When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?*

Kosuke Imai† In Song Kim‡

August 19, 2016

Abstract

Many social scientists use linear fixed effects regression models for causal inference with longitudinal data to account for unobserved time-invariant confounders. We show that these models require two additional causal assumptions, which are not necessary under an alternative selection-on-observables approach. Specifically, the models assume that past treatments do not directly influence current outcome, and past outcomes do not directly affect current treatment. The assumed absence of causal relationships between past outcomes and current treatment may also invalidate some applications of before-and-after and difference-in-differences designs. Furthermore, we propose a new matching framework to further understand and improve one-way and two-way fixed effects regression estimators by relaxing the linearity assumption. Our analysis highlights a key trade-off — the ability of fixed effects regression models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationships between treatment and outcome.

Key Words: before-and-after design, difference-in-differences design, matching, panel data, synthetic control method, 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 (http://cran.r-project.org/package=wfe). This paper subsumes an earlier version of the paper entitled "On the Use of Linear Fixed Effects Regression Estimators for Causal Inference." We thank Alberto Abadie, Mike Bailey, Neal Beck, Matias Cattaneo, Naoki Egami, Erin Hartman, Danny Hidalgo, Yuki Shiraito, and Teppei Yamamoto for helpful comments.

†Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

‡Assistant 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

Linear fixed effects regression models are a primary workhorse for causal inference with longitudinal or panel data in the social sciences (e.g., Angrist and Pischke, 2009). Many researchers use these models in order to draw causal inference from observational data by adjusting for unobserved time-invariant confounders. In spite of this widespread practice, almost all methodological discussions of fixed effects models in political science have taken place from a statistical modeling perspective with little attention to the issues directly related to causal inference (e.g., Beck, 2001; Plümper and Troeger, 2007; Wilson and Butler, 2007; Bell and Jones, 2015; Clark and Linzer, 2015). While we acknowledge the importance of modeling issues, this paper focuses on the causal aspects of fixed effects regression models that are often overlooked by applied researchers.

Specifically, we show that the fixed effects regression models require two additional causal assumptions, which are not necessary under an alternative selection-on-observables approach (Section 2). These models assume that past treatments do not directly influence current outcome and that past outcomes do not directly affect current treatment. In many applications, these additional assumptions may not be credible. In particular, the assumed absence of causal relationships between past outcomes and current treatment may invalidate some applications of popular before-and-after (BA) and difference-in-differences (DiD) designs. Our analysis highlights a key trade-off: the ability of fixed effects regression models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationships between treatment and outcome.

In addition, we propose a novel analytical framework that directly connects these models to matching estimators (Section 3). This framework makes it clear that under linear regression model with unit fixed effects, the treated and control observations are compared across time periods within the same unit in order to account for unobserved time-invariant confounders. Our framework also enables us to develop nonparametric within-unit matching estimators, which relax the linearity assumption and flexibly incorporate various identification strategies. Despite these improvements, the proposed within-

unit matching estimators, like linear fixed effects models, must assume the absence of dynamic causal relationships between treatment and outcome.

We further extend our analysis to the linear regression models with unit and time fixed effects (Section 4). Because for any given observation, no other observation share the same unit and time, it is impossible to nonparametrically adjust for unobserved time-invariant and unit-invariant confounders at the same time. One important exception, however, is the DiD design. Although many researchers use the linear two-way fixed effects regression estimator to implement the DiD estimator (e.g., Bertrand *et al.*, 2004), the general equivalence between the two estimators holds only in the case of two time periods with the treatments administered in the second time period alone. To remedy this problem, we propose a nonparametric multi-period DiD estimator.

Despite the intuitive appeal of the DiD estimators, however, they share the same causal identification assumptions as linear fixed effects regression estimators. In particular, if a difference in the baseline outcome between the treatment and control groups reflects the direct causal effect of the baseline outcome on the treatment assignment, then the validity of the DiD design is compromised. On the other hand, if the difference in the baseline outcome arises from the existence of unobserved time-invariant confounders, then the DiD design remains valid. As before, it is not possible to simultaneously adjust for the baseline outcome difference and unobserved time-invariant confounders. Our argument is also applied to the synthetic control method (Abadie *et al.*, 2010).

We illustrate the proposed methodology by revisiting the controversy about whether GATT (General Agreement on Tariffs and Trade) membership increases international trade (Section 5). We show that the empirical results are sensitive to the underlying causal assumptions, underscoring the importance of causal assumptions when applying fixed effects regression models. The final section provides concluding remarks and suggestions for applied researchers. The open-source software, `wfe: Weighted Linear Fixed Effects Estimators for Causal Inference`, is available as an R package for implementing the proposed methods.

Our work builds upon a small literature on the use of linear fixed effects models for causal inference

with longitudinal data in econometrics and statistics (e.g., Wooldridge, 2005a; Sobel, 2006). Our theoretical results also extend the weighted regression results available in the literature for causal inference with cross-section data to longitudinal studies (e.g., Humphreys, 2009; Aronow and Samii, 2015; Solon *et al.*, 2015). We also contribute to the literature on matching methods (e.g., Rubin, 2006; Ho *et al.*, 2007; Stuart, 2010). Despite their popularity, matching methods are almost exclusively used in the analysis of cross-section data. Little work has been done to develop and apply them for analysis of longitudinal data (see Li *et al.*, 2001, for an exception).

2 Regression Models with Unit Fixed Effects

We study the causal assumptions of regression models with unit fixed effects. While we begin by describing the basic linear regression model with unit fixed effects, our analysis is conducted under a more general, nonparametric setting based on the directed acyclic graphs (DAGs) and potential outcomes frameworks (Pearl, 2009; Imbens and Rubin, 2015).

2.1 The Linear Unit Fixed Effects Regression Model

Throughout this paper, for the sake of simplicity, we assume a balanced longitudinal data set of N units and T time periods with no missing data. We also assume a simple random sampling of units with T fixed. For each unit i at time t, we observe the outcome variable Y_{it} and the binary treatment variable $X_{it} \in \{0, 1\}$. The most basic linear regression model with unit fixed effects is based on the following linear specification.

ASSUMPTION 1 (LINEARITY) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$,*

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

where α_i is a fixed but unknown intercept for unit i and ϵ_{it} is a disturbance term for unit i at time t.

In this model, the unit fixed effect α_i captures a vector of unobserved time-invariant confounders in a flexible manner. Formally, we define each fixed effect as $\alpha_i = h(\mathbf{U}_i)$ where $\mathbf{U}_i$ represents a vector of unobserved time-invariant confounders and $h(\cdot)$ is an arbitrary and unknown function.

Typically, the strict exogeneity of the disturbance term ϵ_{it} is assumed to identify β.

ASSUMPTION 2 (STRICT EXOGENEITY) *For each* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

$$\epsilon_{it} \quad \perp\!\!\!\perp \quad \{\mathbf{X}_i, \mathbf{U}_i\}$$

where $\mathbf{X}_i$ *is a* $T \times 1$ *vector of treatment variables for unit* i

For the sake of generality, the assumption is expressed in terms of statistical independence although technically only the mean independence, i.e., $\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{U}_i) = \mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \alpha_i) = \mathbb{E}(\epsilon_{it})$, is required for the identification of β.

We refer to this model based on Assumptions 1 and 2 as LM-FE. The least squares estimate of β is obtained by regressing the deviation of the outcome variable from its mean on the deviation of the treatment variable from its mean,

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

where $\overline{X}_i = \sum_{t=1}^{T} X_{it}/T$ and $\overline{Y}_i = \sum_{t=1}^{T} Y_{it}/T$. If the data are generated according to LM-FE, then $\hat{\beta}_{\mathsf{FE}}$ is unbiased for β.

The parameter β is interpreted as the average contemporaneous effect of X_{it} on Y_{it}. Formally, let $Y_{it}(x)$ represent the potential outcome for unit i at time t under the treatment status $X_{it} = x$ for $x = 0, 1$ where the observed outcome equals $Y_{it} = Y_{it}(X_{it})$. Equation (2) shows that units with no variation in the treatment variable do not contribute to the estimation of β. Thus, under LM-FE, the causal estimand is the following average treatment effect among the units with some variation in the treatment status.

$$\tau \quad = \quad \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid C_i = 1) \tag{3}$$

where $C_i = \mathbf{1}\{0 < \sum_{t=1}^{T} X_{it} < T\}$. Under LM-FE, this quantity is represented by β, i.e., $\beta = \tau$, because of the assumed linearity for potential outcomes (Assumption 1), i.e., $Y_{it}(x) = \alpha_i + \beta x + \epsilon_{it}$.

