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

Kosuke Imai†  In Song Kim†

January 12, 2020

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

*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 for comments.

†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 the two-way fixed effects regression (2FE), 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 the causal interpretation of the standard two-way fixed effects estimator (see e.g., Borusyak and Jaravel 2017, Abraham and Sun 2018, Athey and Imbens 2018, Chaisemartin and D’Haultfoeuille 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.

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 regression (2FE) model,

\[
Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it} 
\]

for \( i = 1, 2, \ldots, N \) and \( t = 1, 2, \ldots, T \) where \( \alpha_i \) and \( \gamma_t \) 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 \( \alpha_i = h(U_i) \) and \( \gamma_t = f(V_t) \), where \( U_i \) and \( V_t \) represent these unit-specific and time-specific unobserved confounders, 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 $f(\cdot)$.

The least squares estimate of $\beta$ 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,

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

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 here that such an adjustment cannot be done nonparametrically. In particular, we use the matching framework of Imai and Kim (2019) and extend their equivalence results to the 2FE case.

#### 3.1 The Matching Framework

To establish this result, it is useful to consider the 2FE estimator as a nonparametric matching estimator. Intuitively, 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. This implies the impossibility of simultaneous (nonparametric) 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,

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

where for $x = 0, 1$,

$$\overline{Y}_{it}(x) = \frac{1}{T-1} \sum_{t' \neq t} Y_{it'} + \frac{1}{N-1} \sum_{i' \neq i} Y_{i't} - \frac{1}{(T-1)(N-1)} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'}$$

$$K = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ X_{it} \left( \sum_{t' \neq t} \frac{(1 - X_{i't'})}{T-1} + \sum_{i' \neq i} \frac{(1 - X_{i't})}{N-1} - \frac{\sum_{i' \neq i} \sum_{t' \neq t} (1 - X_{i't'})}{(T-1)(N-1)} \right) \right\}$$
The proposition shows that the estimated counterfactual outcome of a given observation, i.e., \( Y_{\overline{1}}t(1 - X_{it}) \), is a function of three averages. First, the average of all the other observations from the same unit, i.e., \( \frac{\sum_{t' \neq t} Y_{it'}}{(T - 1)} \), and the average of all the other observations from the same time period, i.e., \( \frac{\sum_{t' \neq t} Y_{it'}}{(N - 1)} \), are added together. We call them the within-unit and within-time matched sets, i.e., \( \mathcal{M}_{it} \) and \( \mathcal{N}_{it} \), 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., \( \frac{\sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'}}{(T - 1)(N - 1)} \), 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).

Since all of these three averages may include units of the same treatment status, the 2FE 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. 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., \( \frac{\sum_{t' \neq t} (1 - X_{it'})}{(T - 1)} \), and the proportion of matched control observations in the within-time matched set, i.e., \( \frac{\sum_{i' \neq i} (1 - X_{i't})}{(N - 1)} \), and subtract from their sum the proportion of matched control observations in the adjustment set, i.e., \( \frac{\sum_{i' \neq i} \sum_{t' \neq t} (1 - X_{i't'})}{(T - 1)(N - 1)} \).

### 3.2 The Improved Matching Estimator

Given this result, we improve the 2FE estimator by reducing the number of mismatches. We accomplish this 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,

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

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,

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

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

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

Unfortunately, we cannot eliminate mismatches in \(A^*_{it}\) without additional restrictions on the matched sets, \(M^*_{it}\) and \(N^*_{it}\) (see Section 4.1). This point is illustrated by Panel (b) of Figure 1 where the adjustment set \(A^*_{it}\) (triangles) still includes the observations of the same treatment status.

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

**Proposition 2 (The Improved 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 improved matching estimator,

\[ \hat{\beta}^* = \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \left\{ X_{it} \left( Y_{it} - \bar{Y}_{it}(0) \right) + \left( 1 - X_{it} \right) \left( \bar{Y}_{it}(1) - \bar{Y}_{it} \right) \right\} \]

where \( D_{it} = 1\{|M_{it}^*| : |N_{it}^*| > 0\} \), and for \( x = 0,1 \),

\[ \bar{Y}_{it}(x) = \frac{1}{|M_{it}^*|} \sum_{(i',t') \in M_{it}^*} Y_{i't'} + \frac{1}{|N_{it}^*|} \sum_{(i',t') \in N_{it}^*} Y_{i't'} - \frac{1}{|A_{it}^*|} \sum_{(i',t') \in A_{it}^*} Y_{i't'} \]

\[ K_{it} = 1 + \frac{a_{it}}{|A_{it}^*|} \]

