two-way linear fixed effects regression model,

$$Y_{it} \quad = \quad \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it} \tag{1}$$

for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ where α_i and γ_t are unit and time fixed effects, respectively.

The inclusion of unit and time fixed effects accounts for both unit-specific (but time-invariant) and time-specific (but unit-invariant) unobserved confounders in a flexible manner. Let $\mathbf{U}_i$ and $\mathbf{V}_t$ represent these unit-specific and time-specific unobserved confounders, respectively. Then, we can define unit and time effects as $\alpha_i = h(\mathbf{U}_i)$ and $\gamma_t = f(\mathbf{V}_t)$ where $h(\cdot)$ and $f(\cdot)$ are arbitrary functions unknown to researchers. Thus, although the interaction between these two types of unobserved confounders is assumed to be absent, there is no functional-form restriction on $h(\cdot)$ and $f(\cdot)$.

The least squares estimate of β can be computed efficiently by transforming the outcome and treatment variables and then regressing the former on the latter. Formally, the estimator is given by,

$$\hat{\beta}_{\mathsf{FE2}} \ = \ \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\{(Y_{it} - \overline{Y}) - (\overline{Y}_i - \overline{Y}) - (\overline{Y}_t - \overline{Y})\} - \beta\{(X_{it} - \overline{X}) - (\overline{X}_i - \overline{X}) - (\overline{X}_t - \overline{X})\}\right]^2 \tag{2}$$

where $\overline{Y}_i = \sum_{t=1}^{T} Y_{it}/T$ and $\overline{X}_i = \sum_{t=1}^{T} X_{it}/T$ are unit-specific means, $\overline{Y}_t = \sum_{i=1}^{n} Y_{it}/N$ and $\overline{X}_t = \sum_{i=1}^{n} X_{it}/N$ are time-specific means, and $\overline{Y} = \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}/NT$ and $\overline{X} = \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}/NT$ are overall means. The large-sample conditional variance can be estimated using the standard cluster-robust sandwich formula with a multiplicative degrees of freedom adjustment (e.g., Arellano, 1987; Hansen, 2007; Cameron and Miller, 2015),

$$\widehat{\mathbb{V}(\hat{\beta}_{\textsf{FE2}} \mid \mathbf{X}, \mathbf{Y})} \;=\; \frac{N(NT-1)}{(N-1)(NT-N-T-1)} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \widetilde{\mathbf{X}}_i\right)^{-1} \left\{\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\top} \widetilde{\mathbf{X}}_i\right\} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \widetilde{\mathbf{X}}_i\right)^{-1} \tag{3}$$

where $\widetilde{\mathbf{X}}_i$ be an $T \times 1$ vector whose tth element equals $X_{it} - \overline{X}_i - \overline{X}_t + \overline{X}$, $\widetilde{\mathbf{Y}}_i$ is an T-dimensional vector whose tth element equals $Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}$, and $\hat{\boldsymbol{\epsilon}}_i = \widetilde{\mathbf{Y}}_i - \widetilde{\mathbf{X}}_i \hat{\beta}_{\textsf{FE2}}$ is the corresponding vector of residuals.

Equation (2) shows how the two-way fixed effects estimator exploits the covariation in the outcome and treatment variables. Specifically, the equation shows that least squares estimation is applied after the within-unit and within-time variations are subtracted from the overall variation for both outcome and treatment variables. In fact, we can show that the two-way fixed effects estimator can be represented as the weighted average of the following three estimators, i.e., the unit fixed effects, time fixed effects, and pooled regression estimators.

$$\hat{\beta}_{\textsf{FEunit}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \{(Y_{it} - \overline{Y}_i) - \beta(X_{it} - \overline{X}_i)\}^2 \tag{4}$$

$$\hat{\beta}_{\textsf{FEtime}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \{(Y_{it} - \overline{Y}_t) - \beta(X_{it} - \overline{X}_t)\}^2 \tag{5}$$

$$\hat{\beta}_{\textsf{pool}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \beta X_{it})^2 \tag{6}$$

The following proposition presents this result.

Proposition 1 (Two-way Fixed Effects Estimator as a Weighted Average of One-way Fixed Effects and Pooled Regression Estimators) *When N and T are sufficiently large, the two-way fixed effects estimator defined in equation* (2) *is approximated by a weighted average of the unit fixed effects regression estimator defined in equation* (4)*, the time fixed effects estimator defined in equation* (5)*, and the pooled regression estimator defined in equation* (6)*,*

$$\hat{\beta}_{\textsf{FE2}} \;\approx\; \frac{\omega_{\textsf{FEunit}} \times \hat{\beta}_{\textsf{FEunit}} + \omega_{\textsf{FEtime}} \times \hat{\beta}_{\textsf{FEtime}} - \omega_{\textsf{pool}} \times \hat{\beta}_{\textsf{pool}}}{w_{\textsf{FEunit}} + w_{\textsf{FEtime}} - w_{\textsf{pool}}}$$

where

$$\omega_{\mathsf{FEunit}} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} S_i^2, \quad \omega_{\mathsf{FEtime}} \quad = \quad \frac{1}{T}\sum_{t=1}^{T} S_t^2, \quad \omega_{\mathsf{pool}} \quad = \quad S^2$$

with $S_i^2 = \sum_{t=1}^{T}(X_{it} - \overline{X}_i)^2/(T-1)$*,* $S_t^2 = \sum_{i=1}^{N}(X_{it} - \overline{X}_t)^2/(N-1)$*, and* $S^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \overline{X})^2/(NT-1)$*.*

Proof is given in Appendix A. The weights for the unit fixed effects, time fixed effects, and pooled estimators are proportional to the average variances of the treatment variable within unit, within time, and within the entire data set, respectively. This is analogous to the result that the one-way fixed effects estimator is approximately equal to the variance-weighted average of average outcome difference between the treated and control observations within each unit (e.g., Chernozhukov *et al.*, 2013; Imai and Kim, 2019).

The fact that the weight for the pooled estimator is negative illustrates the fundamental difficulty of the two-way fixed effects estimator. In order to adjust for both unit-specific (but time-invariant) and time-specific (but unit-invariant) unobserved confounders, the two-way fixed effects regression estimator combines three estimators: (1) the unit fixed effects estimator, $\hat{\beta}_{\mathsf{FEunit}}$, which adjusts for unit-specific confounders by comparing the observations over time within each unit but fails to adjust for time-specific confounders, (2) the time fixed effects estimator, $\hat{\beta}_{\mathsf{FEtime}}$, which adjusts for time-specific confounders through the comparison of observations within each time period across unit, but is unable to adjust for unit-specific confounders, and (3) the pooled estimator, $\hat{\beta}_{\mathsf{pool}}$, which do not address any of these unobserved confounders. Since $\hat{\beta}_{\mathsf{pool}}$ has both types of biases, one may hope that subtracting it from the sum of the other two estimators cancels out both biases. In general, however, there is no guarantee that this combined estimator eliminates bias.

2.2 Causal Assumptions

Next, we consider a causal interpretation of the two-way fixed effects estimator presented above. First, the two-way fixed effects estimator given in equation (2) assumes the absense of spillover and carry over effects. That is, the outcome of one unit is affected neither by the treatment status of another unit nor by the treatment status of any unit (including itself) from the previous time periods.

Assumption 1 (Absense of Carryover and Spillover Effects) *The potential outcome of unit i at time t is a function of its own contemporaneous treatment status alone, i.e.,*

$$Y_{it}(\{\mathbf{X}_{t'}\}_{t'=1}^{t}) \ = \ Y_{it}(X_{it})$$

where $\mathbf{X}_t$ is a N dimensional vector of treatment status for all units at time t.

In many applications, this assumption may not be credible. For example, an economic shock in one country can directly influence the economy of another country in a future time period. The assumption can be partially relaxed by, for example, adjusting for one's own treatment history as well as the treatment histories of one's neighbors. However, since almost all existing applications of two-way fixed effects regression models require the absense of carryover and spillover effects, we maintain this assumption for the remainder of this paper and leave the challenge of addressing this issue to future research.

In the case of the standard two-way fixed effects model, researchers assume the following strict exogeneity conditional on the unobserved unit-specific and time-specific confounders,

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \alpha_i, \gamma_t) \ = \ 0. \tag{7}$$

We present the following nonparametric analogue of this assumption in the framework of potential outcomes. Unlike the model-based assumption given in equation (7), this formulation facilitates the understanding of the treatment assignment mechanism. Specifically, the treatment assignment for unit i at time t is assumed to be as-if randomized conditional on the unit-specific unobserved confounder $\mathbf{U}_i$, the time-specific confounders $\mathbf{V}_i$, and the past treatment history of the unit $\{X_{it'}\}_{t'=1}^{t-1}$. The assumption, which generalizes Assumption 3 of Imai and Kim (2019) to the two-way fixed effects estimator, is formally written as,

Assumption 2 (Sequential Ignorability with Time-invariant and Unit-invariant

Unobserved Confounders) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the following conditional independence holds,*

$$\begin{aligned}
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i1} \mid \mathbf{U}_i, \mathbf{V}_1 \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i2} \mid X_{i1}, \mathbf{U}_i, \mathbf{V}_2 \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i,T-1} \mid \{X_{it'}\}_{t'=1}^{T-2}, \mathbf{U}_i, \mathbf{V}_{T-1} \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{iT} \mid \{X_{it'}\}_{t'=1}^{T-1}, \mathbf{U}_i, \mathbf{V}_T
\end{aligned}$$

Finally, we also must assume the overlap condition,

Assumption 3 (Overlap) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the following inequalities hold,*

$$0 < \Pr(X_{it} = 1 \mid \{X_{it'}\}_{t'=1}^{t-1}, \mathbf{U}_i, \mathbf{V}_t) < 1.$$

Assumption 2 implies an important restriction that the treatment assignment cannot depend on the past outcomes. Thus, although the two-way fixed effects estimator adjusts for the two types of unobserved confounding under a parametric assumption, it rules out the dynamic causal relationship between the outcome and treatment variables. This fundamental tradeoff between unobservables and dynamics is the same as the one faced by the unit fixed effects regression estimator (Imai and Kim, 2019).

