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

Kosuke Imai* In Song Kim†

Forthcoming in *Political Analysis*

Abstract

The two-way linear fixed effects regression (2FE) has become a default method for estimating causal effects from panel data. Many applied researchers use the 2FE estimator to adjust for unobserved unit-specific and time-specific confounders at the same time. Unfortunately, we demonstrate that the ability of the 2FE model to simultaneously adjust for these two types of unobserved confounders critically relies upon the assumption of linear additive effects. Another common justification for the use of the 2FE estimator is based on its equivalence to the difference-in-differences estimator under the simplest setting with two groups and two time periods. We show that this equivalence does not hold under more general settings commonly encountered in applied research. Instead, we prove that the multi-period difference-in-differences estimator is equivalent to the weighted 2FE estimator but with some observations having negative weights. These analytical results imply that in contrast to the popular belief, the 2FE estimator does not represent a design-based, nonparametric estimation strategy for causal inference. Instead, its validity fundamentally rests on the modeling assumptions.

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

*Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, 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 the two-way fixed effects (2FE) regression, or linear regression with unit and time fixed effects, as the default methodology for estimating causal effects from panel data. Applied researchers often use the 2FE regression to adjust for unobserved unit-specific and time-specific confounders at the same time. Unfortunately, we show that the 2FE's ability to simultaneously adjust for the two types of unobserved confounders critically hinges upon the assumption of linear additive effects. Another common justification is based on the fact that the 2FE estimator is equivalent to the difference-in-differences estimator under the simplest setting with two groups and two time periods (e.g., Bertrand *et al.*, 2004; Angrist and Pischke, 2009). However, we show that this equivalence does not hold under more general settings frequently encountered in applied research. All together, we show that in contrast to the popular belief, the 2FE estimator does not represent a design-based, nonparametric estimation strategy for causal inference. Instead, its validity fundamentally rests on the modeling assumptions.

Our work builds on the growing literature about causal inference with panel data. In particular, we extend the matching representation of one-way fixed effects regression estimator (Imai and Kim, 2019) to the 2FE estimator in order to understand the causal interpretation of these widely used estimators within the nonparametric framework (see e.g., Humphreys, 2009; Aronow and Samii, 2015; Solon *et al.*, 2015, for related work on causal inference with cross-sectional data). In addition, a number of scholars have recently considered causal interpretations of the standard 2FE estimator (see e.g., Borusyak and Jaravel, 2017; Abraham and Sun, 2018; Athey and Imbens, 2018; Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018). While many of these studies assume staggered adoption, our analysis extends to a more general case, in which units can go in and out of the treatment condition at different points in time. Finally, we emphasize that the goal of this paper is to shed new light on two common misunderstandings of the FE estimator rather than to propose an alternative estimator.

2 The Two-way Fixed Effects Regression Estimator

Suppose that we have a panel data set of N units and T time periods. Although our results readily extend to the case of unbalanced panel, for the sake of notational simplicity, we assume a balanced panel data set. 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 (2FE) 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. Specifically, we can define unit and time fixed effects as $\alpha_i = h(\mathbf{U}_i)$ and $\gamma_t = f(\mathbf{V}_t)$, where $\mathbf{U}_i$ and $\mathbf{V}_t$ represent these unit-specific and time-specific unobserved confounders that are common causes of the outcome and treatment variables, and $h(\cdot)$ 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 $f(\cdot)$. In other words, since the treatment is binary, the model makes no restriction other than the additivity and separability of the two types of unobserved confounders.

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,

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

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. Equation (2) shows how the 2FE 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.

3 Adjustment for Unobserved Confounders

Many applied researchers justify the use of the 2FE estimator by its ability to simultaneously adjust for unit-specific and time-specific unobserved confounders. We show below that such a justification is unwarranted without critically relying on the functional-form assumption. Indeed, by extending the matching framework of Imai and Kim (2019), we show that the simultaneous adjustment for the two types of unobserved confounders cannot be done nonparametrically under the 2FE framework.