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

\[ \hat{\beta}^* = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \left\{ (Y_{it} - \bar{Y}_i^* - \bar{Y}_t^* + \bar{Y}^*) - \beta (X_{it} - \bar{X}_i^* - \bar{X}_t^* + \bar{X}^*) \right\}^2 \]

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

\[ W_{it} = \sum_{i'=1}^{N} \sum_{t'=1}^{T} w_{i't'} \quad \text{and} \quad w_{i't'} = \begin{cases} \frac{D_{i't'}}{K_{i't'}} & \text{if } (i,t) = (i',t') \\ \frac{K_{i't'}|M_{i't'}|}{D_{i't'}} & \text{if } (i,t) \in M_{i't'}^* \\ \frac{K_{i't'}|N_{i't'}|}{D_{i't'}} & \text{if } (i,t) \in N_{i't'}^* \\ \frac{D_{i't'}(2X_{i't'}-1)K_{i't'}|M_{i't'}|}{K_{i't'}|A_{i't'}^*|} & \text{if } (i,t) \notin A_{i't'}^* \\ 0 & \text{otherwise.} \end{cases} \]

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 \(A_{it}^*\) (the “pooled” part) is subtracted from the sum of two estimates from \(M_{it}^*\) and \(N_{it}^*\), inflating the estimated treatment effect for a given observation \((i,t)\). In the example of Panel (b) of Figure 1, \(A_{it}^*\) 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., \((i,t) \in M_{i't'}^*\), then its overall weight \(W_{it}\) is increased by \(1/|M_{i't'}^*|\) 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 \(w_{i't'}\) in the proposition).
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, \(M_{\text{DiD}}^{it}\), whereas squares—(2,3) and (5,3)—indicate those from the same time period, \(N_{\text{DiD}}^{it}\). Finally, triangles—(2,2) and (5,2)—represent the set of observations that are used to make adjustment for unit and time effects, \(A_{\text{DiD}}^{it}\). Unlike the examples in Figure 1, \(A_{\text{DiD}}^{it}\) only contains control observations and hence no mismatches (i.e., shaded grey triangles) exist.

4 Relations with the Difference-in-Differences Estimator

In this section, we propose a multi-period difference-in-differences (DiD) estimator under the parallel trend assumption that eliminates the mismatches in the adjustment set. We show that this estimator is equivalent to the weighted two-way fixed effects estimator but with some observations having negative regression weights. This implies that the equivalence between the 2FE estimator and the DiD estimator critically hinges on the modeling assumptions.

4.1 The Multi-period Difference-in-Differences Estimator

We propose a multi-period DiD estimator based on the following parallel trend assumption,

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

\[
E(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = 1, X_{i,t-1} = 0) = E(Y_{it}(0) - Y_{i,t-1}(0) \mid X_{it} = X_{i,t-1} = 0).
\]

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

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

To formulate the proposed multi-period DiD estimator, we 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).
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,

\[ \mathcal{M}_{it}^{\text{DiD}} = \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 0\}. \]  

(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,

\[ \mathcal{N}_{it}^{\text{DiD}} = \{(i', t') : i' \neq i, t' = t, X_{i't'} = X_{i't'-1} = 0\}. \]  

(10)

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

\[ \mathcal{A}_{it}^{\text{DiD}} = \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = X_{i't} = 0\}. \]  

(11)

Thus, the number of observations in this adjustment set is the same as that in \( \mathcal{N}_{it}^{\text{DiD}} \).

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,

\[ \hat{\tau} = \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \left( Y_{it} - \bar{Y}_{it}(0) \right) \]  

(12)

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

\[ \bar{Y}_{it}(0) = Y_{it-1} + \frac{1}{|\mathcal{N}_{it}^{\text{DiD}}|} \sum_{(i', t') \in \mathcal{N}_{it}^{\text{DiD}}} Y_{i't} - \frac{1}{|\mathcal{A}_{it}^{\text{DiD}}|} \sum_{(i', t') \in \mathcal{A}_{it}^{\text{DiD}}} Y_{i't'}. \]  

(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

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, 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. 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 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,

\[
\hat{\tau} = \hat{\beta}_{WFE2} = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \{(Y_{it} - Y_{i}^* - Y_{t}^* + Y^*) - \beta(X_{it} - X_{i}^* - X_{t}^* + X^*)\}^2
\]