2.3 Causal Interpretation

Assumption 2 suggests that we must adjust for unobserved unit-specific and time-specific confounders at the same time. The two-way fixed effects estimator does this by simply including the unit and time fixed effects, i.e., α_i and γ_t, in an additive fashion. However, the assumption also implies that once we relax the linearity assumption, it is impossible to *nonparametrically* adjust for these two types of unobserved confounders at the same time. Using the matching framework, Imai and Kim (2019) shows that the unit-fixed effects estimator, i.e., $\hat{\beta}_{\mathsf{FEunit}}$ in equation (4), is based on the across-time comparison of treated and control observations within each unit. Similarly, the time-fixed effects estimator, i.e., $\hat{\beta}_{\mathsf{FEtime}}$ in equation (5), is based on the across-unit comparison of treated and control observations within each time period. Unfortunately, for any given observation, there is no other observation that belongs to the same unit and same time period. This implies that Assumption 2 cannot be satisfied without an additional parametric restriction.

We can further illustrate this point by expressing the two-way fixed effects estimator as a weighted average of several average treatment effect (ATE) estimators.

Proposition 2 (Two-way Fixed Effects Estimator as a Weighted Average of ATE Estimators) *When N and T are sufficiently large, the two-way fixed effects estimator defined in equation (2) can be approximated by a weighted average of the within-unit ATE estimator, the within-time ATE estimator, and pooled ATE estimator,*

$$\hat{\beta}_{\mathsf{FE2}} \approx \frac{\frac{1}{N}\sum_{i=1}^{N} S_i^2(\widehat{Y_i(1)} - \widehat{Y_i(0)}) + \frac{1}{T}\sum_{t=1}^{T} S_t^2(\widehat{Y_t(1)} - \widehat{Y_t(0)}) - \omega_{\mathsf{pool}}(\widehat{Y(1)} - \widehat{Y(0)})}{\omega_{\mathsf{FEunit}} + \omega_{\mathsf{FEtime}} - \omega_{\mathsf{pool}}}$$

where, for $x = 0, 1$,

$$\widehat{Y_i(x)} = \frac{\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\} Y_{it}}{\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\}}, \quad \widehat{Y_t(x)} = \frac{\sum_{i=1}^{N} \mathbf{1}\{X_{it} = x\} Y_{it}}{\sum_{i=1}^{N} \mathbf{1}\{X_{it} = x\}}, \quad \widehat{Y(x)} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\} Y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\}}.$$

Proof follows directly from Proposition 1 of Section 2.1 and Proposition 1 of Imai and Kim (2019) and hence is omitted. The weights, which are proportional to the variance of treatment variable for each comparison, sum to unity because $\omega_{\mathsf{FEunit}} = \sum_{i=1}^{N} S_i^2/N$ and $\omega_{\mathsf{FEtime}} = \sum_{t=1}^{T} S_t^2/T$. Moreover, the weight for the pooled estimator is negative. The proposition shows that the two-way fixed effects estimator relies upon the linearity assumption for reducing the bias due to the unit-specific unobserved confounding as well as the time-specific unobserved confounding. Doing so requires extrapolation, which is reflected by the negative weight given to the pooled ATE estimator.

As shown by different authors in their recent working papers, this problem of negative weights arises even when the two-way fixed effects estimator is used in much simpler settings. For example, Athey and Imbens (2018) studies the case of staggered adoption, in which all units are in the control condition at the beginning and some of them receive the treatment at different points in time. In this setting, once units are in the treatment condition, they are not allowed to revert to the control group. These authors show that the two-way fixed effects estimator can be written as a weighted average of several ATE estimators where some weights are negative. In the same setting, which is also called the event studies design, Abraham and Sun (2018) obtains different weighted average representations under additional assumptions but show that some weights can be negative (see also Borusyak and Jaravel, 2017). Finally, Chaisemartin and D'Haultfœuille (2018) derives yet another weighted-average representation of the two-way fixed effects estimator in related settings with different assumptions and studies the conditions under which some weights are negative. Proposition 2 above differs from these studies in that it is an algebraic result and does not impose a restriction on the distributions of treatment variable and potential outcomes.

3 The Proposed Methodology

In this section, we propose a multi-period difference-in-differences estimator based upon an alternative identification strategy (i.e., parallel trend assumption) rather than the strict exogeneity assumption described above. We show that this estimator is equivalent to the weighted two-way fixed effects estimator. This equivalence result leads to an efficient computation method, a model-based cluster robust standard error, and a specification test for the standard two-way fixed effects model.

3.1 The Parallel Trend Assumption

The discussion in Section 2 shows that the strict exogeneity assumption required for the two-way fixed effects estimator is problematic because it is impossible to nonparametrically adjust for unit-specific and time-specific unobserved confounders at the same time. To address this problem, we consider an alternative identification assumption based on the difference-in-differences (DiD) design (e.g., Abadie, 2005). Specifically, we assume that the average potential outcomes under the control condition have a parallel time trend for the treatment and control groups.

Assumption 4 (Parallel Trend) *For $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$,*

$$\mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = 1, X_{i,t-1} = 0) \quad = \quad \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = X_{i,t-1} = 0).$$

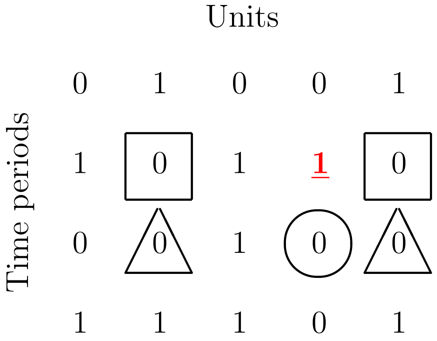


Figure 1: Illustration of how observations are used to estimate counterfactual outcomes for the DiD estimator (Proposition 1). The red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circle indicates the matched observation within the same unit, $Y_{i,t-1}$, whereas squares indicate those from the same time period, $\mathcal{N}_{it}$. Finally, triangles represent the set of observations that are used to make adjustment for unit and time effects, $\mathcal{A}_{it}$.

Under this DiD design, the estimand is the average treatment effect for the treated (ATT),

$$\tau \;=\; \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid X_{it} = 1, X_{i,t-1} = 0). \tag{8}$$

We emphasize that while Assumption 4 is expressed in terms of time trend of average potential outcomes, the DiD estimator still requires the absence of causal relationships between past outcomes and current treatment. For example, suppose that the treatment variable X_{it} has no causal effect on the outcome Y_{it}, but it is influenced by the previous outcome $Y_{i,t-1}$ such that $X_{it} = 1$ when $Y_{i,t-1}$ takes a value greater than its mean. Then, the regression toward the mean phenomenon suggests that the parallel trend assumption between the treatment and control groups is unlikely to hold, leading to a biased DiD estimate (see Allison, 1990, for details).

3.2 The Multi-period Difference-in-Differences Estimator

Next, we propose a multi-period DiD estimator that is consistent for the ATT under the parallel trend assumption and show that this estimator is equivalent to the weighted two-way fixed effects estimator. We use the matching framework of Imai and Kim (2019) and extend their equivalence results to the two-way fixed effects case. To formulate the proposed multi-period DiD estimator, we define three sets of observations as illustrated in Figure 1 — the within-unit matched set (represented by a circle), within-time matched set (represented by squares), and adjustment set (represented by triangles) — for a treated observation (i, t) (represented by the red underlined $\underline{1}$). As shown in Appendix B, this framework enables us to develop a general theory of two-way matching estimators, of which the proposed multi-period DiD estimator is a special case.

Formally, the within-unit matched set contains the observation of the same unit from the previous time period if it is under the control condition, and to be an empty set otherwise,

$$\mathcal{M}_{it} \quad = \quad \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 0\} \tag{9}$$

Similarly, the within-time matched set is defined as a group of control observations in the same time period whose prior observations are also under the control condition,

$$\mathcal{N}_{it} \quad = \quad \{(i', t') : i' \neq i, t' = t, X_{i't'} = X_{i',t'-1} = 0\} \tag{10}$$

Finally, we define the adjustment set $\mathcal{A}_{it}$, which contains the control observations in the previous period that share the same unit as those in $\mathcal{N}_{it}$,

$$\mathcal{A}_{it} \quad = \quad \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = X_{i't} = 0\} \tag{11}$$

Thus, the number of observations in this adjustment set is the same as that in $\mathcal{N}_{it}$.

Using these matched and adjustment sets, we can define the multi-period DiD estimator as the average of two-time-period two-group DiD estimators applied whenever there is a change from the control condition to the treatment condition,

$$\hat{\tau} \quad = \quad \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)} \right) \tag{12}$$

where $D_{i1} = 0$ for all i, $D_{it} = X_{it} \cdot \mathbf{1}\{|\mathcal{M}_{it}| \cdot |\mathcal{N}_{it}| > 0\}$ for $t > 1$, and for $D_{it} = 1$, we define,

$$\widehat{Y_{it}(x)} \quad = \quad \begin{cases} Y_{it} & \text{if } X_{it} = 1 \\ Y_{i,t-1} + \frac{1}{|\mathcal{N}_{it}|} \sum_{(i',t) \in \mathcal{N}_{it}} Y_{i't} - \frac{1}{|\mathcal{A}_{it}|} \sum_{(i',t') \in \mathcal{A}_{it}} Y_{i't'} & \text{if } X_{it} = 0 \end{cases} \tag{13}$$

When the treatment status of a unit changes from the control condition at time $t - 1$ to the treatment condition at time t (and there exists at least one unit i' whose treatment status does not change during the same time periods, i.e., $D_{it} = 1$), the counterfactual outcome for observation (i, t) is estimated by subtracting from Y_{it} its own observed outcome of the previous period $Y_{i,t-1}$ as well as the average outcome difference between the same two time periods among the other units whose treatment status remains unchanged as the control condition (see equation (13)).