3.1 The Matching Framework

To establish the impossibility of nonparametric adjustment for unit-specific and time-specific unobserved confounders, it is useful to consider the 2FE estimator as a matching estimator (Imai and Kim, 2019). An intuitive explanation of this result is as follows. Although one could nonparametrically adjust for unit-specific (time-specific) unobserved confounders by matching a treated observation with control observations of the same unit (time period), no other observation shares the same unit and time indices. Thus, the 2FE estimator critically relies upon the linearity assumption for its simultaneous adjustment for the two types of unobserved confounders. The following proposition formalizes this argument.

Proposition 1 (The Two-way Fixed Effects Regression Estimator as a Two-way Matching Estimator) *The two-way fixed effects estimator defined in equation* (2) *is equivalent to the following matching estimator,*

$$\hat{\beta} \;=\; \frac{1}{K}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\widehat{Y_{it}(0)}\right)+(1-X_{it})\left(\widehat{Y_{it}(1)}-Y_{it}\right)\right\}\right]$$

where for $x = 0, 1$, *the estimate of the potential outcome* $Y_{it}(x)$ *for unit* i *at time* t *under the treatment status* $X_{it} = x$ *is given by,*

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

$$+ (1 - X_{it}) \left(\frac{\sum_{t' \neq t} X_{it'}}{T-1} + \frac{\sum_{i' \neq i} X_{i't}}{N-1} - \frac{\sum_{i' \neq i} \sum_{t' \neq t} X_{i't'}}{(T-1)(N-1)} \right) \Bigg\}.$$

The proposition shows that the estimated counterfactual outcome of a given observation, i.e., $\widehat{Y_{it}(1 - X_{it})}$, is a function of three averages. First, the average of all the other observations from the same unit, i.e., $\sum_{t' \neq t} Y_{it'}/(T-1)$, and the average of all the other observations from the same time period, i.e., $\sum_{i' \neq i} Y_{i't}/(N-1)$, are added together. We call them the *within-unit matched set* $\mathcal{M}_{it}$ and the *within-time matched set* $\mathcal{N}_{it}$, respectively, and formally define them as,

$$\mathcal{M}_{it} \;=\; \{(i', t') : i' = i, t' \neq t\}, \quad \text{and} \quad \mathcal{N}_{it} \;=\; \{(i', t') : i' \neq i, t' = t\}. \tag{3}$$

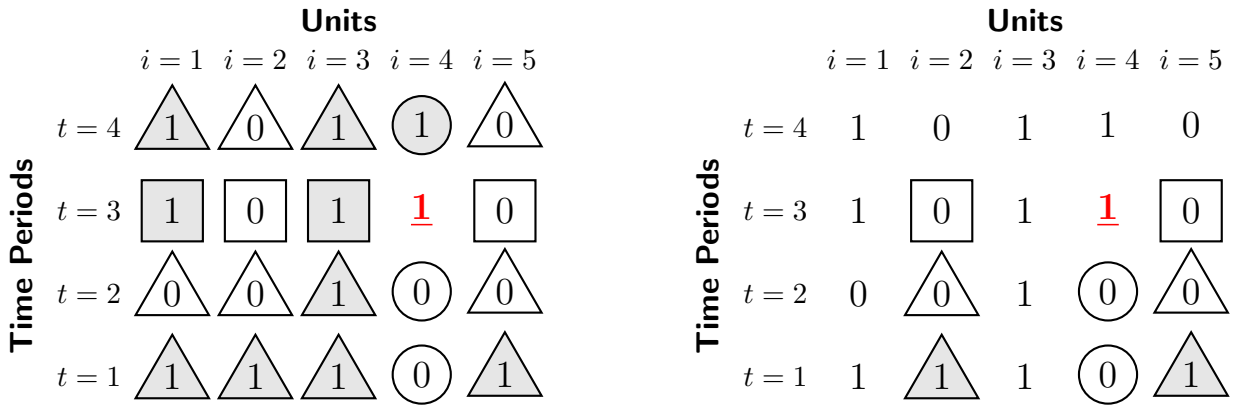
The 2FE estimator then adjusts for unit-specific and time-specific unobserved confounders by using observations that share the same unit or time as those in $\mathcal{N}_{it}$ and $\mathcal{M}_{it}$, respectively, and subtracting their mean, i.e., $\sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'}/(T-1)(N-1)$, from this sum. We use $\mathcal{A}_{it}$ to denote this group of observations and call it the *adjustment set* for observation (i, t) with the following definition,