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

\[
W_{it} = \sum_{i'=1}^{N} \sum_{t'=1}^{T} D_{i't'} \cdot w_{it'i'}\]

\[
w_{it'i'} = \begin{cases} 
1 & \text{if } (i, t) = (i', t') \\
1/|M_{DiD}^{i'i'}| & \text{if } (i, t) \in M_{DiD}^{i'i'} \\
1/|N_{DiD}^{i'i'}| & \text{if } (i, t) \in N_{DiD}^{i'i'} \\
(2X_{it} - 1)(2X_{i't'} - 1)/|A_{DiD}^{i'i'}| & \text{if } (i, t) \in A_{DiD}^{i'i'} \\
0 & \text{otherwise} 
\end{cases}
\]

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.

Thus, Theorem 1 shows that the multi-period DiD estimator is not generally equivalent to the 2FE regression estimator. The algebraic equivalence result also shows that some control observations will have negative regression weights when they frequently enter into the adjustment set, \(A_{DiD}^{i'i'}\), for multiple treated observations (i.e., \((2X_{it} - 1)(2X_{i't'} - 1) = -1\)). All together, the results of this section shows that the 2FE regression does not correspond to the difference-in-differences design unless the linear additive effect assumption is appropriate.

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, the general multi-period difference-in-differences estimator is equivalent to the weighted two-way fixed effects regression estimator, where some observations have 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. In particular, 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). We have also generalized the multi-period difference-in-differences estimator introduced in this paper and propose matching and weighting methods that are applicable to panel data (Imai et al. 2018). More research is needed to improve the existing methods for causal inference with panel data.
References


## Supplemental Appendix

### A Proof of Proposition 1

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

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} \{X_{it}(Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}) - (1 - X_{it})(Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y})\} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ X_{it} \left(1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT}\right) - \left(\frac{1}{T} \sum_{t' \neq t} Y_{it'} - \frac{1}{NT} \sum_{t' \neq t} Y_{it'}\right) \right. \\
\left. - \left(\frac{1}{T} \sum_{t' \neq t} Y_{it'} - \frac{1}{NT} \sum_{t' \neq t} Y_{it'}\right) + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} Y_{it'} \right] \\
-(1 - X_{it}) \left[ Y_{it} \left(1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT}\right) - \left(\frac{1}{T} \sum_{t' \neq t} Y_{it'} - \frac{1}{NT} \sum_{t' \neq t} Y_{it'}\right) \right. \\
\left. - \left(\frac{1}{T} \sum_{t' \neq t} Y_{it'} - \frac{1}{NT} \sum_{t' \neq t} Y_{it'}\right) + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} Y_{it'} \right] \\
= \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ X_{it} \left(\frac{(N-1)(T-1)}{NT} Y_{it} - \frac{N-1}{NT} \sum_{t' \neq t} Y_{it'} - \frac{T-1}{NT} \sum_{t' \neq t} Y_{it'} + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} Y_{it'} \right) \right. \\
\left. -(1 - X_{it}) \left(\frac{(N-1)(T-1)}{NT} Y_{it} - \frac{N-1}{NT} \sum_{t' \neq t} Y_{it'} - \frac{T-1}{NT} \sum_{t' \neq t} Y_{it'} + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} Y_{it'} \right) \right] \\
= \frac{(T-1)(N-1)}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ X_{it} \left(\frac{T}{T-1} Y_{it} - \sum_{t' \neq t} Y_{it'} + \frac{N}{T-1} Y_{it'} - \frac{N}{T-1} \sum_{t' \neq t} Y_{it'} \right) \right. \\
\left. -(1 - X_{it}) \left(\frac{T}{T-1} Y_{it} - \sum_{t' \neq t} Y_{it'} + \frac{N}{T-1} Y_{it'} - \frac{N}{T-1} \sum_{t' \neq t} Y_{it'} \right) \right] \\
= \frac{(T-1)(N-1)}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\bar{Y}_{it}(1) - \bar{Y}_{it}(0) \right) \right)
\]