3.3 Equivalence to the Weighted Two-way Fixed Effects Estimator

It is well known that the standard nonparametric DiD estimator is numerically equivalent to the two-way fixed effects regression estimator in the simplest setting, in which there are only two time periods and the treatment is administered to a group of units only in the second time period. Unfortunately, this equivalence result does not generalize to the current multi-period DiD design, in which the number of time periods may exceed two and each unit may switch in and out of the treatment condition multiple times. In a recent working paper, Goodman-Bacon (2018) shows that in the case of staggered adoption, the two-way fixed effects estimator is equal to the weighted average of various two-time-period two-group DiD estimators. As the author demonstrates, even in the staggered adoption setting, the two-way fixed effects estimator does not consistently estimate the ATT defined in equation (8) under the parallel trend assumption.

The main result of this paper shows that the general multi-period DiD estimator given in equation (12) is numerically equivalent to a weighted two-way fixed effects estimator.

Theorem 1 (Difference-in-Differences Estimator as a Weighted Two-way Fixed Effects Estimator) *Assume that there is at least one treated and control unit, i.e.,* $0 < \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it} < NT$*, and that there is at least one unit with* $D_{it} = 1$*, i.e.,* $0 < \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$*. The difference-in-differences estimator* $\hat{\tau}$*, defined in equation* (12)*, is equivalent to the following weighted two-way fixed effects regression estimator,*

$$\hat{\tau} \;=\; \hat{\beta}_{\mathsf{WFE2}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) - \beta(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)\}^2$$

where the asterisks indicate weighted averages, and the weights are given by,

$$W_{it} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot w_{it}^{i't'} \quad \text{and} \quad w_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t) = (i',t') \\ 1/|\mathcal{M}_{i't'}| & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ 1/|\mathcal{N}_{i't'}| & \text{if } (i,t) \in \mathcal{N}_{i't'} \\ (2X_{it}-1)(2X_{i't'}-1)/|\mathcal{A}_{i't'}| & \text{if } (i,t) \in \mathcal{A}_{i't'} \\ 0 & \text{otherwise.} \end{cases}$$

Proof is in Appendix C. Theorem 1 shows that the DiD estimator can be obtained by calculating the weighted linear two-way fixed effects regression estimator. Although some of the regression weights are negative, the fixed effects projection method can be applied on the complex plane in order to reduce the dimensionality, enabling an efficient computation of the DiD estimator (see Appendix D for details).

In the literature, Arkhangelsky *et al.* (2018) proposed a weighted two-way fixed effects estimator as a bias adjustment to the synthetic control method of Abadie *et al.* (2010) in the settings where a single unit is treated only once at the last time period. Their weights take a multiplicative form combining the unit specific and time specific weights that are designed to accurately predict the outcomes of the treated unit in previous time periods and the outcomes of the last time period for the control units, respectively. In contrast, our weights are constructed such that the resulting estimator is valid under the parallel trend assumption in general settings where units can move in and out of the treatment condition at different points in time.

3.4 Standard Error Calculation

Although weights are negative for some observations, we can view this multi-period DiD estimator as the method of moment estimator with the following moment condition,

$$\mathbb{E}\left\{W_{it}\widetilde{X}_{it}^{*}(\widetilde{Y}_{it}^{*} - \beta\widetilde{X}_{it}^{*})\right\} \;=\; 0$$

where $\widetilde{X}_{it}^{*} = X_{it} - \overline{X}_{i}^{*} - \overline{X}_{t}^{*} + \overline{X}^{*}$ and $\widetilde{Y}_{it}^{*} = Y_{it} - \overline{Y}_{i}^{*} - \overline{Y}_{t}^{*} + \overline{Y}^{*}$. Then, the standard asymptotic properties of the method of moments estimator apply directly to this case (Hansen, 1982). Specifically, let $\mathbf{W}_i$ be a $T \times T$ diagonal matrix whose (t,t) element is W_{it}. Then, the asymptotic conditional variance of $\hat{\tau}$ can be estimated using the standard cluster-robust sandwich formula with a multiplicative degrees of freedom adjustment as,

$$\begin{aligned}\widehat{\mathbb{V}(\hat{\beta}_{\mathsf{WFE2}} \mid \mathbf{X}, \mathbf{Y})} &= \frac{N^*(M^*-1)}{(N^*-1)(M^*-N^*-T^*-1)} \\ &\quad \times \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W}_i \widetilde{\mathbf{X}}_i^*\right)^{-1} \left\{\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W}_i \hat{\boldsymbol{\epsilon}}_i^* \hat{\boldsymbol{\epsilon}}_i^{*\top} \mathbf{W}_i \widetilde{\mathbf{X}}_i^*\right\} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W}_i \widetilde{\mathbf{X}}_i^*\right)^{-1} \quad (14)\end{aligned}$$

where $\widetilde{\mathbf{Y}}_i^*$ and $\widetilde{\mathbf{X}}_i^*$ represent T dimensional vectors of the outcome and treatment variables whose tth elements equal to $\widetilde{Y}_{it}^*$ and $\widetilde{X}_{it}^*$, respectively, $\hat{\boldsymbol{\epsilon}}_i^* = \widetilde{\mathbf{Y}}_i^* - \widetilde{\mathbf{X}}_i^* \hat{\beta}_{\mathsf{WFE2}}$ is the vector of residuals from the weighted regression for unit i, $M^* = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{1}\{W_{it} > 0\}$ is the total number of non-zero weight observations, $N^* = \sum_{i=1}^{N} \mathbf{1}\{\sum_{t=1}^{T} W_{it} > 0\}$ is the number of estimated unit fixed effects, and $T^* = \sum_{t=1}^{T} \mathbf{1}\{\sum_{t=1}^{T} W_{it} > 0\}$ is the number of estimated time fixed effects.

3.5 Specification Test

The equivalence between the multi-period DiD estimator and the weighted two-way fixed effects estimator implies a specification test for the standard two-way fixed effects estimator (White, 1980). This specification test, in which the null hypothesis is that the standard linear regression model with two-way fixed effects is correct, takes the following form,

$$\frac{(\hat{\beta}_{\mathsf{FE2}} - \hat{\beta}_{\mathsf{WFE2}})^2}{\widehat{\Phi}} \overset{\mathcal{D}}{\sim} \chi_1^2 \quad (15)$$

where

$$\begin{aligned}\widehat{\Phi} &= \widehat{\mathbb{V}(\hat{\beta}_{\mathsf{WFE2}} \mid \mathbf{X}, \mathbf{Y})} + \widehat{\mathbb{V}(\hat{\beta}_{\mathsf{FE2}} \mid \mathbf{X}, \mathbf{Y})} \\ &\quad - \frac{NT-1}{M^*-N^*-T^*-1} \left[\left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W} \widetilde{\mathbf{X}}_i^*\right)^{-1} \left\{\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W} \hat{\boldsymbol{\epsilon}}_i^* \hat{\boldsymbol{\epsilon}}_i^{\top} \widetilde{\mathbf{X}}_i\right\} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \widetilde{\mathbf{X}}_i\right)^{-1}\right. \\ &\quad \left. + \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \widetilde{\mathbf{X}}_i\right)^{-1} \left\{\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{*\top} \mathbf{W} \widetilde{\mathbf{X}}_i^*\right\} \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{*\top} \mathbf{W} \widetilde{\mathbf{X}}_i^*\right)^{-1}\right].\end{aligned}$$

This specification test can be used to examine the validity of parametric assumptions implied by the linear two-way fixed effects regression models.

3.6 Adjusting for Observed Time-varying Confounders

So far, we have not considered the existence of time-varying confounders. However, applied researchers often fit the following two-way fixed effects model,

$$Y_{it} \ = \ \alpha_i + \gamma_t + \beta X_{it} + \boldsymbol{\delta}^\top \mathbf{Z}_{it} + \epsilon_{it} \tag{16}$$

where $\mathbf{Z}_{it}$ represents the K-dimensional vector of time-varying covariates for unit i at time t and is causally prior to X_{it}. The corresponding strict exogeneity assumption is given by,

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \alpha_i, \gamma_t) \ = \ 0 \tag{17}$$

where $\mathbf{Z}_i$ is a $T \times K$ matrix whose tth row is $\mathbf{Z}_{it}^\top$. The least squares estimator can be computed in a similar way by transforming $\mathbf{Z}_{it}$ as well as Y_{it} and X_{it}.

In terms of causal assumptions, we maintain Assumption 1 while Assumptions 2 and 3 are generalized to the following,

Assumption 5 (Sequential Ignorability with Time-varying Observed Confounders as well as Time-invariant and Unit-invariant Unobserved Confounders) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the following conditional independence holds,*

$$\begin{aligned}
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T \ &\perp\!\!\!\perp \ X_{i1} \mid \mathbf{Z}_{i1}, \mathbf{U}_i, \mathbf{V}_1 \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T \ &\perp\!\!\!\perp \ X_{i2} \mid X_{i1}, \mathbf{Z}_{i2}, \mathbf{Z}_{i1}, \mathbf{U}_i, \mathbf{V}_2 \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T \ &\perp\!\!\!\perp \ X_{i,T-1} \mid \{X_{it'}\}_{t'=1}^{T-2}, \{\mathbf{Z}_{it'}\}_{t'=1}^{T-1}, \mathbf{U}_i, \mathbf{V}_{T-1} \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T \ &\perp\!\!\!\perp \ X_{iT} \mid \{X_{it'}\}_{t'=1}^{T-1}, \{\mathbf{Z}_{it'}\}_{t'=1}^{T}, \mathbf{U}_i, \mathbf{V}_T
\end{aligned}$$

Assumption 6 (Overlap with Observed Time-varying Confounders) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the following inequalities hold,*

$$0 \ < \ \Pr(X_{it} = 1 \mid \{X_{it'}\}_{t'=1}^{t-1}, \{\mathbf{Z}_{it'}\}_{t'=1}^{t}, \mathbf{U}_i, \mathbf{V}_t) \ < \ 1.$$

These assumptions, which include time-varying confounders in the conditioning set, do not alter the fundamental limitations of the two-way fixed effects model: (1) the past outcomes cannot affect the treatment (directly or indirectly through $\mathbf{Z}_{it}$), and (2) it is impossible to nonparametrically adjust for $\mathbf{U}_i$ and $\mathbf{V}_i$ at the same time.