$$\mathcal{A}_{it} \;=\; \{(i', t') : i' \neq i, t' \neq t, (i, t') \in \mathcal{M}_{it}, (i', t) \in \mathcal{N}_{it}\}. \tag{4}$$

By construction, the number of observations in $\mathcal{A}_{it}$ equals the product of the number of observations in the within-unit and within-time matched sets, i.e., $|\mathcal{A}_{it}| = |\mathcal{M}_{it}| \cdot |\mathcal{N}_{it}|$.

Panel (a) of Figure 1 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 a treated observation of interest, 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 $\mathcal{M}_{it}$ (circles in the figure), plus the average of control observations from the same time period $\mathcal{N}_{it}$ (squares), minus the average of adjustment observations, $\mathcal{A}_{it}$ (triangles).

Note that all of these three averages may include units with the same treatment status as the observation whose counterfactual outcome is being estimated. We refer to these observations as "mismatches" (shaded grey entries in the figure) because for the estimation of causal effects, an observation must be matched with another observation with the *opposite* treatment status. Therefore, mismatches



(a) Two-way fixed effects estimator (b) Two-way matching estimator

Figure 1: **An Example of the Binary Treatment Matrix with Five Units and Four Time Periods**. Panels (a) and (b) illustrate how observations $(i,\ t)$ are used to estimate counterfactual outcomes for the two-way fixed effects estimator (Proposition 1) and the adjusted matching estimator (Proposition 2), respectively. In the figures, the red underlined **1** entry (4,3) represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circles indicate the set of matched observations—(4,1), (4,2), (4,4) in Panel (a) and (4,1), (4,2) in Panel (b)—that are from the same unit, whereas squares indicate those—(1,3), (2,3), (3,3), (5,3) in Panel (a) and (2,3), (5,3) in Panel (b)—from the same time period. Finally, triangles represent the set of observations—(1,1), (1,2), (1,4), (2,1), (2,2), (2,4), (3,1), (3,2), (3,4), (5,1), (5,2), (5,4) in Panel (a) and (2,1), (2,2), (5,1), (5,2) in Panel (b)—that are used to make adjustment for unit and time effects. The shaded grey symbols represent the "mismatches" with the same treatment status, which are prevalent in the two-way fixed effects estimator. The matching estimator in Panel (b) is designed to eliminate the attenuation bias within unit and time, although the adjustment set may still include mismatches (shaded triangles).

imply the (partial) comparison of observations with the same treatment status, which generally leads to an attenuation bias. The 2FE estimator adjusts for this bias via the factor K, which is equal to the net proportion of proper matches between the observations of opposite treatment status. For example, for a treated observation with $X_{it} = 1$, we compute the proportion of matched control observations in the within-unit matched set, i.e., $\sum_{t' \neq t}(1-X_{it'})/(T-1)$, and the proportion of matched control observations in the within-time matched set, i.e., $\sum_{i' \neq i}(1-X_{i't})/(N-1)$, and subtract from their sum the proportion of matched control observations in the adjustment set, i.e., $\sum_{i' \neq i}\sum_{t' \neq t}(1 - X_{i't'})/(T-1)(N-1)$.

3.2 The Impossibility of Nonparametric Adjustment

Given this result, it is natural to ask whether we can eliminate the mismatches and the adjustment set all together within the two-way fixed effects framework. We show below that this is generally impossible.