The second algebraic equality we prove is the following,

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} \{X_{it}(X_{it} - \bar{X}_i - \bar{X}_t + \bar{X}) - (1 - X_{it})(X_{it} - \bar{X}_i - \bar{X}_t + \bar{X})\} = \sum_{i=1}^{N} \sum_{t=1}^{T} \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. \\
\left. - \left(\frac{1}{T} \sum_{t' \neq t} X_{it'} - \frac{1}{NT} \sum_{t' \neq t} X_{it'}\right) + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} X_{it'} \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. \\
\left. - \left(\frac{1}{T} \sum_{t' \neq t} X_{it'} - \frac{1}{NT} \sum_{t' \neq t} X_{it'}\right) + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} X_{it'} \right] \\
= \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_{t' \neq t} X_{it'} + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} X_{it'} \right) \right. \\
\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_{t' \neq t} X_{it'} + \frac{1}{NT} \sum_{t' \neq t} \sum_{t'' \neq t} X_{it'} \right) \right] \\
= \frac{(T-1)(N-1)}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\bar{X}_{it}(1) - \bar{X}_{it}(0) \right) \right)
We first establish the following equality.

Proof of Proposition 2

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

\[
\hat{\beta}_{FE2} = \frac{\sum_{t=1}^{T} \sum_{t'=1}^{T} (X_{it} - X_i - X_t + X)(Y_{it} - Y_i - Y_t + Y)}{\sum_{t=1}^{T} \sum_{t'=1}^{T} (X_{it} - X_i - X_t + X)^2} 
\]

\[
= \frac{\sum_{t=1}^{T} \sum_{t'=1}^{T} X_{it} Y_{it} - T \sum_{i=1}^{N} X_i Y_i - N \sum_{t=1}^{T} X_t Y_t + NT XX}{\sum_{t=1}^{T} \sum_{t'=1}^{T} (X_{it} - X_i - X_t + X)^2}
\]

\[
= \frac{\sum_{t=1}^{T} \sum_{t'=1}^{T} \{X_{it} (Y_{it} - Y_i - Y_t + Y) - (1 - X_{it}) (Y_{it} - Y_i - Y_t + Y)\}}{\sum_{t=1}^{T} \sum_{t'=1}^{T} (X_{it} - X_i - X_t + X)}
\]

\[
= \frac{1}{K} \left\{ \frac{1}{NT} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left( \bar{Y}_{it} - \bar{Y}_{i}(0) \right) \right\}
\]

where the last equality follows from equation (14) and (15).

B Proof of Proposition 2

Proof We first establish the following equality.

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} = \sum_{i=1}^{N} \sum_{t'=1}^{T} \sum_{t=1}^{T} \left( \sum_{i=1}^{N} w_{it}^{i'} \right) = \sum_{t'=1}^{T} \sum_{t=1}^{T} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} X_{i't'} w_{it}^{i'} + (1 - X_{i't'}) w_{it}^{i'} \right)
\]

\[
= \sum_{t'=1}^{T} \sum_{t=1}^{T} D_{i't'} \left\{ X_{i't'} \left( \frac{\#A^*_{i't'} + a_{i't'}}{\#A^*_{i't'} + a_{i't'} + \#M_{i't'} + \#N_{i't'} + \#A_{i't'} + a_{i't'}} + \frac{\#M_{i't'} + \#N_{i't'}}{\#A^*_{i't'} + a_{i't'}} \right) + \frac{\#A^*_{i't'} + a_{i't'}}{\#A^*_{i't'} + a_{i't'} + \#M_{i't'} + \#N_{i't'} + \#A_{i't'} + a_{i't'}} \right) 
\]

\[
= \sum_{t'=1}^{T} \sum_{t=1}^{T} D_{i't'} (2X_{i't'} + 2(1 - X_{i't'})) = 2 \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}.
\]
The unit itself gets weight equal to \( \frac{D_{i,t} \#K_{i,t}}{\#M_{i,t}^*} \) with the same treatment status with \((i', t')\) and \(A_{i,t}^* - a_{i,t'}\) observations with the same treatment status with \((i', t')\). The final matched set \(A_{i,t}^*\) is composed of \(a_{i,t'}\) observations with \( \#A_{i,t}^* + a_{i,t'} \) while the latter is weighted by \( \frac{D_{i,t'} \#A_{i,t'}^*}{\#A_{i,t'}^* + a_{i,t'}} \). All the other observations will get zero weight.