To address this problem, we generalize Assumption 4 by assuming the parallel trend conditional on the observed time-varying confounders,

Assumption 7 (Parallel Trend with Observed Time-varying Confounders) *For* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

$$\mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = 1, X_{i,t-1} = 0, \mathbf{Z}_{it}) \ = \ \mathbb{E}(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = X_{i,t-1} = 0, \mathbf{Z}_{it}).$$

This assumption then leads to the multi-period DiD estimator of equation (12), in which the counterfactual outcomes are estimated conditional on time-varying confounders. Imai *et al.* (2018) uses matching and weighting methods for the estimation under a similar assumption.

Here, we consider an alternative model-based approach by exploiting the fact that the proposed DiD estimator is equivalent to the weighted two-way fixed effects estimator as shown in Section 3.3. This estimator is given by,

$$(\hat{\beta}_{\mathsf{WFE2}}, \hat{\boldsymbol{\delta}}_{\mathsf{WFE2}}) \ = \ \underset{(\beta,\boldsymbol{\delta})}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(\widetilde{Y}^*_{it} - \beta\widetilde{X}^*_{it} - \boldsymbol{\delta}^\top\widetilde{\mathbf{Z}}^*_{it})^2 \tag{18}$$

where $\widetilde{\mathbf{Z}}^*_{it} = \mathbf{Z}_{it} - \overline{\mathbf{Z}}^*_i - \overline{\mathbf{Z}}^*_t + \overline{\mathbf{Z}}^*$ with $\overline{\mathbf{Z}}^*_i = \sum_{t=1}^T W_{it}\mathbf{Z}_{it} / \sum_{t=1}^T W_{it}$, $\overline{\mathbf{Z}}^*_t = \sum_{i=1}^N W_{it}\mathbf{Z}_{it} / \sum_{i=1}^N W_{it}$, and $\overline{\mathbf{Z}}^* = \sum_{i=1}^N \sum_{t=1}^T W_{it}\mathbf{Z}_{it} / \sum_{i=1}^N \sum_{t=1}^T W_{it}$. This estimator can be interpreted as the multi-period DiD estimator with a model-based adjustment for observed time-varying confounders. Specifically, equation (13) is modified as,

$$\widehat{Y_{it}(x)} = \begin{cases} Y_{it} - \hat{\boldsymbol{\delta}}^\top_{\mathsf{WFE2}}\mathbf{Z}_{it} & \text{if } X_{it} = 1 \\ (Y_{i,t-1} - \hat{\boldsymbol{\delta}}^\top_{\mathsf{WFE2}} Z_{i,t-1}) + \frac{1}{|\mathcal{N}_{it}|}\sum_{(i',t)\in\mathcal{N}_{it}}(Y_{i't} - \hat{\boldsymbol{\delta}}^\top_{\mathsf{WFE2}}\mathbf{Z}_{i't}) - \frac{1}{|\mathcal{A}_{it}|}\sum_{(i',t')\in\mathcal{A}_{it}}(Y_{i't'} - \hat{\boldsymbol{\delta}}^\top_{\mathsf{WFE2}}\mathbf{Z}_{i't'}) & \text{if } X_{it} = 0 \end{cases} \tag{19}$$

That is, the outcome variable is adjusted using the observed time-varying confounders. Under this formulation, all the results regarding the standard error calculation and specification test discussed in Sections 3.4 and 3.5 are directly applicable by including $\mathbf{Z}_{it}$ as additional predictors and making appropriate changes to the degrees of freedom.

4 An Empirical Illustration

In this section, we illustrate the proposed methodology by applying it to the controversy regarding the effects of the General Agreement on Tariffs and Trade (GATT), which was succeeded by the World Trade Organization (WTO) in 1995, on bilateral trade. Although GATT/WTO has been touted as one of the most successful international agreements in reducing barriers to global trade, Rose (2004) finds little evidence that countries' formal membership in the institution increases trade. In contrast, Tomz *et al.* (2007) found the effect to be positive and statistically significant so long as the definition of membership includes *nonmember participants* such as former colonies who benefit from essentially identical benefits and are required to meet similar obligations (see Rose, 2007, for a rebuttal). While they disagree on empirical findings, all researchers embrace the use of the linear fixed effects estimators.

4.1 Data and Models

Here, we compare the results of the standard two-way fixed effects estimators with those of the multi-period DiD estimators implemented as the weighted two-way fixed effects estimators. We analyze the dyadic data set of annual trade volume, which extends across 163 countries and 47 years. We restrict our analysis to the GATT period (1948–1994) because the WTO relies on different institutional rules and institutions such as its dispute settlement system; our study is based on 196,207 dyad-year observations.[1] To facilitate the comparison of our empirical findings against existing studies, we use the two different definitions of GATT/WTO memberships: One is formal memberships used by Rose (2004), and the other is participants as defined by Tomz *et al.* (2007).

We begin by fitting the standard two-way fixed effects model as done by Tomz *et al.* (2007)

[1] We follow the literature to focus our analysis on the dyads with positive trade volume. See, Kim *et al.* (2019) for the importance of including dyads with zero trade.

and Rose (2007),

$$\log Y_{it} \quad = \quad \alpha_i + \gamma_t + \beta X_{it} + \delta^\top \mathbf{Z}_{it} + \epsilon_{it}, \qquad (20)$$

where Y_{it} is the bilateral trade volume for dyad i in year t, α_i and γ_t represent dyad and year fixed effects, respectively, and X_{it} indicates whether both countries of the dyad are GATT/WTO members. Finally, $\mathbf{Z}_{it}$ represents a vector of observed time-varying confounders including the presence of GSP (Generalized System of Preferences) preferential rates, log product real GDP, log product real GDP per capita, joint memberships in regional FTA (Free Trade Agreement) and currency union, and an indicator of current colonial relationship between the countries in the dyad. Throughout our analyses, we use the cluster-robust standard error that allows for arbitrary autocorrelation as well as heteroskedasticity as defined in equations (3) and (14).

4.2 Findings

Table 1 summarizes the results. Consistent with the findings from Rose (2004) and Tomz *et al.* (2007), we find conflicting evidence for the effects of GATT on bilateral trade from the standard two-way fixed effects model depending on the definition of membership. Specifically, the effect of formal membership is statistically indistinguishable from zero while it becomes positive and statistically significant when we consider nonmember participants as members.[2] The estimated effect sizes are also substantively different. For example, the models with time-varying covariates show that pairs of two participants are likely to trade 25% ($\approx \exp(0.227) - 1$) more than dyads with at least one non-participant whereas the analogous estimate for the formal membership is only 3.6%($\approx \exp(0.036) - 1$).

Next, we fit the proposed multi-period DiD estimator. Table 1 shows that under the parallel trend assumption we find little effect of the GATT on trade regardless of the inclusion of time-

[2] Note that we find a significant effect of formal membership when we do not include time-varying covariates (see the second column). However, the null hypothesis that this specification is correct is rejected via the specification test (p-value $= 0.025$).

	Formal Membership				Participants			
	Standard		Diff-in-Diffs		Standard		Diff-in-Diffs	
Estimate	0.036	0.098	0.003	0.019	0.227	0.320	0.043	0.010
(SE)	(0.025)	(0.028)	(0.040)	(0.033)	(0.031)	(0.034)	(0.047)	(0.029)
Specification test			8.229	5.032			37.703	118.252
(p-value)			0.313	0.025			0.000	0.000
N	196,207	196,207	92,043	92,043	196,207	196,207	59,732	59,732
Covariates	✓		✓		✓		✓	

Table 1: **Estimated Effects of GATT/WTO Memberships on the Logarithm of Bilateral Trade**. The "Standard" column presents the estimates based on the standard linear regression model with dyadic and time fixed effects. The "Diff-in-Diffs" column presents the estimates based on the multi-period difference-in-differences estimator. We also present the test statistic for the specification test along with its p-value. These results compare the dyads of two GATT/WTO members with those consisting of either one or no GATT/WTO member. The year-varying dyadic covariates include GSP (Generalized System of Preferences), log product real GDP, log product real GDP per capita, regional FTA (Free Trade Agreement), currency union, and currently colonized. We also report the results without these covariates (see the columns without ✓). Cluster-robust standard errors are in parentheses. The results suggest that different causal assumptions, which imply different regression weights, can yield different results, and that the standard linear fixed effects models are likely to be misspecified.

varying covariates and under different membership measures. The estimated effect sizes are also much smaller and less variable than the estimates from the standard two-way fixed effects model, ranging from 0.003% to at most 0.043% increase in trade. Indeed, the specification test shows that we reject the null hypothesis that the standard linear regression model with two-way fixed effects is correct for all cases when the estimated effect is found to be positive and statistically significant under the standard model.

Finally, the equivalence between the multi-period DiD estimator and the weighted two-way fixed effects estimator enables researchers to examine the weights used for each observation in estimation. Figure 2 shows the distribution of non-zero weights across the two different membership measures. First, many observations receive zero weights as the treatment status tends not to change over time. For example, the U.S.-U.K. dyad gets zero weight across the entire period. In this regard, more observations will get zero weights for the participant $(136, 475)$ than formal