In particular, although we can construct a weighted 2FE estimator that has fewer mismatches, this estimator in general still suffers from some mismatches and has an adjustment set.

To develop a weighted 2FE estimator with fewer mismatches, we begin by matching each observation only with other observations of the *opposite* treatment status to estimate the counterfactual outcome. That is, we use the following *within-unit matched set* $\mathcal{M}_{it}^*$, which consists of the observations within the same unit but with the opposite treatment status,

$$\mathcal{M}_{it}^* \;=\; \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}\}. \tag{5}$$

Similarly, we restrict the *within-time matched set* so that its observations belong to the same time period t but have the opposite treatment status,

$$\mathcal{N}_{it}^* \;=\; \{(i', t') : t' = t, X_{i't'} = 1 - X_{it}\}. \tag{6}$$

Then, using equation (4), we can define the corresponding adjustment set $\mathcal{A}_{it}^*$.

$$\mathcal{A}_{it}^* \;=\; \{(i', t') : i' \neq i, t' \neq t, (i, t') \in \mathcal{M}_{it}^*, (i', t) \in \mathcal{N}_{it}^*\}. \tag{7}$$

The next proposition establishes that this two-way matching estimator, which eliminates mismatches within-unit and within-time dimension, can be written as a weighted 2FE estimator.

Proposition 2 (The Two-way Matching Estimator with Fewer Mismatches 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 matching estimator,*

$$\hat{\beta}^* \;=\; \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{D_{it}}{K_{it}} \left\{ X_{it} \left(Y_{it} - \widehat{Y_{it}(0)} \right) + (1 - X_{it}) \left(\widehat{Y_{it}(1)} - Y_{it} \right) \right\}$$

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

$$\begin{aligned} \widehat{Y_{it}(x)} \;&=\; \frac{1}{|\mathcal{M}_{it}^*|} \sum_{(i,t') \in \mathcal{M}_{it}^*} Y_{it'} + \frac{1}{|\mathcal{N}_{it}^*|} \sum_{(i',t) \in \mathcal{N}_{it}^*} Y_{i't} - \frac{1}{|\mathcal{A}_{it}^*|} \sum_{(i',t') \in \mathcal{A}_{it}^*} Y_{i't'} \\ K_{it} \;&=\; 1 + \frac{a_{it}}{|\mathcal{A}_{it}^*|} \end{aligned}$$

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

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

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

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

Proof is given in Appendix B. Unlike Proposition 1, the adjustment is done by deflating the estimated treatment effect for each treated observation (i,t) by $1/K_{it}$. This is because the attenuation bias from $\mathcal{A}^*_{it}$ (the "pooled" part) is *subtracted* from the sum of two estimates from $\mathcal{M}^*_{it}$ and $\mathcal{N}^*_{it}$, inflating the estimated treatment effect for a given observation (i,t). In the example of Panel (b) of Figure 1, $\mathcal{A}^*_{it}$ contains two mismatches (shaded grey entries in triangles), i.e., $a_{it} = 2$, and hence the adjustment factor is $K_{it} = 3/2 = 1 + 2/4$. Note that such adjustment is not necessary (i.e., $K_{it} = 1$) when there are no mismatches in the adjustment set, i.e. $a_{it} = 0$.

The algebraic equivalence result given in Proposition 2 clarifies the set of observations that are used to estimate the counterfactual for each unit and how the adjustments due to mismatches are reflected in the weighted two-way fixed effects estimator. Specifically, it shows that each observation (i,t) is weighted differently according to the number of times it serves as a control unit. For example, if an observation (i,t) has the treatment status opposite to another observation within-unit (i',t'), i.e., $(i,t) \in \mathcal{M}^*_{i't'}$, then its overall weight W_{it} is increased by $1/|\mathcal{M}^*_{i't'}|$ along with other observations in the within-unit matched set. This contribution to the weight is then deflated by the adjustment factor $K_{i't'}$, correcting the attenuation bias due to mismatches (see the formula for computing $w_{it}^{i't'}$ in the proposition).