2.2 Causal Assumptions

We examine causal assumptions of LM-FE using the two causal inference frameworks: Directed Acyclic Graphs (DAGs) and potential outcomes. Pearl (2009) shows that a DAG can formally represent a non-

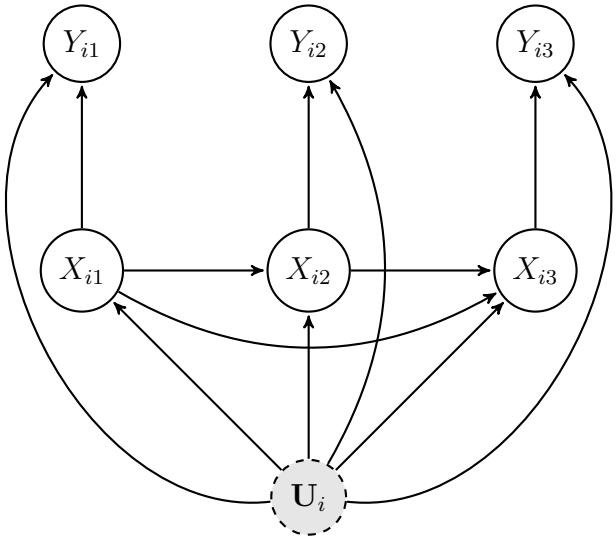


Figure 1: Directed Acyclic Graph for Regression Models with Unit Fixed Effects based on Three Time Periods. Solid circles represent observed outcome Y_{it} and treatment X_{it} variables whereas a grey dashed circle represents a vector of unobserved time-invariant confounders $\mathbf{U}_i$. The solid arrows indicate the possible existence of causal relationships whereas the absence of such arrows represents the lack of causal relationships. DAGs are also assumed to contain all relevant, observed and unobserved, variables.

parametric structural equation model (NPSEM), enabling researchers to derive the causal assumptions and testable implications without imposing parametric assumptions. We consider the following NPSEM,

$$Y_{it} = g_1(X_{it}, \mathbf{U}_i, \epsilon_{it}) \tag{4}$$

$$X_{it} = g_2(X_{i1}, \ldots, X_{i,t-1}, \mathbf{U}_i, \eta_{it}) \tag{5}$$

where η_{it} is the exogenous disturbance term. We call this model NPSEM-FE. Unlike LM-FE, NPSEM-FE does not assume a functional form and enables all effects to vary across observations (Pearl, 2009).

The DAG in Figure 1 graphically represents NPSEM-FE. In this DAG, the observed variables, X_{it} and Y_{it}, are represented by solid circles whereas a dashed grey circle represents the unobserved time-invariant confounders, $\mathbf{U}_i$.[1] The solid black arrows indicate the possible existence of direct causal effects whereas the absence of such arrows represents the assumption of no direct causal effect. In addition, DAGs are assumed to contain all relevant, observed and unobserved, variables. Therefore, this DAG assumes the absence of unobserved time-varying confounder. Finally, following the convention, we omit

[1]For simplicity, the DAG only describes the causal relationships for three time periods, but we assume that the same relationships apply to all time periods even when there are more than three time periods, i.e., $T > 3$.

exogenous disturbance terms, ϵ_{it} and η_{it}, from DAGs.

We show that NPSEM-FE is a nonparametric generalization of LM-FE. That is, no additional arrows can be added to the DAG without making NPSEM-FE inconsistent with LM-FE. First, note that equation (4) includes Assumption 1 as a special case, i.e., $g_1(X_{it}, \mathbf{U}_i, \epsilon_{it}) = h(\mathbf{U}_i) + \beta X_{it} + \epsilon_{it}$. Both models assume that neither past treatments nor past outcomes directly affect current outcome because these variables are not included as covariates. Second, while LM-FE does not directly specify the data generating process for the treatment variable X_{it}, Assumption 2 also holds under NPSEM-FE.[2] Moreover, no additional arrows that point to X_{it} can be included in the DAG without violating Assumption 2. The existence of any such arrow, which must originate from past outcomes $Y_{it'}$ where $t' < t$, would imply a possible correlation between $\epsilon_{it'}$ and X_{it}.

In sum, LM-FE and its nonparametric generalization NPSEM-FE require the following assumptions, each of which is represented by the absence of corresponding arrows: no unobserved time-varying confounder exists (Assumption (a)), past outcomes do not directly affect current outcome (Assumption (b)), past outcomes do not directly affect current treatment (Assumption (c)), and past treatments do not directly affect current outcome (Assumption (d)).

Next, we adopt the potential outcomes framework. While DAGs illuminate the entire causal structure, the potential outcomes framework clarifies the assumptions about treatment assignment mechanisms. First, the right hand sides of equations (1) and (4) include the contemporaneous value of the treatment but not its past values, implying that past treatments do not directly affect current outcome. We call this restriction the assumption of no carryover effect,[3] corresponding to Assumption (d) described above.

ASSUMPTION 3 (NO CARRYOVER EFFECT) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the potential outcome is given by,*

$$Y_{it}(X_{i1}, X_{i2}, \ldots, X_{i,t-1}, X_{it}) \;=\; Y_{it}(X_{it})$$

[2]This is because Y_{it} acts as a collider on any path between ϵ_{it} and $\{\mathbf{X}_i, \mathbf{U}_i\}$.

[3]These models are based on the usual assumption of no spillover effect that the outcome of a unit is not affected by the treatments of other units (Rubin, 1990). The assumption of no spillover effect is made throughout this paper.

To better understand the assumed treatment assignment mechanism, we consider a randomized experiment, in which Assumption 2 is satisfied: for any given unit i, we randomize the treatment X_{i1} at time 1, and for the next time period 2, we randomize the treatment X_{i2} conditional on the realized treatment at time 1, i.e., X_{i1}. More generally, at time t, we randomize the current treatment X_{it} conditional on the past treatments $X_{i1}, X_{i2}, \ldots, X_{i,t-1}$. The critical assumptions are that there exists no unobserved time-varying confounder (Assumption (a)) and that the treatment assignment probability at time t cannot depend on its past realized outcomes $Y_{it'}$ where $t' < t$ (Assumption (c)). However, the treatment assignment probability may vary across units as a function of unobserved time-invariant characteristics $\mathbf{U}_i$. We formalize this treatment assignment mechanism as follows.

ASSUMPTION 4 (SEQUENTIAL IGNORABILITY WITH UNOBSERVED TIME-INVARIANT CONFOUNDERS) *For each* $i = 1, 2, \ldots, N$,

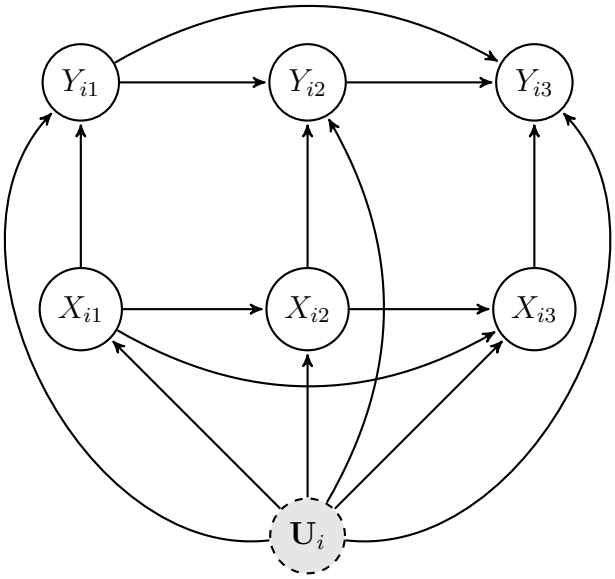
$$
\begin{aligned}
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i1} \mid \mathbf{U}_i \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{it'} \mid X_{i1}, \ldots, X_{i,t'-1}, \mathbf{U}_i \\
&\vdots \\
\{Y_{it}(1), Y_{it}(0)\}_{t=1}^{T} &\perp\!\!\!\perp X_{iT} \mid X_{i1},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

In Appendix A.1, we prove that under LM-FE, Assumption 4 is equivalent to Assumption 2. Thus, Assumption 3 corresponds to Assumption (d) of NPSEM-FE and Assumption 1 of LM-FE whereas Assumption 4 corresponds to Assumptions (a) and (c) of NPSEM-FE and Assumption 2 of LM-FE.

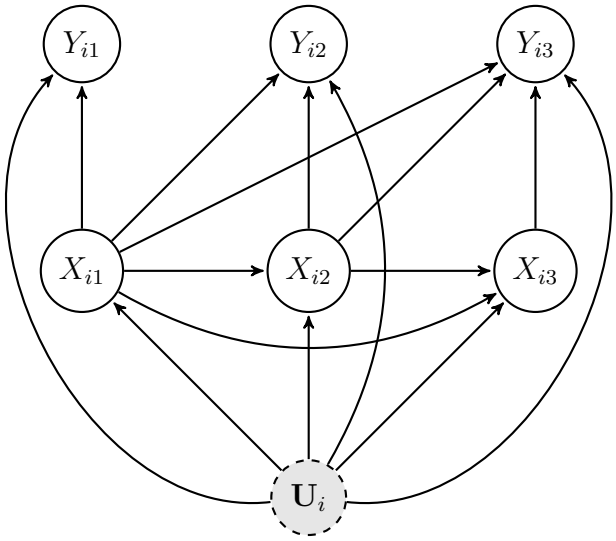
2.3 Relaxing the Causal Assumptions

It is well known that the assumption of no unobserved time-varying confounder (Assumption (a)) is difficult to relax. Therefore, we consider the other three identification assumptions shared by LM-FE and NPSEM-FE (Assumptions (b), (c), and (d)) in turn.