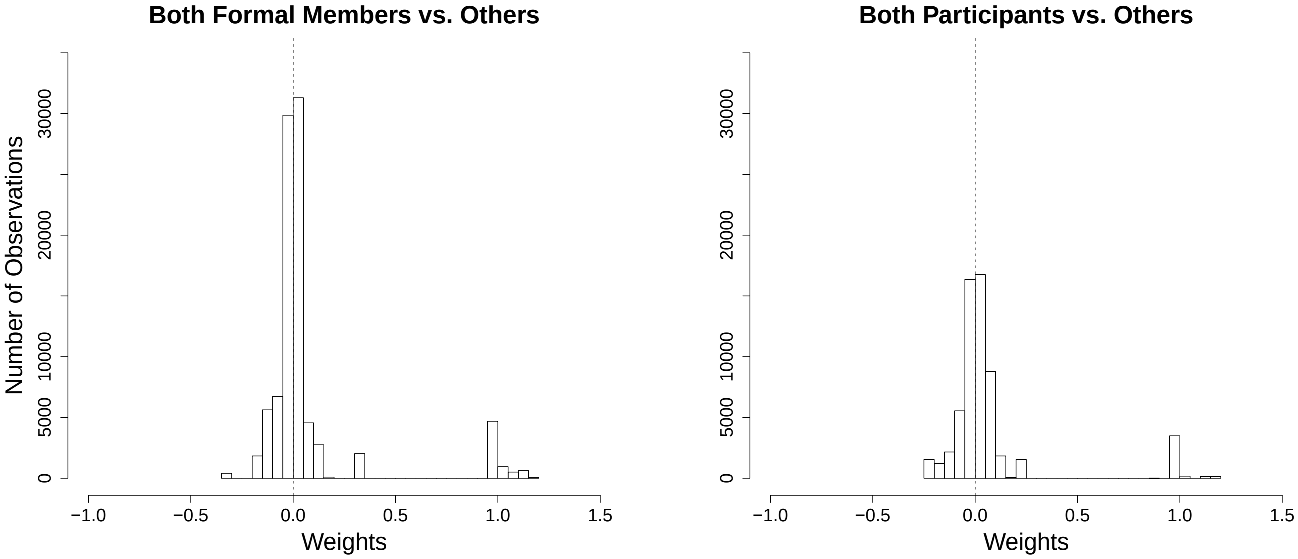


Figure 2: **Distribution of Weights:** This figure shows the distribution W_{it}. The observations with negative and positive weights are on the left- and right-hand side of the dotted vertical line at zero. There are 104,164 (136,475) observations with zero weights when we analyze formal membership (nonmember participants.

membership $(104, 164)$ since dyads are under the treatment condition for longer periods when the broader definition of membership is used. Second, we observe an approximately symmetric distribution of positive and negative weights. This is because the weights assigned to the dyads with a change in the treatment status will be similar across the two consecutive periods with different signs (see Theorem 1 and notice that $|\mathcal{N}_{i't'}|$ and $|\mathcal{A}_{i't'}|$ are the same in DiD for a given treated observation (i', t')). Finally, the dyads that change the treatment status (e.g., U.S.-Austria in 1951) will get a weight of 1 at the time of treatment while they receive a weight of 1 ($\mathcal{A}_{i't'}$ as its own within-unit control) in addition to the weights that they accrue as a control ($\mathcal{N}_{j,t'-1}$) for other treated dyads j in the previous year (e.g., U.S.-Italy in 1950). This explains the cluster of a few observations at 1 in Figure 2. Overall, the proposed estimator provides a robust estimate of the causal quantity of interest and additional information about the contribution of each observation to the estimation of causal effects.

5 Concluding Remarks

In this paper, we study the use of linear regression models with unit and time fixed effects for causal inference with panel data. Although these models have been used extensively in applied research, little has been understood about how these models can be used to identify causal effects. A number of researchers have recently tackled this question (e.g., Borusyak and Jaravel, 2017; Abraham and Sun, 2018; Athey and Imbens, 2018; Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018), and we contribute to this growing body of methodological research.

Many extensions of two-way fixed effects estimators beyond connecting them to the multi-period difference-in-differences estimator as done in this paper are possible. For example, Imai *et al.* (2018) develop matching and weighting methods for causal inference with panel data. They generalize the equivalence result of this paper and show that the proposed matching and weighting estimators are equivalent to the weighted two-way fixed effects estimators. Their result represents a nonparametric generalization of covariate adjustment for time-varying confounders described in Section 3.6 of this paper.

References

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

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* **105**, 490, 493–505.

Abraham, S. and Sun, L. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Tech. rep., Department of Economics, Massachusetts Institute of Technology.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology* **20**, 93–114.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* **49**, 4, 431–434.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2018). Synthetic difference in differences. Tech. rep., Stanford Graduate School of Business.

Aronow, P. M. and Samii, C. (2015). Does regression produce representative estimates of causal effects? *American Journal of Political Science* **60**, 1, 250–267.

Athey, S. and Imbens, G. (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Tech. rep., Stanford Graduate School of Business, `https://arxiv.org/abs/1808.05293`.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* **119**, 1, 249–275.

Borusyak, K. and Jaravel, X. (2017). Revisiting event study designs, with an application to the estimation of the marginal propensity to consume. Tech. rep., Department of Economics, Harvard University.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources* **50**, 2, 317–372.

Chaisemartin, C. d. and D'Haultfœuille, X. (2018). Two-way fixed effects estimators with heterogeneous treatment effects. Tech. rep., Department of Economics, University of California, Santa Barbara, `https://arxiv.org/abs/1803.08807`.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica* **81**, 2, 535–580.

Davis, P. (2002). Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics* **106**, 67–95.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Working Paper 25018, National Bureau of Economic Research.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics* **141**, 597–620.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 4, 1029–1054.

Humphreys, M. (2009). Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Tech. rep., Department of Political Science, Columbia University. `http://www.columbia.edu/~mh2245/papers1/monotonicity7.pdf`.

Imai, K. and Kim, I. S. (2011). On the use of linear fixed effects regression models for causal inference. Tech. rep., Princeton University.

Imai, K. and Kim, I. S. (2019). When should we use linear unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science* Forthcoming.

Imai, K., Kim, I. S., and Wang, E. (2018). Matching methods for time-series cross-section data. Working paper available at `https://imai.fas.harvard.edu/research/tscs.html`.

Kim, I. S., Londregan, J., and Ratkovic, M. (2019). The effects of political institutions on the extensive and intensive margins of trade. *International Organization* Forthcoming.

Rose, A. K. (2004). Do we really know that the WTO increases trade? *The American Economic Review* **94**, 1, 98–114.

Rose, A. K. (2007). Do we really know that the WTO increases trade? Reply. *The American Economic Review* **97**, 5, 2019–2025.

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources* **50**, 2, 301–316.

Tomz, M., Goldstein, J. L., and Rivers, D. (2007). Do we really know that the WTO increases trade? Comment. *The American Economic Review* **97**, 5, 2005–2018.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review* **21**, 1, 149–170.

A Proof of Proposition 1

Define:

$$\begin{aligned}
\overline{Y}_{1i} &= \frac{\sum_{t=1}^{T} X_{it} Y_{it}}{\sum_{t=1}^{T} X_{it}}, \quad \overline{Y}_{0i} = \frac{\sum_{t=1}^{T}(1-X_{it}) Y_{it}}{\sum_{t=1}^{T}(1-X_{it})}, \quad \overline{Y}_{1t} = \frac{\sum_{i=1}^{N} X_{it} Y_{it}}{\sum_{i=1}^{N} X_{it}}, \\
\overline{Y}_{0t} &= \frac{\sum_{i=1}^{N}(1-X_{it}) Y_{it}}{\sum_{i=1}^{N}(1-X_{it})}, \quad \overline{Y}_{1} = \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} X_{it} Y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}}, \quad \overline{Y}_{0} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(1-X_{it}) Y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}(1-X_{it})}.
\end{aligned}$$

Then, we have,

$$\begin{aligned}
\hat{\beta}_{\textsf{FE2}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\{(X_{it}-\overline{X}_i)(Y_{it}-\overline{Y}_i)+(X_{it}-\overline{X}_t)(Y_{it}-\overline{Y}_t)-(X_{it}-\overline{X})(Y_{it}-\overline{Y})\}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\{(X_{it}-\overline{X}_i)^2+(X_{it}-\overline{X}_t)^2-(X_{it}-\overline{X})^2\}} \\
&= \frac{(T-1)\sum_{i=1}^{N} S_i^2(\overline{Y}_{1i}-\overline{Y}_{0i})+(N-1)\sum_{t=1}^{T} S_t^2(\overline{Y}_{1t}-\overline{Y}_{0t})-(NT-1)S^2(\overline{Y}_1-\overline{Y}_0)}{(T-1)\sum_{i=1}^{N} S_i^2+(N-1)\sum_{t=1}^{T} S_t^2-(N-1)(T-1)S^2} \\
&= \frac{\omega^*_{\textsf{FEunit}} \times \hat{\beta}_{\textsf{FEunit}} + \omega^*_{\textsf{FEtime}} \times \hat{\beta}_{\textsf{FEtime}} - \omega^*_{\textsf{pool}} \times \hat{\beta}_{\textsf{pool}}}{w^*_{\textsf{FE}} + w^*_{\textsf{FEtime}} - w^*_{\textsf{pool}}}
\end{aligned}$$

where

$$\begin{aligned}
\omega^*_{\textsf{FEunit}} &= \frac{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{NT}\right)S^2} \\
\omega^*_{\textsf{FEtime}} &= \frac{\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{NT}\right)S^2} \\
\omega^*_{\textsf{pool}} &= \frac{\left(1-\frac{1}{NT}\right)S^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{NT}\right)S^2}
\end{aligned}$$

Assuming that N and T are sufficiently large, the desired results follow. □

B The Two-way Matching Estimators

In this appendix, in order to prove Theorem 1, we build a more general theory of two-way matching estimators, building on the one-way matching estimators of Imai and Kim (2019).

B.1 The Two-way Fixed Effects Estimator as a Matching Estimator

First, we directly connect the two-way fixed effects estimator in equation (1) to the two-way matching estimators. The discussion in Section 2 shows that the two-way fixed effects estimator is unable to nonparametrically adjust for unit-specific and time-specific unobserved confounders at the same time since no other observation shares the same unit and time indices. We show that this difficulty yields a matched set that contains some observations of the same treatment status. The following proposition shows that the two-way fixed effects estimator can be written as a matching estimator with a correction for such "mismatches."