Unfortunately, we cannot eliminate mismatches in $\mathcal{A}^*_{it}$ without additional restrictions on the matched sets, $\mathcal{M}^*_{it}$ and $\mathcal{N}^*_{it}$ (see Section 4.1). This point is illustrated by Panel (b) of Figure 1 where the adjustment set $\mathcal{A}^*_{it}$ (triangles) still includes the observations of the same treatment status. Therefore, even the weighted 2FE estimator, which has fewer mismatches than the standard 2FE estimator, suffers

from some mismatches. The estimator also has an adjustment set whose observations belong to neither the same unit nor the same time period as the observation being matched with. This implies that it is impossible to simultaneously and nonparametrically adjust for unit-specific and time-specific unobserved confounders under the two-way fixed effects framework.

4 The Difference-in-Differences Design

Although it is generally impossible to eliminate all mismatches, in this section we show that we can do so under the difference-in-differences (DiD) design. In contrast to a common belief among applied researchers, we also show that under the general panel data settings, the DiD estimator is not equivalent to the standard 2FE estimator. Instead, the multi-period DiD estimator is equal to the weighted 2FE estimator with some observations having invalid (i.e., negatie) regression weights. This implies that the equivalence between the 2FE estimator and the DiD estimator critically hinges on the linearity assumption.

4.1 The Multi-period Difference-in-Differences Estimator

To establish the relations between the 2FE and DiD estimators, we begin by considering the following parallel trend assumption,

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

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

We emphasize that this assumption may not be credible in some settings (see e.g., Bilinski and Hatfield, 2018; Kahn-Lang and Lang, 2019; Rambachan and Roth, 2019). The goal of our analysis, however, is to shed new light on a popular justification of the 2FE estimator as the DiD estimator under the simplest setting.[1] Under this parallel trend assumption, the estimand is the average treatment effect for the treated (ATT),

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

[1]For example, Bertrand *et al.* (2004) describe the linear regression model with two-way fixed effects as "a common generalization of the most basic DiD setup (with two periods and two groups)" (p. 251).

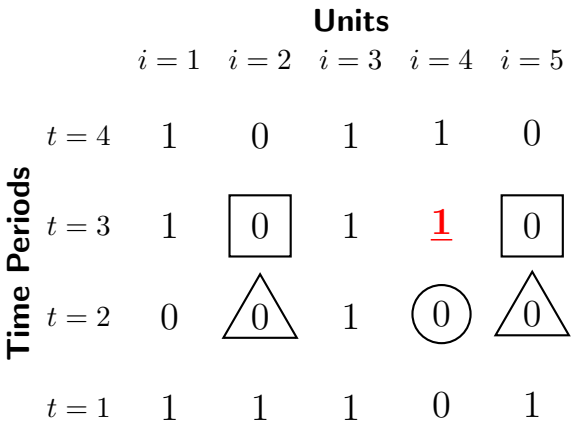


Figure 2: **Illustration of how observations are used to estimate counterfactual outcomes for the DiD estimator (equation** (12)**).** The red underlined 1 entry represents the treated observation $(4,3)$, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circle indicates the matched observation (4,2) within the same unit, $\mathcal{M}_{it}^{\mathsf{DiD}}$, whereas squares—(2,3) and (5,3)—indicate those from the same time period, $\mathcal{N}_{it}^{\mathsf{DiD}}$. Finally, triangles—(2,2) and (5,2)—represent the set of observations that are used to make adjustment for unit and time effects, $\mathcal{A}_{it}^{\mathsf{DiD}}$. Unlike the examples in Figure 1, $\mathcal{A}_{it}^{\mathsf{DiD}}$ only contains control observations and hence no mismatches (i.e., shaded grey triangles) exist.