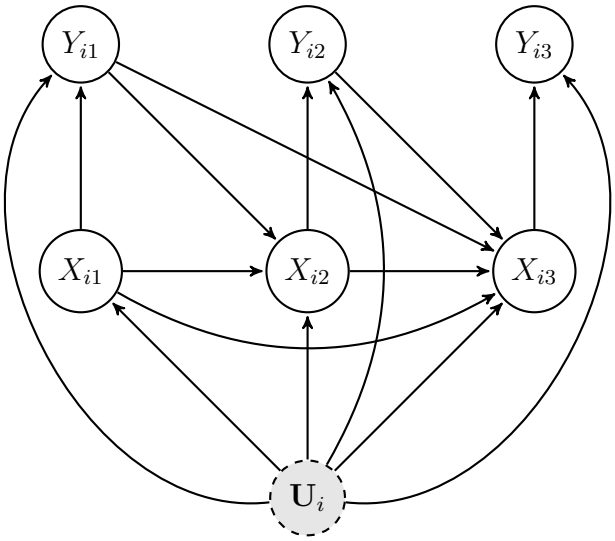
We did not mention Assumption (b) under the potential outcomes framework. Indeed, this assumption — past outcomes do not directly affect current outcome — can be relaxed without compromising identification. To see this, suppose that past outcomes directly affect current outcome as in Figure 2(a).



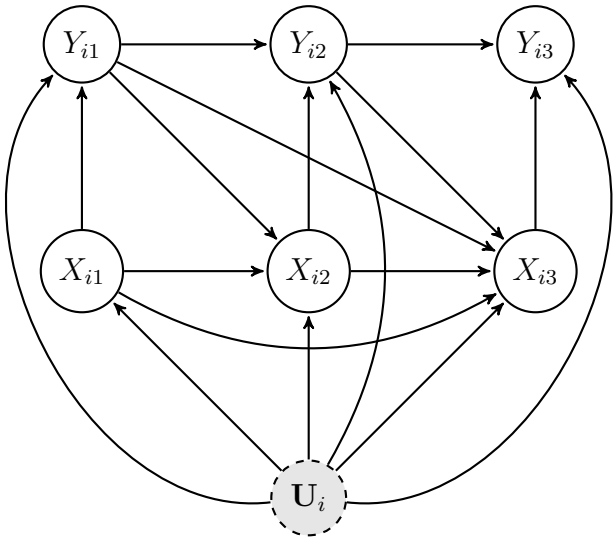
(a) past outcome affects current outcome



(b) past treatments affect current outcome



(c) past outcomes affect current treatment



(d) instrumental variables

Figure 2: Directed Acyclic Graphs with the Relaxation of Various Identification Assumptions of Regression Models with Unit Fixed Effects (shown in Figure 1). Identification is not compromised when past outcomes affect current outcome (panel (a)). However, the other two scenarios (panels (b) and (c)) violate the strict exogeneity assumption. To address the possible violation of strict exogeneity shown in panel (c), researchers often use an instrumental variable approach shown in panel (d).

Even in this scenario, past outcomes do not confound the causal relationship between current treatment and current outcome so long as we condition on past treatments and unobserved time-invariant confounders. The reason is that past outcomes do not directly affect current treatment. Thus, there is no need to adjust for past outcomes even when they directly affect current outcome.[4]

Next, we entertain the scenario in which past treatments directly affect current outcome. Typically, applied researchers address this possibility by including lagged treatment variables in LM-FE. Here, we consider the following model with one period lag.

$$Y_{it} \quad = \quad \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \epsilon_{it} \tag{6}$$

The model implies that the potential outcome can be written as a function of the contemporaneous and previous treatments, i.e., $Y_{it}(X_{i,t-1}, X_{it})$, rather than the contemporaneous treatment alone, slightly relaxing Assumption 3.

The DAG in Figure 2(b) generalizes the above model and depicts an NPSEM where a treatment possibly affects all future outcomes as well as current outcome. This NPSEM is a modification of NPSEM-FE replacing equation (5) with the following alternative model for the outcome,

$$Y_{it} \quad = \quad g_1(X_{i1}, \ldots, X_{it}, \mathbf{U}_i, \epsilon_{it}) \tag{7}$$

It can be shown that under this NPSEM Assumption 4 still holds.[5] The only difference between the DAGs in Figures 1 and 2(b) is that in the latter we must adjust for the past treatments because they confound the causal relationship between the current treatment and outcome.

In general, however, we cannot nonparametrically adjust for all past treatments and unobserved time-invariant confounders $\mathbf{U}_i$ at the same time. By nonparametric adjustment, we mean that researchers match exactly on confounders. To nonparametrically adjust for $\mathbf{U}_i$, the comparison of treated and control

[4]The application of the adjustment criteria (Shpitser *et al.*, 2010) implies that these additional causal relationships do not violate Assumption 4 since every non-causal path between the treatment X_{it} and any outcome $Y_{it'}$ is blocked where $t \neq t'$.

[5]The result follows from the application of the adjustment criteria (Shpitser *et al.*, 2010) where any non-causal path between ϵ_{it} and $\{\mathbf{X}_i, \mathbf{U}_i\}$ contains a collider Y_{it}.

observations must be done across different time periods within the same unit. The problem is that no two observations within a unit, measured at different time periods, share the same treatment history. Such adjustment must be done by comparing observations across units within the same time period, and yet, doing so makes it impossible to adjust for unobserved time-invariant variables.

Therefore, in practice, researchers assume that only a small number of past treatments matter. Under this assumption, multiple observations within the same unit may share the identical but partial treatment history even though they are measured at different points in time. Under the linear regression framework, researchers conduct a parametric adjustment by simply including a small number of past treatments as done in equation (6). However, the number of lagged treatments to be included is arbitrarily chosen and is rarely justified on a substantive ground.[6]

Finally, we consider relaxing the assumption that past outcomes do not directly affect current treatment. This scenario is depicted as Figure 2(c). It is immediate that Assumptions 4 is violated because the existence of causal relationships between past outcomes and current treatment implies a correlation between past disturbance terms and current treatment.[7] This lack of feedback effects over time represents another key causal assumption required for LM-FE.

To address this issue, the model that has attracted much attention is the following linear fixed effects model with a lagged outcome variable,

$$Y_{it} \quad = \quad \alpha_i + \beta X_{it} + \rho Y_{i,t-1} + \epsilon_{it} \tag{8}$$

Recall that if past outcomes do not affect current treatment, it is unnecessary to adjust for past outcomes even if they affect current outcome. Figure 2(d) presents a DAG that corresponds to this model. The standard identification strategy commonly employed for this model is based on instrumental variables (e.g., Arellano and Bond, 1991). Applying the result of Brito and Pearl (2002) shows that we can identify

[6]One exception is the setting where the treatment status changes only once in the same direction, e.g., from the control to treatment condition. While adjusting for the previous treatment is sufficient in this case, there may exist a time trend in outcome, which confounds the causal relationship between treatment and outcome.

[7]For example, there is a unblocked path from X_{i2} to ϵ_{i1} through Y_{i1}.

the average causal effect of X_{i3} on Y_{i3} by using X_{i1}, X_{i2}, and Y_{i1} as instrumental variables conditional on $\mathbf{U}_i$ and Y_{i2}.[8] However, the validity of each instrument depends on the assumed absence of its direct causal effect on the outcome variable (i.e., direct effects of Y_{i1}, X_{i1}, and X_{i2} on Y_{i3}). Unfortunately, these assumptions are often made without a substantive justification.

In sum, three key identification assumptions are required for LM-FE and its nonparametric generalization NPSEM-FE. The assumption of no unobserved time-varying confounder is well appreciated by applied researchers. However, many fail to recognize two additional assumptions required for unit fixed effects regression models: past treatments do not affect current outcome and past outcomes do not affect current treatment. The former can be partially relaxed by assuming that only a small number of lagged treatment variables affect the outcome while the use of instrumental variables is a popular approach to relax the latter assumption. However, such approaches are rarely justified on substantive grounds.

2.4 Comparison with the Selection-on-Observables Approach

It is instructive to compare NPSEM-FE with a selection-on-observables approach, which does not permit any type of unobserved confounders. A prominent causal model based on this alternative approach is the marginal structural models (MSMs) developed in epidemiology (Robins *et al.*, 2000; Blackwell, 2013). Unlike NPSEM-FE, the MSMs assume the absence of unobserved time-invariant confounders as well as that of unobserved time-varying confounders. However, the MSMs are able to relax the other two assumptions required for NPSEM-FE.

Figure 3 presents a DAG for the MSMs. First, the MSMs relax Assumption 3, allowing past treatments to directly affect current outcome. These causal relationships are represented by the arrows pointing from past treatments to current outcome in the DAG. Thus, the potential outcome at time t for unit i can be written as a function of the unit's entire treatment sequence up to that point in time, i.e., $Y_{it}(x_1, \ldots, x_t)$ for a given treatment sequence $(X_{i1}, \ldots, X_{it}) = (x_1, \ldots, x_t)$.

Second, under the MSMs, past outcomes can affect current treatment as well as current outcome. In

[8] If X_{i1} and X_{i2} directly affect Y_{i3}, then only Y_{i1} can serve as a valid instrument. If past outcomes do not affect current outcome, then we can use both Y_{i1} and Y_{i2} as instruments.