Proposition 3 (The Two-way Fixed Effects Estimator as a Two-Way Matching Estimator) *The two-way fixed effects estimator defined in equation* (2) *is equivalent to the following* *matching estimator,*

$$\hat{\beta}_{\mathsf{FE2}} \quad = \quad \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right)\right\}$$

where for $x = 0, 1$,

$$\begin{aligned}
\widehat{Y_{it}(x)} &= \left\{\begin{array}{ll} Y_{it} & \text{if } X_{it}=x \\ \frac{1}{T-1}\sum_{t'\neq t}Y_{it'}+\frac{1}{N-1}\sum_{i'\neq i}Y_{i't}-\frac{1}{(T-1)(N-1)}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'} & \text{if } X_{it}=1-x \end{array}\right. \\
K &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(\frac{\sum_{t'\neq t}(1-X_{it'})}{T-1}+\frac{\sum_{i'\neq i}(1-X_{i't})}{N-1}-\frac{\sum_{i'\neq i}\sum_{t'\neq t}(1-X_{i't'})}{(T-1)(N-1)}\right)\right. \\
&\qquad \left. +\,(1-X_{it})\left(\frac{\sum_{t'\neq t}X_{it'}}{T-1}+\frac{\sum_{i'\neq i}X_{i't}}{N-1}-\frac{\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}}{(T-1)(N-1)}\right)\right\}.
\end{aligned}$$

Proof We begin by establishing two algebraic equalities. First, we prove the following equality,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})-(1-X_{it})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})\right\} \\
= &\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}-\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right.\right. \\
&\qquad\qquad\qquad \left. -\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}-\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\} \\
&\quad -(1-X_{it})\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}-\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right. \\
&\qquad\qquad\qquad \left.\left. -\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}-\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right] \\
= &\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right. \\
&\qquad \left. -(1-X_{it})\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right] \\
= &\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{\sum_{t'=1}^{T}Y_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}Y_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}}{(T-1)(N-1)}\right)\right. \\
&\qquad \left. -\,(1-X_{it})\left(\frac{\sum_{t'=1}^{T}Y_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}Y_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}-Y_{it}}{(T-1)(N-1)}\right)\right\}. \\
= &\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right) \qquad (21)
\end{aligned}$$

The second algebraic equality we prove is the following,

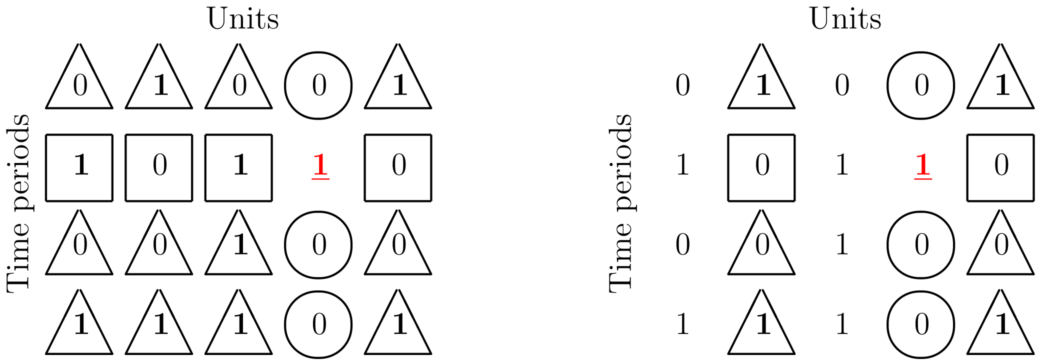
$$
\begin{aligned}
& \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})-(1-X_{it})(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})\right\} \\
= \; & \sum_{i=1}^{N}\sum_{t=1}^{T}\Bigg[X_{it}\Bigg\{X_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\Bigg(\frac{1}{T}\sum_{t'\neq t}X_{it'}-\frac{1}{NT}\sum_{t'\neq t}X_{it'}\Bigg) \\
& \qquad -\Bigg(\frac{1}{N}\sum_{i'\neq i}X_{i't}-\frac{1}{NT}\sum_{i'\neq i}X_{i't}\Bigg)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\Bigg\} \\
& \quad -(1-X_{it})\Bigg\{X_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\Bigg(\frac{1}{T}\sum_{t'\neq t}X_{it'}-\frac{1}{NT}\sum_{t'\neq t}X_{it'}\Bigg) \\
& \qquad -\Bigg(\frac{1}{N}\sum_{i'\neq i}X_{i't}-\frac{1}{NT}\sum_{i'\neq i}X_{i't}\Bigg)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\Bigg\}\Bigg] \\
= \; & \sum_{i=1}^{N}\sum_{t=1}^{T}\Bigg[X_{it}\Bigg\{\frac{(N-1)(T-1)}{NT}X_{it}-\frac{N-1}{NT}\sum_{t'\neq t}X_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\Bigg\} \\
& \quad -(1-X_{it})\Bigg\{\frac{(N-1)(T-1)}{NT}X_{it}-\frac{N-1}{NT}\sum_{t'\neq t}X_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\Bigg\}\Bigg] \\
= \; & \frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Bigg[\Bigg\{X_{it}\Bigg(\frac{\sum_{t'=1}^{T}(1-X_{it'})}{T-1}+\frac{\sum_{i'=1}^{N}(1-X_{i't})}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}(1-X_{i't'})}{(T-1)(N-1)}\Bigg) \\
& \quad +(1-X_{it})\Bigg(\frac{\sum_{t'=1}^{T}X_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}X_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}X_{i't'}}{(T-1)(N-1)}\Bigg)\Bigg\}\Bigg] \\
= \; & K(T-1)(N-1) \qquad (22)
\end{aligned}
$$

Finally, using the above algebraic equalities, we can derive the desired result as follows,

$$
\begin{aligned}
\hat{\beta}_{\mathsf{FE2}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}Y_{it}-T\sum_{i=1}^{N}\overline{X}_i\overline{Y}_i-N\sum_{t=1}^{T}\overline{X}_t\overline{Y}_t+NT\overline{X}\overline{Y}}{NT\overline{X}-T\sum_{i=1}^{N}\overline{X}_i^2-N\sum_{t=1}T\overline{X}_t^2+NT\overline{X}^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(2X_{it}-1)(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(2X_{it}-1)(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})-(1-X_{it})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})-(1-X_{it})(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})\right\}} \\
&= \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right)\right\}
\end{aligned}
$$

where the last equality follows from equation (21) and (22). □

The proposition shows that the estimated counterfactual outcome of a given unit is a function of three averages. First, the average of all the other observations from the same unit and the average of



(a) Two-way fixed effects estimator

(b) Two-way matching estimator

Figure 3: An Example of the Binary Treatment Matrix with Five Units and Four Time Periods. Panels (a) and (b) illustrate how observations are used to estimate counterfactual outcomes for the two-way fixed effects estimator (Proposition 3) and the adjusted matching estimator (Proposition 4), respectively. In the figures, the red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circles indicate the set of matched observations that are from the same unit, $\mathcal{M}_{it}^{\mathsf{FE2}}$ for Panel (a) and $\mathcal{M}_{it}^{\mathsf{match}}$ for Panel (b), whereas squares indicate those from the same time period, $\mathcal{N}_{it}^{\mathsf{FE2}}$ for Panel (a) and $\mathcal{N}_{it}^{\mathsf{match}}$ for Panel (b). Finally, triangles represent the set of observations that are used to make adjustment for unit and time effects, $\mathcal{A}_{it}$. The set includes the observations of the same treatment status in both cases (bold **1** entries in triangles), leading to an adjustment in the matching estimator.

all the other observations from the same time period are added together. We call them the *within-unit matched set* and the *within-time matched set*, respectively. In the case of the linear two-way fixed effects estimator, these matched sets are defined as,

$$\mathcal{M}_{it}^{\mathsf{FE2}} = \{(i', t') : i' = i, t' \neq t\} \tag{23}$$

$$\mathcal{N}_{it}^{\mathsf{FE2}} = \{(i', t') : i' \neq i, t' = t\} \tag{24}$$

Finally, to adjust for unit and time fixed effects, we use observations that share the same unit or time as those in $\mathcal{N}_{it}^{\mathsf{FE2}}$ and $\mathcal{M}_{it}^{\mathsf{FE2}}$, respectively, and subtract their mean from this sum. We use $\mathcal{A}_{it}^{\mathsf{FE2}}$ to denote this group of observations and call it the *adjustment set* for observation (i, t) with the following definition,

$$\mathcal{A}_{it}^{\mathsf{FE2}} = \{(i', t') : i' \neq i, t' \neq t, (i, t') \in \mathcal{M}_{it}^{\mathsf{FE2}}, (i', t) \in \mathcal{N}_{it}^{\mathsf{FE2}}\} \tag{25}$$

Therefore, by construction, the number of observations in $\mathcal{A}_{it}^{\mathsf{FE2}}$ equals the product of the number of observations in the within-unit and within-time matched sets, i.e., $\#\mathcal{A}_{it}^{\mathsf{FE2}} = \#\mathcal{M}_{it}^{\mathsf{FE2}} \cdot \#\mathcal{N}_{it}^{\mathsf{FE2}}$.

Panel (a) of Figure 3 presents an example of the binary treatment matrix with five units and four time periods, i.e., $N = 5$ and $T = 4$. In the figure, the red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated using other observations. This counterfactual quantity is estimated as the average of control observations from the same unit $\mathcal{M}_{it}^{\mathsf{FE2}}$ (circles in the figure), plus the average of control observations from the same time period $\mathcal{N}_{it}^{\mathsf{FE2}}$ (squares), minus the average of adjustment observations, $\mathcal{A}_{it}^{\mathsf{FE2}}$ defined as $\{(i', t') : i' \neq i, t' \neq t, (i, t') \in \mathcal{M}_{it}^{\mathsf{FE2}}, (i', t) \in \mathcal{N}_{it}^{\mathsf{FE2}}\}$ (triangles).

Since all of these three averages include units of the same treatment status, the two-way fixed effects estimator adjusts for the attenuation bias due to these "mismatches." This is done via the factor K, which is equal to the net proportion of proper matches between the observations of opposite treatment status. The nonparametric matching representation given in Proposition 3 identifies the exact information used by the two-way fixed effects estimator to estimate counterfactual outcomes. Specifically, to estimate the counterfactual outcome of each unit, all the other observations are used, including the observations of the same treatment status from different units and different years. This makes the causal interpretation of the standard two-way fixed effects estimator difficult.