To formulate a multi-period DiD estimator under the 2FE estimator framework, we follow the analytical strategy used in the previous section and define three sets of observations as illustrated in Figure 2 — 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 $(4,3)$ (represented by the red underlined 1). We next show that the DiD design eliminates mismatches from these three sets.

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,

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

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,

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

Finally, we define the adjustment set $\mathcal{A}_{it}^{\mathsf{DiD}}$, which contains the control observations in the previous

period that share the same unit as those in $\mathcal{N}_{it}^{\mathsf{DiD}}$,

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

Thus, the number of observations in this adjustment set is the same as that in $\mathcal{N}_{it}^{\mathsf{DiD}}$. It is worth noting that all three sets only contain control observations, thereby eliminating all mismatches.

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,

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

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

$$\widehat{Y_{it}(0)} \quad = \quad Y_{i,t-1} + \frac{1}{|\mathcal{N}_{it}^{\mathsf{DiD}}|} \sum_{(i',t)\in\mathcal{N}_{it}^{\mathsf{DiD}}} Y_{i't} - \frac{1}{|\mathcal{A}_{it}^{\mathsf{DiD}}|} \sum_{(i',t')\in\mathcal{A}_{it}^{\mathsf{DiD}}} Y_{i't'} \tag{13}$$

Thus, 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 as follows. We subtract 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.

4.2 Equivalence to the Weighted Two-way Fixed Effects Estimator with Some Negative Regression Weights

It is well known that the standard nonparametric DiD estimator is numerically equivalent to the 2FE estimator in the simplest setting, in which there are only two time periods and the treatment is administered only to one group of units in the second time period. Unfortunately, we show that this equivalence result does not generalize to the current multi-period DiD design, in which the number of time periods may exceed two and different units may switch in and out of the treatment condition

multiple times and at different points in time.[2] Instead, the following theorem establishes that the general multi-period DiD estimator given in equation (12) is equivalent to a *weighted* two-way fixed effects regression 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,

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

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.

Theorem 1 has two important implications. First, in contrast to a common belief held among applied researchers, the (unweighted) 2FE estimator is not in general equivalent to the multi-period DiD estimator. Second, although the multi-period DiD estimator can be shown to be equivalent to the weighted 2FE estimator, some control observations will have negative regression weights. This occurs when they frequently enter into the adjustment set, $\mathcal{A}_{i't'}^{\mathsf{DiD}}$, for multiple treated observations (i.e., $(2X_{it}-1)(2X_{i't'}-1) = -1$). Since the regression weights should generally positive, the results of this section shows that the justification of the 2FE estimator as the DiD estimator is not warranted unless the linearity assumption is imposed.

[2] If the model in equation (1) is assumed to be correct, then the 2FE estimator is consistent for τ under the multi-period DiD design. That is, if we rewrite the 2FE model specified using the potential outcome notation, i.e., $Y_{it}(x) = \alpha_i + \gamma_t + \beta x + \epsilon_{it}$, we have $\beta = \tau$.

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. We show that contrary to the common belief, the standard two-way fixed effects regression estimator does not represent a design-based, nonparametric causal estimator. It is impossible to simultaneously adjust for unobserved unit-specific and time-specific confounders. In addition, a general multi-period difference-in-differences estimator is equivalent to the weighted two-way fixed effects regression estimator, but some observations have invalid (i.e., negative) weights.

Given the problems of the standard two-way fixed effects regression estimator identified in this paper, future research should develop design-based estimators for causal inference with panel data. Recently, a number of researchers have extended the synthetical control method of Abadie *et al.* (2010) to more general settings (e.g., Xu, 2017; Ben-Michael *et al.*, 2019). In a separate paper, we have also generalized the multi-period difference-in-differences estimator introduced in this paper and proposed matching and weighting methods that are applicable to panel data (Imai *et al.*, 2018). In that paper, we show how to apply matching methods to time-series cross section data by explicitly comparing each treated observation with a set of control observations that are matched based on certain criteria. An advantage of such a method is the fact that it allows researchers to assess the quality of matches by examining the balance of confounders. Much research is needed to improve the existing methods for causal inference with panel data. While we have focused on a binary treatment variable, causal inference with general treatment regimes in panel data settings is of particular interest to many researchers.