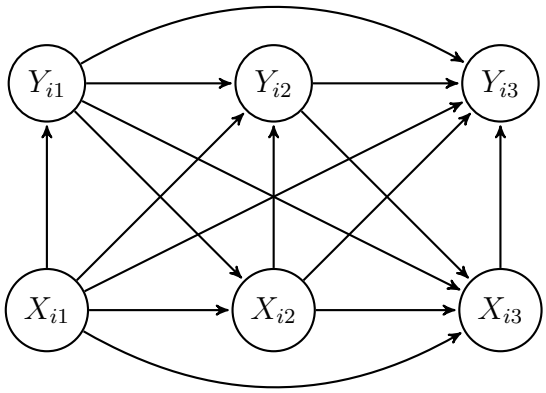


Figure 3: Directed Acyclic Graph for Marginal Structural Models (MSMs). When compared to regression models with unit fixed effects (Figure 1), MSMs assume the absence of unobserved time-invariant confounders $\mathbf{U}_i$ but relax the other assumptions by allowing the past treatments to affect the current outcome and the past outcomes to affect the current treatment.

the DAG, this scenario is represented by the arrows that point to current treatment from past outcomes. If we wish to identify the average contemporaneous treatment effect, we adjust for past outcomes. Using the potential outcomes framework, it can also be shown that under the following sequential ignorability assumption, the MSMs can identify the average outcome under any given treatment sequence, i.e., $\mathbb{E}(Y_{it}(x_1, \ldots, x_t))$, going beyond the average contemporaneous treatment effect.[9]

ASSUMPTION 5 (SEQUENTIAL IGNORABILITY WITH PAST OUTCOMES (ROBINS *et al.*, 2000)) *For* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

$$
\begin{aligned}
\{Y_{it}(x_1, \ldots, x_t)\}_{t=1}^{T} &\perp\!\!\!\perp X_{i1} \\
&\vdots \\
\{Y_{it}(x_1, \ldots, x_t)\}_{t=t'}^{T} &\perp\!\!\!\perp X_{it'} \mid X_{i1} = x_1, \ldots, X_{i,t'-1} = x_{t'-1}, Y_{i1}, \ldots, Y_{i,t'-1} \\
&\vdots \\
Y_{iT}(x_1, \ldots, x_T) &\perp\!\!\!\perp X_{iT} \mid X_{i1} = x_1, \ldots, X_{i,T-1} = x_{T-1}, Y_{i1}, \ldots, Y_{i,T-1}
\end{aligned}
$$

Unlike Assumption 4, Assumption 5 conditions on past outcomes as well as past treatments. However, under Assumption 5, we cannot adjust for unobserved time-invariant confounders $\mathbf{U}_i$.

Thus, the ability of LM-FE and its nonparametric generalization NPSEM-FE to adjust for unobserved time-invariant confounders comes at the cost: we must assume that past treatments do not di-

[9] We also assume that the treatment assignment probability at each time period for any unit is bounded away from 0 and 1.

rectly affect current outcome and past outcomes do not directly affect current treatment. In contrast, the selection-on-observables approach, while it cannot account for unobserved time-invariant confounders, can relax both of these assumptions and identify the average causal effect of an entire treatment sequence.

2.5 Adjusting for Observed Time-varying Confounders

Finally, we consider the adjustment of observed time-varying confounders under regression models with unit fixed effects. Since the unobserved confounders $\mathbf{U}_i$ must be time-invariant, applied researchers often adjust for a vector of observed time-varying covariates $\mathbf{Z}_{it}$ to improve the credibility of assumptions for regression models with unit fixed effects[10] In particular, it is a common practice to include these time-varying confounders in LM-FE,

$$Y_{it} \quad = \quad \alpha_i + \beta X_{it} + \boldsymbol{\delta}^\top \mathbf{Z}_{it} + \epsilon_{it} \tag{9}$$

The model is completed with the following version of strict exogeneity assumption,

ASSUMPTION 6 (STRICT EXOGENEITY WITH OBSERVED TIME-VARYING CONFOUNDERS) *For each* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$*,*

$$\epsilon_{it} \quad \perp\!\!\!\perp \quad \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i\}$$

where $\mathbf{Z}_i = (\mathbf{Z}_{i1}\ \mathbf{Z}_{i2}\ \ldots\ \mathbf{Z}_{iT})$.

Figure 4 presents a DAG that is a nonparametric generalization of the model given in equation (9) under Assumption 6. The difference between the DAGs shown in Figures 1 and 4 is the addition of $\mathbf{Z}_{it}$, which directly affects the contemporaneous outcome Y_{it}, current and future treatments $\{X_{it}, X_{i,t+1}, \ldots, X_{iT}\}$, and their own future values $\{\mathbf{Z}_{i,t+1}, \ldots, \mathbf{Z}_{iT}\}$. Moreover, the unobserved time-invariant confounders can directly affect these observed time-varying confounders. Under this NPSEM, only the contemporaneous time-varying confounders $\mathbf{Z}_{it}$ and the unobserved time-invariant confounders $\mathbf{U}_i$ confound the contemporaneous causal relationship between X_{it} and Y_{it}. Neither past treatments nor past time-varying confounders are confounders because they do not directly affect current outcome Y_{it}.[11]

[10] Note that $\mathbf{Z}_{it}$ is assumed to be causally prior to the current treatment X_{it}.

[11] Appendix A.2 consider the potential outcome representation of this model where the treatment assignment mechanism

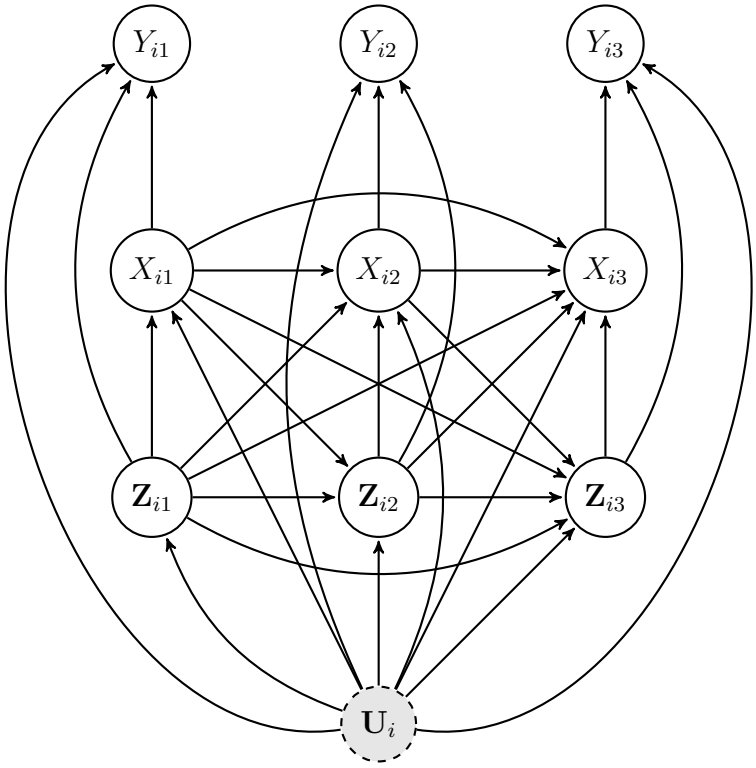


Figure 4: Directed Acyclic Graph for Regression Models with Unit Fixed Effects and Observed Time-varying Confounders based on Three Time Periods.

Now, suppose that the observed time-varying confounders $\mathbf{Z}_{it}$ directly affect future and current outcomes $Y_{it'}$ where $t' \geq t$. In this case, we need to adjust for the past values of the observed time-varying confounders as well as their contemporaneous values. This can be done by including the relevant lagged confounding variables in regression models with unit fixed effects. However, for the same reason as the one explained in Section 2.3, it is impossible to nonparametrically adjust for the entire sequence of past time-varying confounders and unobserved time-invariant confounders $\mathbf{U}_i$ at the same time. The problem is that while the nonparametric adjustment of $\mathbf{U}_i$ requires the comparison of observations across different time periods within each unit, no two observations measured at different points in time share an identical history of time-varying confounders.

Furthermore, similar to the case of NPSEM-FE, the average contemporaneous treatment effect of

is formalized as the sequential ignorability with unobserved time-invariant and observed time-varying confounders (see Assumption 10). The assumption makes it clear that the treatment assignment cannot depend on past outcomes.

X_{it} on Y_{it} becomes unidentifiable if the outcome Y_{it} affects future treatments $X_{it'}$ either directly or indirectly through $\mathbf{Z}_{it'}$ where $t' > t$. This is because the existence of causal relationship between Y_{it} and $\mathbf{Z}_{it'}$ implies a correlation between ϵ_{it} and $\mathbf{Z}_{it'}$, thereby violating Assumption 6. In Section 2.3, we pointed out the difficulty of assuming the lack of causal relationships between past outcomes and current treatment. In many applications, we expect feedback effects to occur over time between the outcome and treatment. For the same reason, assuming the absence of causal effects of past outcomes on current time-varying confounders may not be realistic.

The above discussion implies that researchers face the same key tradeoff regardless of whether or not time-varying confounders are adjusted. To adjust for unobserved time-invariant confounders, researchers must assume the absence of dynamic causal relationships among the outcome, treatment, and observed time-varying confounders. In contrast, the MSMs discussed in Section 2.4 can relax these assumptions under the assumption of no unobserved time-invariant confounder. Under the MSMs, past treatments can directly affect current outcome and past outcomes can either directly or indirectly (through time-varying confounders) affect current treatment.

3 A Matching Framework for Longitudinal Causal Inference

We consider estimating counterfactual outcomes for causal inference using LM-FE. We propose a non-parametric within-unit matching estimator that improves LM-FE by relaxing its linearity assumption. We also show that this matching framework can accommodate a variety of identification strategies and clarifies the comparisons between treated and control observations.