B.2 The Adjusted Two-way Matching Estimator

Given this result, we improve the two-way fixed effects estimator by only matching each observation with other observations of the *opposite* treatment status to estimate the counterfactual outcome. That is, we use the following *within-unit matched set* $\mathcal{M}_{it}^{\mathsf{match}}$, which consists of the observations within the same unit but with the opposite treatment status,

$$\mathcal{M}_{it}^{\mathsf{match}} \;=\; \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}\}. \tag{26}$$

Similarly, we restrict the *within-time matched set* $\mathcal{N}_{it}$ so that its observations belong to the same time period t but have the opposite treatment status,

$$\mathcal{N}_{it}^{\mathsf{match}} \;=\; \{(i', t') : t' = t, X_{i't'} = 1 - X_{it}\}. \tag{27}$$

Then, using equation (25), we can define the corresponding adjustment set $\mathcal{A}_{it}^{\mathsf{match}}$. Unlike the one-way case studied in Imai and Kim (2019), however, we cannot eliminate mismatches in $\mathcal{A}_{it}^{\mathsf{match}}$ without additional restrictions on the matched sets, $\mathcal{M}_{it}^{\mathsf{match}}$ and $\mathcal{N}_{it}^{\mathsf{match}}$ (see Section 3.1). This point is illustrated by Panel (b) of Figure 3 where the adjustment set $\mathcal{A}_{it}^{\mathsf{match}}$ (triangles) includes the observations of the same treatment status.

The next proposition establishes that *any* adjusted two-way matching estimator that eliminates mismatches within-unit and within-time dimension can be written as a weighted linear regression estimator with unit and time fixed effects.

Proposition 4 (The Adjusted Two-way Matching Estimator as a Weighted Two-way Fixed Effects Regression Estimator) *Assume that the treatment varies within each unit as well as within each time period, i.e.,* $0 < \sum_{t=1}^{T} X_{it} < T$ *for each* i *and* $0 < \sum_{i=1}^{N} X_{it} < N$ *for each* t. *Consider the following adjusted matching estimator,*

$$\hat{\tau}_{\mathsf{match2}} \;=\; \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{D_{it}}{K_{it}} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where $D_{it} = \mathbf{1}\{\#\mathcal{M}_{it}^{match} \cdot \#\mathcal{N}_{it}^{match} > 0\}$, *and for* $x = 0, 1$,

$$\widehat{Y_{it}(x)} \;=\; \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{\#\mathcal{M}_{it}^{match}} \sum_{(i,t') \in \mathcal{M}_{it}^{match}} Y_{it'} + \frac{1}{\#\mathcal{N}_{it}^{match}} \sum_{(i',t) \in \mathcal{N}_{it}^{match}} Y_{i't} - \frac{1}{\#\mathcal{A}_{it}^{match}} \sum_{(i',t') \in \mathcal{A}_{it}^{match}} Y_{i't'} & \text{if } X_{it} = 1 - x \end{cases}$$

$$K_{it} \;=\; \frac{\#\mathcal{A}_{it}^{match} + a_{it}}{\#\mathcal{A}_{it}^{match}}$$

and $a_{it} = \#\{(i', t') \in \mathcal{A}_{it}^{match} : X_{i't'} = X_{it}\}$. *Then, this adjusted matching estimator is equivalent to the following weighted two-way fixed effects estimator,*

$$\hat{\beta}_{\mathsf{WFE2}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) - \beta(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)\}^2$$

where the asterisks indicate weighted averages, $\overline{Y}_i^* = \sum_{t=1}^T W_{it}Y_{it}/\sum_{t=1}^T W_{it}$*,* $\overline{Y}_t^* = \sum_{i=1}^N W_{it}Y_{it}/\sum_{i=1}^N W_{it}$*,* $\overline{X}_i^* = \sum_{t=1}^T W_{it}X_{it}/\sum_{t=1}^T W_{it}$*,* $\overline{X}_t^* = \sum_{i=1}^N W_{it}X_{it}/\sum_{i=1}^N W_{it}$*,* $\overline{Y}^* = \sum_{i=1}^N\sum_{t=1}^T W_{it}Y_{it}/\sum_{i=1}^N\sum_{t=1}^T W_{it}$*,* $\overline{X}^* = \sum_{i=1}^N\sum_{t=1}^T W_{it}X_{it}/\sum_{i=1}^N\sum_{t=1}^T W_{it}$*, and*

$$W_{it} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{and} \quad w_{it}^{i't'} \;=\; \begin{cases} \frac{D_{i't'}}{K_{i't'}} & \text{if } (i,t) = (i',t') \\ \frac{D_{i't'}}{K_{i't'}\cdot \#\mathcal{M}_{i't'}^{\mathit{match}}} & \text{if } (i,t) \in \mathcal{M}_{i't'}^{\mathit{match}} \\ \frac{D_{i't'}}{K_{i't'}\cdot \#\mathcal{N}_{i't'}^{\mathit{match}}} & \text{if } (i,t) \in \mathcal{N}_{i't'}^{\mathit{match}} \\ \frac{D_{i't'}(2X_{it}-1)(2X_{i't'}-1)}{K_{i't'}\cdot \#\mathcal{A}_{i't'}^{\mathit{match}}} & \text{if } (i,t) \in \mathcal{A}_{i't'}^{\mathit{match}} \\ 0 & \text{otherwise.} \end{cases}$$

Proof We first establish the following equality.

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
= \;& \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}\right) \\
= \;& \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
= \;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{M}_{i't'}^{\mathsf{match}}\cdot\#\mathcal{N}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{N}_{i't'}^{\mathsf{match}}\cdot\#\mathcal{M}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\left(a_{i't'} - \#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}\right)}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}}\right)\right. \\
& \left. +(1-X_{i't'})\left(\frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{M}_{i't'}^{\mathsf{match}}\cdot\#\mathcal{N}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{N}_{i't'}^{\mathsf{match}}\cdot\#\mathcal{M}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} - \frac{\left(\#\mathcal{A}_{i't'}^{\mathsf{match}} - a_{i't'} - a_{i't'}\right)}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}}\right)\right\} \\
= \;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\left(a_{i't'} - \#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}\right)}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}}\right)\right. \\
& \left. +(1-X_{i't'})\left(\frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} - \frac{\left(\#\mathcal{A}_{i't'}^{\mathsf{match}} - a_{i't'} - a_{i't'}\right)}{\#\mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}}\right)\right\} \\
= \;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(2X_{i't'} + 2(1-X_{i't'})\right) = 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}.
\end{aligned} \tag{28}$$

The third equality follows from the fact that for a given unit (i',t') there are $\#\mathcal{M}_{i't'}^{\mathsf{match}}$ matched observations $(i,t) \in \mathcal{M}_{i't'}^{\mathsf{match}}$ with weights equal to $\frac{D_{i't'}K_{i't'}}{\#\mathcal{M}_{i't'}^{\mathsf{match}}} = \frac{D_{i't'}\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{M}_{i't'}^{\mathsf{match}}(\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'})} = \frac{D_{i't'}\#\mathcal{N}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'}}$. Similarly, there are $\#\mathcal{N}_{i't'}^{\mathsf{match}}$ observations $(i,t) \in \mathcal{N}_{i't'}^{\mathsf{match}}$ with weights $\frac{D_{i't'}\#\mathcal{M}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'}}$. The final matched set $\mathcal{A}_{i't'}^{\mathsf{match}}$ is composed of $a_{i't'}$ observations with the same treatment status with (i',t') and $\mathcal{A}_{i't'}^{\mathsf{match}} - a_{i't'}$ observations with the opposite treatment status. When $X_{i't'}$, the former type gets weight equal to $\frac{D_{i't'}}{\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'}}$ while the latter type is weighted by $-\frac{D_{i't'}}{\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'}}$. The unit itself gets weight equal to $\frac{\#\mathcal{A}_{i't'}^{\mathsf{match}}}{\#\mathcal{A}_{i't'}^{\mathsf{match}}+a_{i't'}}$. All the other observations will get zero weight.

Following the same logic from above, it is straightforward to show that $\overline{X}_i^* = \overline{X}_t^* = \overline{X}^* = \frac{1}{2}$, and thus

$$X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^* \quad = \quad \left\{ \begin{array}{ll} \frac{1}{2} & \text{if } X_{it} = 1 \\ -\frac{1}{2} & \text{if } X_{it} = 0 \end{array} \right. \tag{29}$$

For instance,

$$\begin{aligned}
\overline{X}^* &= \frac{\sum_{i=1}^N \sum_{t=1}^T W_{it} X_{it}}{\sum_{i=1}^N \sum_{t=1}^T W_{it}} \\
&= \frac{\sum_{i'=1}^N \sum_{t'=1}^T \left(\sum_{i=1}^N \sum_{t=1}^T X_{it} w_{it}^{i't'} \right)}{2 \sum_{i=1}^N \sum_{t=1}^T D_{it}} \\
&= \frac{\sum_{i'=1}^N \sum_{t'=1}^T \left(\sum_{i=1}^N \sum_{t=1}^T X_{i't'} X_{it} w_{it}^{i't'} + (1 - X_{i't'}) X_{it} w_{it}^{i't'} \right)}{2 \sum_{i=1}^N \sum_{t=1}^T D_{it}} \\
&= \frac{\sum_{i'=1}^N \sum_{t'=1}^T D_{i't'} X_{i't'} \left(\frac{\# \mathcal{A}_{i't'}^{\mathsf{match}}}{\# \mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} + \frac{a_{i't'}}{\# \mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} \right) + D_{i't'} (1 - X_{i't'}) \left(\frac{\# \mathcal{A}_{i't'}^{\mathsf{match}}}{\# \mathcal{A}_{i't'}^{\mathsf{match}} + a_{i't'}} - \frac{-a_{i't'}}{\# \mathcal{A}_{i't'}^{\mathsf{match}} a_{i't'}} \right)}{2 \sum_{i=1}^N \sum_{t=1}^T D_{it}} \\
&= \frac{\sum_{i=1}^N \sum_{t=1}^T D_{it}}{2 \sum_{i=1}^N \sum_{t=1}^T D_{it}} = \frac{1}{2}
\end{aligned}$$

We can derive the desired result.