Acknowledgements 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). We thank Clement de Chaisemartin and anonymous reviewers for helpful comments.

References

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.

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

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`.

Ben-Michael, E., Feller, A., and Rothstein, J. (2019). Synthetic controls and weighted event studies with staggered adoption. Tech. rep., arXiv:1912.03290.

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.

Bilinski, A. and Hatfield, L. A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273* .

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.

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`.

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

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* **63**, 2, 467–490.

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

Kahn-Lang, A. and Lang, K. (2019). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics* 1–14.

Rambachan, A. and Roth, J. (2019). An honest approach to parallel trends. Working paper available at `https://scholar.harvard.edu/jroth/publications/Roth_JMP_Honest_Parallel_Trends`.

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

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

Supplemental Appendix

A Proof of Proposition 1

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

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})-(1-X_{it})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})\right\}\\
=&\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}-\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right.\right.\\
&\qquad\left.-\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}-\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\\
&-(1-X_{it})\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}-\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right.\\
&\qquad\left.\left.-\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}-\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right]\\
=&\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right.\\
&\qquad\left.-(1-X_{it})\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right]\\
=&\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{\sum_{t'=1}^{T}Y_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}Y_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}}{(T-1)(N-1)}\right)\right.\\
&\qquad\left.-(1-X_{it})\left(\frac{\sum_{t'=1}^{T}Y_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}Y_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}-Y_{it}}{(T-1)(N-1)}\right)\right\}.\\
=&\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right)
\end{aligned}
\tag{14}
$$

The second algebraic equality we prove is the following,

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

$$
\begin{aligned}
&= \sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{\frac{(N-1)(T-1)}{NT}X_{it} - \frac{N-1}{NT}\sum_{t'\neq t}X_{it'} - \frac{T-1}{NT}\sum_{i'\neq i}X_{i't} + \frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\right. \\
&\qquad \left. -(1-X_{it})\left\{\frac{(N-1)(T-1)}{NT}X_{it} - \frac{N-1}{NT}\sum_{t'\neq t}X_{it'} - \frac{T-1}{NT}\sum_{i'\neq i}X_{i't} + \frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\right] \\
&= \frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left\{X_{it}\left(\frac{\sum_{t'=1}^{T}(1-X_{it'})}{T-1} + \frac{\sum_{i'=1}^{N}(1-X_{i't})}{N-1} - \frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}(1-X_{i't'})}{(T-1)(N-1)}\right)\right.\right. \\
&\qquad \left.\left. + (1-X_{it})\left(\frac{\sum_{t'=1}^{T}X_{it'}}{T-1} + \frac{\sum_{i'=1}^{N}X_{i't}}{N-1} - \frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}X_{i't'}}{(T-1)(N-1)}\right)\right\}\right] \\
&= K(T-1)(N-1) \qquad (15)
\end{aligned}
$$

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 (14) and (15). □

B Proof of Proposition 2

Proof We first establish the following equality.