Following the same logic from above, it is straightforward to show that \( X_t^* = X_t - X_t^* + X^* \) and thus

\[
X_t - X_t^* + X^* = \begin{cases} 
\frac{1}{2} & \text{if } X_{it} = 1 \\
-\frac{1}{2} & \text{if } X_{it} = 0 
\end{cases}
\]

(17)

For instance,

\[
X^* = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} X_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}} \\
= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} w_{it}^{i,t'} \right)}{2 \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \\
= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{i,t'} X_{i,t'} \left( \frac{\#A_{i,t'}^*}{\#A_{i,t'}^* + a_{i,t'}} + \frac{a_{i,t'}}{\#A_{i,t'} + a_{i,t'}} \right) + D_{i,t'} (1 - X_{i,t'}) \left( \frac{\#A_{i,t'}^*}{\#A_{i,t'}^* + a_{i,t'}} - \frac{-a_{i,t'}}{\#A_{i,t'}^* + a_{i,t'}} \right)}{2 \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}} \\
= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{i,t} \sum_{i=1}^{T} (2X_{it} - 1)(Y_{it} - Y_{i}^* - Y_{t}^* + Y^*)}{2 \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}}
\]

We can derive the desired result.

\[
\hat{\beta}_{WFE2} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(X_{it} - X_{i}^* - X_t^* + X^*) (Y_{it} - Y_{i}^* - Y_{t}^* + Y^*)}{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(X_{it} - X_{i}^* - X_t^* + X^*)^2} \\
= \frac{1}{2} \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(2X_{it} - 1)(Y_{it} - Y_{i}^* - Y_{t}^* + Y^*)}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \sum_{i=1}^{T} (2X_{it} - 1)Y_{it}} \\
= \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \sum_{i=1}^{T} (2X_{it} - 1)Y_{it}} \\
= \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \sum_{i=1}^{T} (2X_{it} - 1)Y_{it}} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \sum_{i=1}^{T} w_{i,t'}^{i,t'} \right) (2X_{it} - 1)Y_{it} \right\} \\
= \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{i,t'} \sum_{i=1}^{T} \sum_{t=1}^{T} \left( \sum_{i=1}^{T} w_{i,t'}^{i,t'} (2X_{it} - 1)Y_{it} \right) + (1 - X_{i,t'}) \left( \sum_{i=1}^{N} \sum_{t=1}^{T} w_{i,t'}^{i,t'} (2X_{it} - 1)Y_{it} \right)} \\
= \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} D_{i,t'} \sum_{i=1}^{T} \sum_{t=1}^{T} \left( \sum_{i=1}^{T} w_{i,t'}^{i,t'} (2X_{it} - 1)Y_{it} \right) + (1 - X_{i,t'}) \left( \sum_{i=1}^{N} \sum_{t=1}^{T} w_{i,t'}^{i,t'} (2X_{it} - 1)Y_{it} - \frac{Y_{it}(1) - Y_{it}(0)}{\#_{match2}} \right)}
where the second and third equality follows from equation (10) and (17). The last two equalities follow from applying the definition of $K_{it}$, $W_{it}$, $Y_{it}(1)$ and $Y_{it}(0)$ given in Proposition 2.

\[ \hat{\gamma}_{\text{DD}} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(X_{it} - X_{i}^{*} - X_{i}^{*})Y_{it} - Y_{it}^{*} + Y_{it}^{*})}{\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(X_{it} - X_{i}^{*} + X_{i}^{*})^2} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(2X_{it} - 1)(Y_{it} - Y_{it}^{*} - Y_{it}^{*} + Y_{it}^{*}) \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(2X_{it} - 1)Y_{it} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left( \sum_{j=1}^{N} \sum_{t'=1}^{T} w_{it}^{j,t'} \right) (2X_{it} - 1)Y_{it} \right\} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i} \left\{ X_{i',t'} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it}^{i',t'} (2X_{it} - 1)Y_{it} \right) + (1 - X_{i',t'}) \left( \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it}^{i',t'} (2X_{it} - 1)Y_{it} \right) \right\} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i} \left\{ X_{i',t'} \left( Y_{i',t'} - Y_{i',t'}^{*} - Y_{i',t'}^{*} + Y_{i',t'}^{*} \right) \right\} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i} \left\{ X_{i',t'} \left( Y_{i',t'} - Y_{i',t'}^{*} - Y_{i',t'}^{*} + Y_{i',t'}^{*} \right) \right\} \]

\[ \hat{\gamma}_{\text{DD}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i} \left\{ X_{i',t'} \left( Y_{i',t'} - Y_{i',t'}^{*} - Y_{i',t'}^{*} + Y_{i',t'}^{*} \right) \right\} \]

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