3.1 The Within-Unit Matching Estimator

Despite its popularity, LM-FE does not consistently estimate the average treatment effect (ATE) defined in equation (3) even when Assumptions 3 and 4 are satisfied. This is because LM-FE additionally requires the linearity assumption. Indeed, the linear unit fixed effects regression estimator given in equation (2) converges to a weighted average of unit-specific ATEs where the weights are proportional to the

within-unit variance of the treatment assignment variable (see Chernozhukov *et al.*, 2013, Theorem 1). This result is restated here as a proposition.

PROPOSITION 1 (INCONSISTENCY OF THE LINEAR FIXED EFFECTS REGRESSION ESTIMATOR (CHERNOZHUKOV *et al.*, 2013)) *Suppose* $\mathbb{E}(Y_{it}^2) < \infty$ *and* $\mathbb{E}(C_i S_i^2) > 0$ *where* $S_i^2 = \sum_{t=1}^T (X_{it} - \overline{X}_i)^2/(T-1)$. *Under Assumptions 3 and 4 as well as simple random sampling of units with* T *fixed, the linear fixed effects regression estimator given in equation* (2) *is inconsistent for the average treatment effect* τ *defined in equation* (3)*,*

$$\hat{\beta}_{\mathsf{FE}} \xrightarrow{p} \frac{\mathbb{E}\left\{C_i \left(\frac{\sum_{t=1}^T X_{it}Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1-X_{it})Y_{it}}{\sum_{t=1}^T 1-X_{it}}\right) S_i^2\right\}}{\mathbb{E}(C_i S_i^2)} \neq \tau$$

Thus, in general, under Assumptions 3 and 4, the linear fixed effects estimator fail to consistently estimate the ATE unless either the within-unit ATE or the within-unit proportion of treated observations is constant across units. This result also applies to the use of LM-FE in a cross-sectional context. For example, LM-FE is often used to analyze stratified randomized experiments (Duflo *et al.*, 2007). Even in this case, if the proportion of treated units and the ATE vary across strata, then the resulting least squares estimator will be inconsistent.

We consider a nonparametric matching estimator that eliminates this bias. The key insight from an earlier discussion is that under Assumptions 3 and 4 even though a set of time-invariant confounders $\mathbf{U}_i$ are not observed, we can nonparametrically adjust for them by comparing the treated and control observations measured at different time periods within the same unit. This within-unit comparison motivates the following matching estimator, which computes the difference of means between the treated and control observations within each unit and then averages it across units.

$$\hat{\tau}^{\mathsf{match}} = \frac{1}{\sum_{i=1}^N C_i} \sum_{i=1}^N C_i \left(\frac{\sum_{t=1}^T X_{it}Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1-X_{it})Y_{it}}{\sum_{t=1}^T (1-X_{it})}\right) \tag{10}$$

This matching estimator is attractive because unlike the estimator in equation (2), it does not require the linearity assumption. Under Assumptions 3 and 4, this within-unit matching estimator is consistent for the ATE defined in equation (3).

PROPOSITION 2 (CONSISTENCY OF THE WITHIN-UNIT MATCHING ESTIMATOR) *Under the same set of assumptions as Proposition 1, the within-unit matching estimator defined in equation* (10) *is consistent for the average treatment effect defined in equation* (3).

Proof is in Appendix A.3.

We make the connection to matching methods more explicit by defining a *matched set* $\mathcal{M}_{it}$ for each observation (i, t) as a group of other observations that are matched with it. For example, under the estimator proposed above, a treated (control) observation is matched with all control (treated) observations within the same unit, and hence the matched set is given by,

$$\mathcal{M}_{it}^{\mathsf{match}} \quad = \quad \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}\} \tag{11}$$

Thus far, we focused on the average treatment effect as a parameter of interest given that researchers often interpret the parameter β of LM-FE as the average contemporaneous treatment effect of X_{it} on Y_{it}. However, our matching framework can accomodate various identification strategies for different causal quantities of interest using different matched sets. That is, given any matched set $\mathcal{M}_{it}$, we can define the corresponding within-unit matching estimator $\hat{\tau}$ as,

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

where $Y_{it}(x)$ is observed when $X_{it} = x$ and is estimated using the average of outcomes among the units of its matched set when $X_{it} = 1 - x$,

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

Note that $\#\mathcal{M}_{it}$ represents the number of observations in the matched set and that D_{it} indicates whether the matched set $\mathcal{M}_{it}$ contains at least one observation, i.e., $D_{it} = \mathbf{1}\{\#\mathcal{M}_{it} > 0\}$. In the case of $\mathcal{M}_{it}^{\mathsf{match}}$, we have $D_{it} = C_i$ for any t.

Unlike the matching estimator given in equation (10), under LM-FE, the within-unit comparison is done through the de-meaning process where each observation is compared with the average of all other

observations within the same unit regardless of their treatment status (see equation (2)). This corresponds to a naive within-unit matching estimator based on the following matched set,

$$\mathcal{M}_{it}^{\text{LM-FE}} \;=\; \{(i', t') : i' = i, t' \neq t\} \tag{14}$$

In Appendix A.4, we discuss how this matching estimator suffers from the bias due to "mismatches" where the observations with the same treatment status are matched to each other.

3.2 Identification Strategies based on Within-Unit Comparison

The framework described above can accommodate diverse matching estimators through their corresponding matched sets $\mathcal{M}_{it}$. Here, we illustrate the generality of the proposed framework. First, we incorporate time-varying confounders $\mathbf{Z}_{it}$ by matching observations within each unit based on the values of $\mathbf{Z}_{it}$. For example, the within-unit nearest neighbor matching leads to the following matched set,

$$\mathcal{M}_{it}^{\text{NN}} \;=\; \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}, \mathcal{D}(\mathbf{Z}_{it}, \mathbf{Z}_{i't'}) = J_{it}\} \tag{15}$$

where $\mathcal{D}(\cdot, \cdot)$ is a distance measure (e.g., Mahalanobis distance), and

$$J_{it} \;=\; \min_{(i,t') \in \mathcal{M}_{it}^{\text{match}}} \mathcal{D}(\mathbf{Z}_{it}, \mathbf{Z}_{it'}) \tag{16}$$

represents the minimum distance between the time-varying confounders of this observation and another observation from the same unit whose treatment status is opposite. With this definition of matched set, we can construct the within-unit nearest neighbor matching estimator using equation (12). The argument of Proposition 2 suggests that this within-unit nearest neighbor matching estimator is consistent for the ATE so long as matching on $\mathbf{Z}_{it}$ eliminates the confounding bias.

Second, we consider the before-and-after (BA) design where each average potential outcome is assumed to have no time trend over a short time period. Since the BA design also requires the assumption of no carryover effect, the BA design may be most useful when for a given unit the change in the treatment status happens only once. Under the BA design, we simply compare the outcome right before and immediately after a change in the treatment status. Formally, the assumption can be written as,

ASSUMPTION 7 (BEFORE-AND-AFTER DESIGN) *For* $i = 1, 2, \ldots, N$ *and* $t = 2, \ldots, T$,

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

where $x \in \{0, 1\}$.

Under Assumptions 3 and 7, the average difference in outcome between before and after a change in the treatment status is a valid estimate of the local ATE, i.e., $\mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid X_{it} \neq X_{i,t-1})$.

To implement the BA design within our framework, we restrict the matched set and compare the observations within two subsequent time periods that have opposite treatment status. Formally, the resulting matched set can be written as,

$$\mathcal{M}_{it}^{\mathsf{BA}} \quad = \quad \{(i', t') : i' = i, t' \in \{t-1, t+1\}, X_{i't'} = 1 - X_{it}\} \tag{17}$$

It is straightforward to show that this matching estimator is equivalent to the following first difference (FD) estimator,

$$\hat{\beta}_{\mathsf{FD}} \quad = \quad \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=2}^{T} \{(Y_{it} - Y_{i,t-1}) - \beta(X_{it} - X_{i,t-1})\}^2 \tag{18}$$

In standard econometrics textbooks, the FD estimator defined in equation (18) is presented as an alternative estimation method for the LM-FE estimator. For example, Wooldridge (2010) writes "We emphasize that the model and the interpretation of β are exactly as in [the linear fixed effects model]. What differs is our method for estimating β" (p. 316; italics original). However, as can be seen from the above discussion, the difference lies in the identification assumption and the population for which the ATE is identified. Both the LM-FE and FD estimators match observations within the same unit. However, the FD estimator matches observations only from subsequent time periods whereas the LM-FE estimator matches observations from all time periods regardless of the treatment status.

Finally, a word of caution about the BA design is warranted. While Assumption 7 is written in terms of time trend of potential outcomes, the assumption is violated if past outcomes affect current treatment. Consider a scenario where the treatment variable X_{it} is set to 1 when the lagged outcome $Y_{i,t-1}$ takes a

value greater than its mean. In this case, even if the treatment effect is zero, the outcome difference between the two periods, $Y_{it} - Y_{i,t-1}$, is likely to be negative because of the so-called "regression toward the mean" phenomenon. James (1973) derives an expression for this bias under the normality assumption.

3.3 Estimation, Inference, and Specification Test

As a main analytical result of this paper, we show that *any* within-unit matching estimator can be written as a weighted linear regression estimator with unit fixed effects. The following theorem establishes this result and shows how to compute regression weights for a given matched set (see Gibbons *et al.*, 2011; Solon *et al.*, 2015, for related results).