$$\begin{aligned}
\hat{\beta}_{\mathsf{WFE2}} &= \frac{\sum_{i=1}^N \sum_{t=1}^T W_{it} (X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*)}{\sum_{i=1}^N \sum_{t=1}^T W_{it} (X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)^2} \\
&= \frac{\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T W_{it} (2X_{it} - 1)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*)}{\frac{1}{4} \sum_{i=1}^N \sum_{t=1}^T W_{it}} \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i=1}^N \sum_{t=1}^T W_{it} (2X_{it} - 1)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i=1}^N \sum_{t=1}^T W_{it} (2X_{it} - 1) Y_{it} \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i=1}^N \sum_{t=1}^T \left\{ \left(\sum_{i'=1}^N \sum_{t'=1}^T w_{it}^{i't'} \right) (2X_{it} - 1) Y_{it} \right\} \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i'=1}^N \sum_{t'=1}^T \left\{ X_{i't'} \left(\sum_{i=1}^N \sum_{t=1}^T w_{it}^{i't'} (2X_{it} - 1) Y_{it} \right) + (1 - X_{i't'}) \left(\sum_{i=1}^N \sum_{t=1}^T w_{it}^{i't'} (2X_{it} - 1) Y_{it} \right) \right\} \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i'=1}^N \sum_{t'=1}^T \frac{D_{i't'}}{K_{i't'}} \left\{ X_{i't'} \left(Y_{it} - \frac{\sum_{(i,t) \in \mathcal{M}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{M}_{i't'}^{\mathsf{match}}} - \frac{\sum_{(i,t) \in \mathcal{N}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{N}_{i't'}^{\mathsf{match}}} + \frac{\sum_{(i,t) \in \mathcal{A}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{A}_{i't'}^{\mathsf{match}}} \right) \right. \\
&\qquad \left. + (1 - X_{i't'}) \left(\frac{\sum_{(i,t) \in \mathcal{M}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{M}_{i't'}^{\mathsf{match}}} + \frac{\sum_{(i,t) \in \mathcal{N}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{N}_{i't'}^{\mathsf{match}}} - \frac{\sum_{(i,t) \in \mathcal{A}_{i't'}^{\mathsf{match}}} Y_{it}}{\# \mathcal{A}_{i't'}^{\mathsf{match}}} - Y_{it} \right) \right\} \\
&= \frac{1}{\sum_{i=1}^N \sum_{t=1}^T D_{it}} \sum_{i=1}^N \sum_{t=1}^T \frac{D_{it}}{K_{it}} (\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}) = \hat{\tau}_{\mathsf{match2}}
\end{aligned}$$

where the second and third equality follows from equation (28) and (29). The last two equalities follow from applying the definition of K_{it}, W_{it} $\widehat{Y_{it}(1)}$ and $\widehat{Y_{it}(0)}$ given in Proposition 4. □

Unlike Proposition 3, the adjustment is done by deflating the estimated treatment effect by $1/K_{it}$. This is because the attenuation bias from $\mathcal{A}_{it}^{\textsf{match}}$ (the "pooled" part) is *subtracted* from the sum of two unbiased estimates from $\mathcal{M}_{it}^{\textsf{match}}$ and $\mathcal{N}_{it}^{\textsf{match}}$, amplifying the estimated treatment effect for the observation. In the example of Panel (b) of Figure 3, $\mathcal{A}_{it}^{\textsf{match}}$ contains four mismatches (bold **1** entries in triangles), i.e., $a_{it} = 4$, and hence the adjustment factor is $(6+4)/6$ where the denominator represents the total number of adjustments observations.

C Proof of Theorem 1

The proof of this theorem follows directly from Proposition 4 as the within-unit and within-time matched sets are subsets of $\mathcal{M}_{it}^{\textsf{match}}$ and $\mathcal{N}_{it}^{\textsf{match}}$. Specifically, $\mathcal{M}_{it}^{\textsf{DiD}}$ consists of up to one observation $(i, t-1)$ that is under the opposite treatment status, i.e., $\{(i', t') : i' = i, t' = t-1, X_{i't'} = 0\}$, while $\mathcal{N}_{it}^{\textsf{DiD}}$ is limited to the observations in the same time period whose prior observation is also under the control condition.

$$
\begin{aligned}
\hat{\beta}_{\textsf{DiD}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*)}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)^2} \\
&= \frac{\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*)}{\frac{1}{4}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)Y_{it} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\{ \left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'} \right) (2X_{it}-1)Y_{it} \right\} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} \left\{ X_{i't'} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)Y_{it} \right) + (1 - X_{i't'}) \left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)Y_{it} \right) \right\} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left\{ X_{i't'} \left(Y_{i't'} - Y_{i',t'-1} - \frac{\sum_{(i,t')\in\mathcal{N}_{i't'}^{\textsf{DiD}}} Y_{it'}}{\#\mathcal{N}_{i't'}^{\textsf{DiD}}} + \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}^{\textsf{DiD}}} Y_{it}}{\#\mathcal{A}_{i't'}^{\textsf{DiD}}} \right) \right. \\
&\qquad \left. + (1 - X_{i't'}) \left(Y_{i',t'-1} + \frac{\sum_{(i,t')\in\mathcal{N}_{i't'}^{\textsf{DiD}}} Y_{it'}}{\#\mathcal{N}_{i't'}^{\textsf{DiD}}} - \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}^{\textsf{DiD}}} Y_{it}}{\#\mathcal{A}_{i't'}^{\textsf{DiD}}} - Y_{i't'} \right) \right\} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}) = \hat{\tau}_{\textsf{DID}}
\end{aligned}
$$

where the seventh equality follows from the fact that, given $\mathcal{M}_{i't'}^{\textsf{DiD}}$ and $\mathcal{N}_{i't'}^{\textsf{DiD}}$, all the units in $\mathcal{A}_{i't'}^{\textsf{DiD}}$ are under the opposite treatment status (i.e., $a_{i't'} = 0$), and thus $K_{i't'} = 1$ (see Proposition 4). □

D Efficient Computation of the Weighted Two-way Fixed Effects Estimator

We consider the efficient computation of the weighted two-way fixed effects estimator in a general case with time-varying covariates $\mathbf{Z}_{it}$. Define the following matrices without observations that have zero weights,

$$\mathbf{Y} \;=\; \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_N \end{pmatrix}, \quad \mathbf{U} \;=\; [\mathbf{U}_1 \; \mathbf{U}_2] \;=\; \begin{bmatrix} \mathbf{D}_1 \, \iota_T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{D}_1 \\ \mathbf{0} & \mathbf{D}_2 \, \iota_T & \cdots & \mathbf{0} & \mathbf{D}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_N \, \iota_T & \mathbf{D}_N \end{bmatrix},$$

$$\mathbf{V} \;=\; \begin{bmatrix} \mathbf{X}_1 & \mathbf{Z}_1 \\ \mathbf{X}_2 & \mathbf{Z}_2 \\ \vdots & \vdots \\ \mathbf{X}_N & \mathbf{Z}_N \end{bmatrix}, \quad \text{and} \quad \mathbf{W} \;=\; \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_N \end{bmatrix},$$

where $\mathbf{Y}_i$ and $\mathbf{X}_i$ are T_i-dimensional vectors of the outcome and treatment variables for unit i which include time periods with non-zero weights respectively (T_i is the number of time periods for unit i with non-zero weights), $\mathbf{D}_i$ is the $(T_i \times T)$ matrix obtained by deleting the rows corresponding to time periods with zero regression weights for unit i, i.e., $W_{it} = 0$, from the $(T \times T)$ identity matrix, $\mathbf{Z}_i$ is the $(T_i \times M)$ matrix of observed covariates for unit i that vary across units and time periods, and $\mathbf{W}_i$ is an $(T_i \times T_i)$ diagonal matrix of non-zero regression weights for unit i. Thus, $\mathbf{U}_1$ and $\mathbf{U}_2$ represent the matrix of unit and time indicators, respectively. Under this setup, the weighted two-way fixed effects estimator with covariates can be in general written as $(\mathbf{S}^\top \mathbf{W} \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{W} \mathbf{Y}$ where $\mathbf{S} = (\mathbf{U} \; \mathbf{V})$.

The computation of this weighted two-way fixed effects estimator is expensive when the number of units and/or time periods is large. Since some of the regression weights can be negative, this means that the projection is done on the complex plane. Specifically, we first define $\mathbf{W}^{1/2}$ where we use the imaginary unit i so that $\sqrt{W_{it}} = i\sqrt{-W_{it}}$ if $W_{it} < 0$. Then, the projection matrix for $\mathbf{U}_1^* = \mathbf{W}^{1/2}\mathbf{U}_1$ is given by $\mathbf{P}_1 = \mathbf{U}_1^*(\mathbf{U}_1^\top \mathbf{W} \mathbf{U}_1)^- \mathbf{U}_1^{*\top}$ where $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$. We note that $\mathbf{U}_1^\top \mathbf{W} \mathbf{U}_1$ is an $(N \times N)$ diagonal matrix whose ith diagonal element equals the sum of weights across time, i.e., $\sum_{t=1}^T W_{it}$ and hence its generalized inverse is an $(N \times N)$ diagonal matrix whose ith diagonal element equals $1/\sum_{t=1}^T W_{it}$ if $\sum_{t=1}^T W_{it} \neq 0$ and 0 otherwise. This implies that the computation of $\mathbf{P}_1$ does not require the inversion of a large matrix.

Now, applying Lemma 1 of Davis (2002), the projection matrix on $\mathbf{U}^* = \mathbf{W}^{1/2}\mathbf{U}$ is given by $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ where $\mathbf{P}_2 = \mathbf{Q}\left(\mathbf{Q}^\top \mathbf{Q}\right)^- \mathbf{Q}^\top$ and $\mathbf{Q} = (\mathbf{I} - \mathbf{P}_1)\mathbf{W}^{1/2}\mathbf{U}_2$. The computation of $\mathbf{P}_2$ requires the inversion of $\mathbf{Q}^\top \mathbf{Q}$, which is a $(T \times T)$ matrix. Thus, the computation strategy outlined here is particularly effective when $N \gg T$ (if $N \ll T$, then one can devise a similar strategy by flipping the roles of unit and time dimensions). Finally, the weighted two-way fixed effects estimators for the coefficients of the treatment variable X_{it} and covariates $\mathbf{Z}_{it}$ can be obtained by regressing $\mathbf{Y}^* = (\mathbf{I} - \mathbf{P})\mathbf{W}^{1/2}\mathbf{Y}$ on $\mathbf{V}^* = (\mathbf{I} - \mathbf{P})\mathbf{W}^{1/2}\mathbf{V}$.