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}W_{it} \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{M}^*_{i't'}\cdot\#\mathcal{N}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{N}^*_{i't'}\cdot\#\mathcal{M}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{(a_{i't'}-\#\mathcal{A}^*_{i't'}+a_{i't'})}{\#\mathcal{A}^*_{i't'}+a_{i't'}}\right)\right. \\
&\quad \left.+(1-X_{i't'})\left(\frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{M}^*_{i't'}\cdot\#\mathcal{N}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{N}^*_{i't'}\cdot\#\mathcal{M}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} - \frac{(\#\mathcal{A}^*_{i't'}-a_{i't'}-a_{i't'})}{\#\mathcal{A}^*_{i't'}+a_{i't'}}\right)\right\} \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{(a_{i't'}-\#\mathcal{A}^*_{i't'}+a_{i't'})}{\#\mathcal{A}^*_{i't'}+a_{i't'}}\right)\right. \\
&\quad \left.+(1-X_{i't'})\left(\frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}^*_{i't'}}{\#\mathcal{A}^*_{i't'}+a_{i't'}} - \frac{(\#\mathcal{A}^*_{i't'}-a_{i't'}-a_{i't'})}{\#\mathcal{A}^*_{i't'}+a_{i't'}}\right)\right\} \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}D_{i't'}\left(2X_{i't'}+2(1-X_{i't'})\right) = 2\sum_{i=1}^{N}\sum_{t=1}^{T}D_{it}. \qquad (16)
\end{aligned}
$$

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

$$X_{it} - \overline{X}^*_i - \overline{X}^*_t + \overline{X}^* = \begin{cases} \frac{1}{2} & \text{if } X_{it} = 1 \\ -\frac{1}{2} & \text{if } X_{it} = 0 \end{cases} \tag{17}$$

For instance,

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

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'}} Y_{it}}{\#\mathcal{M}^*_{i't'}} - \frac{\sum_{(i,t) \in \mathcal{N}^*_{i't'}} Y_{it}}{\#\mathcal{N}^*_{i't'}} + \frac{\sum_{(i,t) \in \mathcal{A}^*_{i't'}} Y_{it}}{\#\mathcal{A}^*_{i't'}} \right) \right. \\
&\qquad \left. + (1 - X_{i't'}) \left(\frac{\sum_{(i,t) \in \mathcal{M}^*_{i't'}} Y_{it}}{\#\mathcal{M}^*_{i't'}} + \frac{\sum_{(i,t) \in \mathcal{N}^*_{i't'}} Y_{it}}{\#\mathcal{N}^*_{i't'}} - \frac{\sum_{(i,t) \in \mathcal{A}^*_{i't'}} Y_{it}}{\#\mathcal{A}^*_{i't'}} - 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 (16) and (17). 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 2. □

C Proof of Theorem 1

The proof of this theorem follows directly from Proposition 2 as the within-unit and within-time matched sets are subsets of $\mathcal{M}^*_{it}$ and $\mathcal{N}^*_{it}$. Specifically, $\mathcal{M}^{\mathsf{DiD}}_{it}$ 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 $\mathcal{N}^{\mathsf{DiD}}_{it}$ is limited to the observations in the same time period whose prior observation is also under the control condition.

$$
\begin{aligned}
\hat{\beta}_{\mathsf{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}^{\mathsf{DiD}}_{i't'}} Y_{it'}}{\#\mathcal{N}^{\mathsf{DiD}}_{i't'}} + \frac{\sum_{(i,t) \in \mathcal{A}^{\mathsf{DiD}}_{i't'}} Y_{it}}{\#\mathcal{A}^{\mathsf{DiD}}_{i't'}} \right) \right. \\
&\qquad \left. + (1 - X_{i't'}) \left(Y_{i',t'-1} + \frac{\sum_{(i,t') \in \mathcal{N}^{\mathsf{DiD}}_{i't'}} Y_{it'}}{\#\mathcal{N}^{\mathsf{DiD}}_{i't'}} - \frac{\sum_{(i,t) \in \mathcal{A}^{\mathsf{DiD}}_{i't'}} Y_{it}}{\#\mathcal{A}^{\mathsf{DiD}}_{i't'}} - 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}_{\mathsf{DiD}}
\end{aligned}
$$

where the seventh equality follows from the fact that, given $\mathcal{M}^{\mathsf{DiD}}_{i't'}$ and $\mathcal{N}^{\mathsf{DiD}}_{i't'}$, all the units in $\mathcal{A}^{\mathsf{DiD}}_{i't'}$ are under the opposite treatment status (i.e., $a_{i't'} = 0$), and thus $K_{i't'} = 1$ (see Proposition 2). □