THEOREM 1 (WITHIN-UNIT MATCHING ESTIMATOR AS A WEIGHTED FIXED EFFECTS ESTIMATOR) *Any within-unit matching estimator $\hat{\tau}$ defined by a matched set $\mathcal{M}_{it}$ equals the weighted linear fixed effects estimator, which can be efficiently computed as,*

$$\hat{\beta}_{\mathsf{WFE}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \{ (Y_{it} - \overline{Y}_i^*) - \beta (X_{it} - \overline{X}_i^*) \}^2 \tag{19}$$

where $\overline{X}_i^* = \sum_{t=1}^{T} W_{it} X_{it} / \sum_{t=1}^{T} W_{it}$, $\overline{Y}_i^* = \sum_{t=1}^{T} W_{it} Y_{it} / \sum_{t=1}^{T} W_{it}$, and the weights are given by,

$$W_{it} \;=\; D_{it} \sum_{i'=1}^{N} \sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{where} \quad w_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t) = (i',t') \\ 1/\#\mathcal{M}_{i't'} & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Proof is in Appendix A.5. In this theorem, $w_{it}^{i't'}$ represents the amount of contribution or "matching weight" of observation (i,t) for the estimation of treatment effect of observation (i',t'). For any observation (i,t), its regression weight is given by the sum of the matching weights across all observations.

Theorem 1 yields several practically useful implications. First, one can efficiently compute within-unit matching estimators even when the number of units is large. Specifically, a weighted linear fixed effects estimator can be computed by first subtracting its within-unit weighted average from each of the variables and then running another weighted regression using these "weighted demeaned" variables. Second, taking advantage of this equivalence, we can use various model-based robust standard errors for within-unit matching estimators (e.g., White, 1980a; Stock and Watson, 2008). Third, a within-unit

matching estimator is consistent for the ATE even when LM-FE is the true model (i.e., the linearity assumption holds). This observation leads to a simple specification test based on the difference between the unweighted and weighted least squares (White, 1980b) where the null hypothesis is that the linear unit fixed effects regression model is correct.

Finally, the theorem provides a simple way to adjust for time-varying confounders by directly including them in the weighted linear regression with unit fixed effects rather than using them to construct the matched set. This adjusted weighted linear fixed effects regression estimator is given by,

$$(\hat{\beta}_{\mathsf{WFEadj}}, \hat{\delta}_{\mathsf{WFEadj}}) \;=\; \underset{\beta,\delta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^*) - \beta(X_{it} - \overline{X}_i^*) - \delta^\top(\mathbf{Z}_{it} - \overline{\mathbf{Z}}_i^*)\}^2 \quad (21)$$

where $\overline{Y}_i^* = \sum_{t=1}^T W_{it}Y_{it}/\sum_{t=1}^T W_{it}$, $\overline{X}_i^* = \sum_{t=1}^T W_{it}X_{it}/\sum_{t=1}^T W_{it}$, $\overline{\mathbf{Z}}_i^* = \sum_{t=1}^T W_{it}\mathbf{Z}_{it}/\sum_{t=1}^T W_{it}$. Then, Theorem 1 implies that $\hat{\beta}_{\mathsf{WFEadj}}$ can be written as a model-adjusted within-unit matching estimator where the estimated potential outcome given in equation (13) can be written as,

$$\widehat{Y_{it}(x)} \;=\; \begin{cases} Y_{it} - \hat{\delta}_{\mathsf{WFEadj}}^\top \mathbf{Z}_{it} & \text{if } X_{it} = x \\ \frac{1}{\#\mathcal{M}_{it}} \sum_{(i',t')\in\mathcal{M}_{it}} Y_{i't'} - \hat{\delta}_{\mathsf{WFEadj}}^\top \mathbf{Z}_{i't'} & \text{if } X_{it} = 1 - x \end{cases} \quad (22)$$

Thus, this model-adjusted within-unit matching estimator matches the covariate-adjusted outcome with that of another observation with the opposite treatment status within the same unit.

4 Regression Models with Unit and Time Fixed Effects

Next, we extend the above analysis to the linear regression models with unit and time fixed effects. We begin by analyzing the simplest such model,

$$Y_{it} \;=\; \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it} \quad (23)$$

where α_i and γ_t are unit and time fixed effects, respectively. The inclusion of unit and time fixed effects accounts for both time-invariant and unit-invariant unobserved confounders in a flexible manner. Let $\mathbf{U}_i$ and $\mathbf{V}_t$ represent these time-invariant and unit-invariant 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.

The least squares estimate of β can be computed efficiently. To do this, we transform the outcome and treatment variables by subtracting their corresponding unit and time specific sample means from the original variables and then adding their overall sample means. Formally, the estimator is given by,

$$\hat{\beta}_{\mathsf{FE2}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\{(Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}) - \beta(X_{it} - \overline{X}_i - \overline{X}_t + \overline{X})\}^2 \tag{24}$$

where $\overline{Y}_t = \sum_{i=1}^{n} Y_{it}/N$, $\overline{X}_t = \sum_{i=1}^{n} X_{it}/N$, $\overline{Y} = \sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}/NT$, and $\overline{X} = \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}/NT$. In what follows, we first identify several problems of this two-way fixed effects estimator and propose ways to address these problems.

4.1 The Problems of the Two-way Fixed Effects Estimator

Like the one-way fixed effects regression model described in Section 2.1, the linear regression model with unit and time fixed effects defined in equation (23) assumes no carryover effect (Assumption 3). Typically, researchers also assume the following strict exogeneity assumption.

ASSUMPTION 8 (STRICT EXOGENEITY WITH UNOBSERVED TIME-INVARIANT AND UNIT-INVARIANT CONFOUNDERS) *For each* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

$$\epsilon_{it} \;\perp\!\!\!\perp\; \{\mathbf{X}_i, \mathbf{U}_i, \mathbf{V}_t\}$$

The key difference between the one-way and two-way fixed effects estimators is that the identification strategy of within-group comparison used for the one-way fixed effects estimator is not applicable to the two-way fixed effects estimator. The reason is simple. For any given observation, there is no other observation that belongs to the same unit and time. As a result, it is not possible to directly adjust for unobserved time-invariant and unit-invariant confounders at the same time. Thus, to identify the ATE, the two-way fixed effects estimator relies heavily on the linearity assumption. We investigate this problem in detail below. As in the case of one-way fixed effect model, the model assumes that past treatments do not affect current outcome and past outcomes do not affect current treatment.

We first extend the result of Proposition 1 to the two-way fixed effects estimator. We show that the two-way fixed effects estimator can be represented as the weighted average of three estimators: the

unit fixed effects regression estimator $\hat{\beta}_{\mathsf{FE}}$, the time fixed effects regression estimator $\hat{\beta}_{\mathsf{FEtime}}$, and *minus* the pooled regression estimator $\hat{\beta}_{\mathsf{pool}}$. Note that $\hat{\beta}_{\mathsf{FE}}$ and $\hat{\beta}_{\mathsf{FEtime}}$ themselves can be written as weighted averages of unit-specific and time-specific ATEs, respectively.

Although its validity requires the linearity assumption, the intuition is that $\hat{\beta}_{\mathsf{FE}}$ possibly suffers from the bias due to time effects whereas $\hat{\beta}_{\mathsf{FEtime}}$ is potentially biased because of unit effects. Since $\hat{\beta}_{\mathsf{pool}}$ has both types of biases, one may hope that subtracting it from the sum of the other two estimators cancel out both biases. For completeness, we define the time fixed effects and pooled regression estimators as,

$$\hat{\beta}_{\mathsf{FEtime}} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \{(Y_{it} - \overline{Y}_t) - \beta(X_{it} - \overline{X}_t)\}^2 \tag{25}$$

$$\hat{\beta}_{\mathsf{pool}} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \beta X_{it})^2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} Y_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}} - \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (1 - X_{it}) Y_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} (1 - X_{it})} \tag{26}$$

In this case of no covariate, the pooled regression estimator is equivalent to the difference of means between the treatment and control groups based on the entire sample. Now, we formally state this result.

PROPOSITION 3 (TWO-WAY FIXED EFFECTS ESTIMATOR AS A WEIGHTED AVERAGE) *The two-way fixed effects estimator defined in equation* (24) *can be written as a weighted average of the unit fixed effects regression estimator defined in equation* (2)*, the time fixed effects estimator defined in equation* (25)*, and the pooled regression estimator defined in equation* (26)*,*

$$\hat{\beta}_{\mathsf{FE2}} = \frac{\omega_{\mathsf{FE}} \times \hat{\beta}_{\mathsf{FE}} + \omega_{\mathsf{FEtime}} \times \hat{\beta}_{\mathsf{FEtime}} - \omega_{\mathsf{pool}} \times \hat{\beta}_{\mathsf{pool}}}{w_{\mathsf{FE}} + w_{\mathsf{FEtime}} - w_{\mathsf{pool}}}$$

When N and T are sufficiently large, the weights are approximated by,

$$\begin{aligned} \omega_{\mathsf{FE}} &\approx \mathbb{E}(S_i^2) \\ \omega_{\mathsf{FEtime}} &\approx \mathbb{E}(S_t^2) \\ \omega_{\mathsf{pool}} &\approx S^2 \end{aligned}$$

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

Proof is given in Appendix A.6. The weights for $\hat{\beta}_{\mathsf{FE}}$, $\hat{\beta}_{\mathsf{FEtime}}$, and $\hat{\beta}_{\mathsf{pool}}$, are approximately equal to the average variance of treatment within unit, within time, and within the entire data set, respectively.

Proposition 3 illustrates the difficulty of the two-way fixed effects regression estimator. Direct adjustment of unit-invariant and time-invariant unobserved confounders is impossible because no more than

one observation shares the same unit and time. To overcome this difficulty, the two-way fixed effects regression estimator combines three estimators: (1) the unit fixed effects estimator, which addresses unit-specific time-invariant confounders but not time-specific unit-invariant confounders, (2) the time fixed effects estimator, which adjusts time-specific unit-invariant confounders but not unit-specific time-invariant, and (3) the pooled estimator, which do not address any of these unobserved confounders. In general, however, there is no guarantee that this combined estimator eliminates bias.

4.2 Two-way Matching Estimators

Next, we partially address the difficulties of the two-way fixed effects regression by considering matching estimators with unit and time fixed effects. Proposition 5 in Appendix A.4 shows that the one-way fixed effects regression estimator can be written as a naive within-unit matching estimator with the adjustment for mismatches. Here, we extend this result to the two-way fixed effects regression estimator and show that it can also be written as a naive two-way matching estimator with the adjustment for mismatches.

PROPOSITION 4 (THE TWO-WAY FIXED EFFECTS ESTIMATOR AS A NAIVE TWO-WAY MATCHING ESTIMATOR WITH ADJUSTMENT) *The two-way fixed effects estimator defined in equation* (24) *is equivalent to the following adjusted matching estimator,*

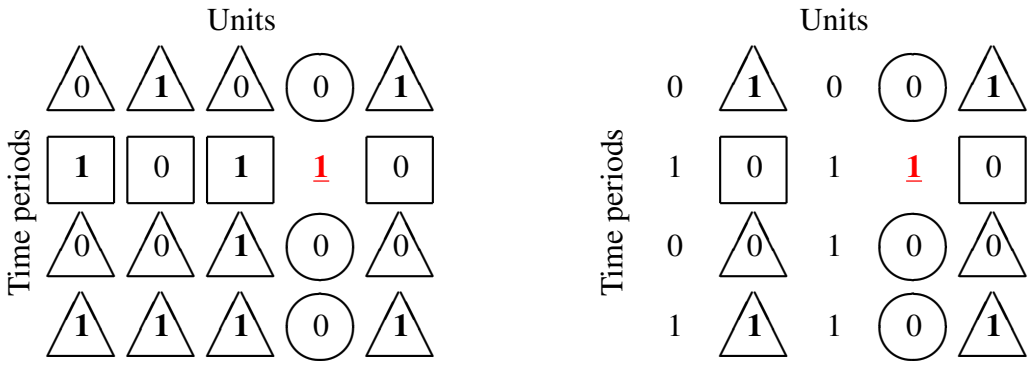
$$\hat{\beta}_{\mathsf{FE2}} \;=\; \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right)\right\}$$

where for $x=0,1$,

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

Proof is given in Appendix A.7.

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 all the



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

Figure 5: 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 4) and the adjusted matching estimator (Proposition 6), 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.

other observations from the same time period are added together. We call them the *within-unit matched set* $\mathcal{M}_{it}$ and the *within-time matched set* $\mathcal{N}_{it}$, respectively. In the case of the linear two-way fixed effects estimator, these matched sets are defined as,

$$\mathcal{M}_{it}^{\mathsf{FE2}} = \{(i', t') : i' = i, t' \neq t\} \tag{27}$$

$$\mathcal{N}_{it}^{\mathsf{FE2}} = \{(i', t') : i' \neq i, t' = t\} \tag{28}$$

Finally, to adjust for unit and time fixed effects, we use observations that share the same unit or time as those in $\mathcal{N}_{it}$ and $\mathcal{M}_{it}$, respectively, and subtract their mean 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,

$$\mathcal{A}_{it} = \{(i', t') : i' \neq i, t' \neq t, (i, t') \in \mathcal{M}_{it}, (i', t) \in \mathcal{N}_{it}\} \tag{29}$$

Therefore, 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 5 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, as in the one-way case (see Proposition 5 in Appendix A.4), 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 4 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.

Given this result, we improve the two-way fixed effects estimator in the same manner as done in Section 3.1 by only matching each observation with other observations of the opposite treatment status to estimate the counterfactual outcome. That is, we use the *within-unit matched set* $\mathcal{M}_{it}^{\mathsf{match}}$ defined in equation (11), which consists of the observations within the same unit but with the opposite treatment status. 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,

$$\mathcal{N}_{it}^{\mathsf{match}} \quad = \quad \{(i', t') : t' = t, X_{i't'} = 1 - X_{it}\}. \tag{30}$$

Then, using equation (29), we can define the corresponding adjustment set $\mathcal{A}_{it}^{\mathsf{match}}$.

Unlike the one-way case (see Section 3.1), 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 4.3). This point is illustrated by Panel (b) of Figure 5 where the adjustment set $\mathcal{A}_{it}^{\mathsf{match}}$ (triangles) includes the observations

of the same treatment status. A formal statement of this result is given as Proposition 6 in Appendix A.8, which also establishes the equivalence between this adjusted matching estimator and the weighted two-way fixed effects estimator.

4.3 The General Difference-in-Differences Estimator

The two-way fixed effects estimator is often motivated by the difference-in-differences (DiD) design (e.g., Angrist and Pischke, 2009). 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). Under the DiD design, we assume that the average potential outcomes under the control condition have parallel time trends for the treatment and control groups (Abadie, 2005).

ASSUMPTION 9 (DIFFERENCE-IN-DIFFERENCES DESIGN) *For* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

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

where the estimand is the local ATE for the treated, i.e., $\mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid X_{it} = 1, X_{i,t-1} = 0)$.

It is well known that the standard DiD estimator is numerically equivalent to the linear two-way fixed effects regression estimator if there are two time periods and the treatment is administered to some units only in the second time period. Unfortunately, this equivalence result does not generalize to the multi-period DiD design, in which the number of time periods may exceed two and each unit may receive the treatment multiple times. Here, we prove that the general multi-period DiD estimator is a special case of the two-way matching estimators introduced in Section 4.2. We establish that by further restricting the matched sets, $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$, we can construct a two-way matching estimator, which is equivalent to the multi-period DiD estimator. Furthermore, we show that this estimator can be represented as a weighted linear two-way fixed effects regression estimator.

To formulate the DiD estimator as a two-way matching estimator, we first define the within-unit matched set of this treated observation to be the observation of the previous period if it belongs to the

control group, and to be an empty set otherwise,

$$\mathcal{M}_{it}^{\mathsf{DiD}} \;=\; \{(i',t') : i' = i, t' = t-1, X_{i't'} = 0\} \tag{31}$$

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.

$$\mathcal{N}_{it}^{\mathsf{DiD}} \;=\; \{(i',t') : i' \neq i, t' = t, X_{i't'} = X_{i',t'-1} = 0\} \tag{32}$$

In addition, one may restrict the within-time matched sets further by only considering observations that have similar observed values of past outcomes, time-invariant covariates or more generally any pre-treatment variables. So long as Assumption 9 holds conditional on these variables, the DiD design should be valid.

Once $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$ are defined in this way, the adjustment set $\mathcal{A}_{it}$ is given by equation (29). This set contains the control observations in the previous period that share the same unit as those in $\mathcal{N}_{it}^{\mathsf{DiD}}$.

$$\mathcal{A}_{it}^{\mathsf{DiD}} \;=\; \{(i',t') : i' \neq i, t' = t-1, X_{i't'} = X_{i't} = 0\} \tag{33}$$

In this case, the number of observations in this adjustment set is the same as that in $\mathcal{N}_{it}^{\mathsf{DiD}}$.

Using these matched and adjustment sets, we can define the multi-period DiD estimator as the following two-way matching estimator,

$$\hat{\tau}_{\mathsf{DiD}} \;=\; \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right) \tag{34}$$

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

$$\widehat{Y_{it}(x)} \;=\; \begin{cases} Y_{it} & \text{if } X_{it} = 1 \\ Y_{i,t-1} + \frac{1}{\#\mathcal{N}_{it}^{\mathsf{DiD}}} \sum_{(i',t)\in\mathcal{N}_{it}^{\mathsf{DiD}}} Y_{i't} - \frac{1}{\#\mathcal{A}_{it}^{\mathsf{DiD}}} \sum_{(i',t')\in\mathcal{A}_{it}^{\mathsf{DiD}}} Y_{i't'} & \text{if } X_{it} = 0 \end{cases} \tag{35}$$

When the treatment status of a unit changes from the control 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., $D_{it} = 1$, this estimator estimates the counterfactual outcome for observation (i,t) by

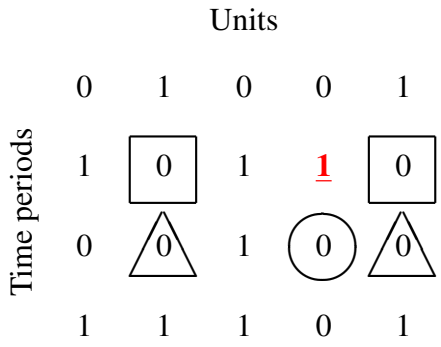


Figure 6: Illustration of how observations are used to estimate counterfactual outcomes for the DiD estimator (Proposition 2). The red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circle indicates the matched observation within the same unit, $Y_{i,t-1}$, whereas squares indicate those from the same time period, $\mathcal{N}_{it}^{\mathsf{DiD}}$. Finally, triangles 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 5, $\mathcal{A}_{it}^{\mathsf{DiD}}$ only contains control observations and hence no adjustment is required.

subtracting from its observed outcome of the previous period $Y_{i,t-1}$, the average difference in outcomes between the same two time periods among the other units whose treatment status remains unchanged as the control condition.

This matching estimator is illustrated in Figure 6. In this example, the counterfactual outcome for the treated unit (represented by the red underlined 1) is estimated as the difference between the outcome under the control condition from the previous period (circle) and the average difference (square minus triangle) between the current and previous periods for the two units which receive the control condition in both time periods. Thus, the DiD estimator, by construction, avoids mismatches and therefore eliminates the need for additional adjustment. Finally, the following proposition shows that the DiD estimator can be represented as a weighted linear two-way fixed effects regression estimator.

THEOREM 2 (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* $\hat{\tau}_{\mathsf{DiD}}$*, defined in equation* (34)*, is equivalent to the following weighted two-way fixed effects regression estimator,*

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

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 A.9.

Theorem 2 shows that the DiD estimator of the average treatment effect for the treated can be obtained by calculating the weighted linear two-way fixed effects regression estimator.[12] Under this framework, it is also possible to incorporate pre-treatment confounders as additional regressors in the weighted two-way fixed effects regression. This method of covariate adjustment can be justified as a matching estimator based on covariate-adjusted outcomes (see Section 2.5).

We emphasize that while Assumption 9 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 (see Ashenfelter and Card, 1985, who explore this issue in the case of job training program).[13] As discussed for the before-and-after design in Section 3.2, suppose that the treatment effect does not exist but the treatment assignment variable X_{it} is determined 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).

Suppose that a researcher finds a statistically significant difference in the baseline outcome between the treatment and control groups. The key question is whether this difference is a cause of the treatment

[12] Since some of the regression weights can be negative, we view the weighted two-way fixed effects estimators given in Proposition 6 and Theorem 2 as the method of moments estimators where the moment conditions are given by the first order condition. Then, the standard asymptotic properties of the method of moments estimator apply directly to this case (Hansen, 1982). In addition, while the regression weights can be negative, the fixed effects projection method can be applied on the complex plane in order to reduce the dimensionality as before.

[13] The assumption of no carryover effect (Assumption 3) is slightly relaxed because the lagged treatment is adjusted (i.e., $X_{i,t-1} = 0$ for both treated and control observations).

assignment. Assume that there is no causal relationship between the baseline outcome and the treatment. Then, the DiD estimator is appropriate even if unobserved time-varying confounders are causes of this baseline outcome difference. On the other hand, if past outcomes affect current treatment, then the baseline outcome difference reflects the causal effect of the baseline outcome on the treatment assignment. In this case, the validity of the DiD design is compromised. Adjusting for the baseline outcome difference is sufficient in the absence of unobserved time-invariant confounders.

Unfortunately, the simultaneous presence of unobserved time-invariant confounders and causal relationships between past outcomes and current treatment makes the identification of average treatment effects impossible. This can be seen from the fact that if the adjustment completely eliminates the baseline outcome difference, the DiD estimator is equivalent to the difference-in-means estimator based on the post-treatment outcome variable, which does not adjust for unobserved time-invariant confounders. Thus, in a given application, researchers must decide whether to focus on unobserved time-invariant confounders or causal effects of past outcomes on current treatment.

4.4 The Synthetic Control Method

The above discussion applies to the synthetic control method (Abadie *et al.*, 2010). Consider the setting where there is one treated unit, i^*, which received the treatment at time T. The rest of $N-1$ units belong to the control group and do not receive the treatment. The idea of the synthetic control method is to match the treated unit with a weighted average of the control units using past outcomes $Y_{i1}, \ldots, Y_{i,T-1}$ and possibly past time-varying confounders $\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{i,T-1}$. Doing so leads to an estimate of the potential outcome at time T under the control condition for the treated unit, i.e., $Y_{i^*T}(0)$, based on a weighted average of the observed outcome of the control units at time T.

Abadie *et al.* (2010) show that the synthetic control method can be applied when the data generation process is based on the following autoregressive model,

$$Y_{iT}(0) = \rho_T Y_{i,T-1}(0) + \delta_T^\top \mathbf{Z}_{iT} + \epsilon_{iT} \tag{36}$$

$$\mathbf{Z}_{iT} = \lambda_{T-1} Y_{i,T-1}(0) + \mathbf{\Delta}_T \mathbf{Z}_{i,T-1} + \nu_{iT} \tag{37}$$

where $\boldsymbol{\Delta}_T$ is a matrix of coefficients. The errors of a unit are assumed to be independent of all previous outcomes and time-varying confounders including those of the other units. This assumption is formalized as $\mathbb{E}(\epsilon_{iT} \mid \{Y_{i't'}, \mathbf{Z}_{i't'}\}_{1 \leq i' \leq N, 1 \leq t' \leq T-1}) = \mathbb{E}(\nu_{iT} \mid \{Y_{i't'}, \mathbf{Z}_{i't'}\}_{1 \leq i' \leq N, 1 \leq t' \leq T-1}) = 0$. This model adjusts for the previous outcome $Y_{i,T-1}(0)$ while assuming the absence of unobserved time-invariant confounders $\mathbf{U}_i$. Under this model, the previous outcome does affect current time-varying confounders $\mathbf{Z}_{iT}$, which is not allowed under the fixed effects regression models (see Section 2.5).

While the synthetic control method directly adjusts for past outcomes, Abadie *et al.* (2010) propose the following generalization of linear fixed effects regression model as a main model to motivate the method. This model contains unobserved time-invariant confounders $\mathbf{U}_i$,

$$Y_{it}(0) \quad = \quad \gamma_t + \delta_t^\top \mathbf{Z}_{it} + \xi_t^\top \mathbf{U}_i + \epsilon_{it} \tag{38}$$

where $\mathbb{E}(\epsilon_{it} \mid \{\mathbf{Z}_{i't'}, \mathbf{U}_{i'}\}_{1 \leq i' \leq N, 1 \leq t' \leq T}) = 0$, implying that the error term of an observation is mean independent of all time-varying confounders and unobserved time-invariant confounders including those of the other observations. This assumption suggests that, unlike the above autoregressive model (but like the fixed effects regression models), this model assumes that past outcomes do not affect current time-varying confounders.

Moreover, in addition to the linearity assumption, Abadie *et al.* (2010) assume that there exist a set of weights $\{w_i\}_{i \neq i^*}$ such that $\sum_{i \neq i^*} w_i \mathbf{Z}_{it} = \mathbf{Z}_{i^*t}$ for all $t \leq T-1$ and $\sum_{i \neq i^*} w_i \mathbf{U}_i = \mathbf{U}_{i^*}$ hold. This assumption implies that one set of weights can balance time-varying confounders and unobserved time-invariant confounders at the same time. The authors show that if this assumption holds, then the synthetic control method yields an unbiased estimate of $Y_{i^*T}(0)$ under the model given in equation (38). In general, however, balancing observed confounders does not necessarily balance unobserved confounders. Therefore, we view the synthetic control method as a selection-on-observables approach. The method adjusts for past outcomes and past time-varying confounders, but it may not be able to adjust for unobserved time-invariant confounders.

In our framework, we can follow the idea of the synthetic control method and adjust for past outcomes and past time-varying confounders when employing the multi-period DiD design. Specifically, we can

restrict the within-time matched set such that only units similar to a treated unit are included in the comparison set. For example, we can use a caliper c, which represents the maximum deviation allowed for all treated units. Then, the within-time matched set becomes,

$$\mathcal{N}_{it}^{\text{DiD-caliper}} \;=\; \{(i', t') : i' \neq i, t' = t, X_{i't'} = X_{i',t'-1} = 0, |Y_{i,t-1} - Y_{i',t'-1}| \leq c\}. \tag{39}$$

Yet another possibility is to use the Mahalanobis distance measure so that we can match on past outcomes and/or past time-varying confounders from more than one time period. These and other matching methods can be combined with the multi-period DiD design as weighted linear two-way fixed effects regression models in our framework.

5 An Empirical Illustration

In this section, we illustrate our proposed methodology by estimating the effects of General Agreement on Tariffs and Trade (GATT) membership on bilateral trade with various fixed effects models. We show that different causal assumptions can yield substantively different results.

5.1 Effects of GATT Membership on Bilateral Trade

Does GATT membership increase international trade? Rose (2004) finds the answer to this question is negative. Based on the standard gravity model with year fixed effects applied to dyadic trade data, he finds economically and statistically insignificant effect of GATT membership (and its successor World Trade Organization or WTO) on bilateral trade. This finding led to subsequent debates among empirical researchers as to whether or not GATT actually promotes trade (e.g., Gowa and Kim, 2005; Tomz *et al.*, 2007; Rose, 2007). In particular, Tomz *et al.* (2007) find a substantial effect of GATT/WTO on trade when a broader definition of membership is employed. They argue that *nonmember participants* such as former colonies, *de facto* members, and provisional members should also be included in empirical analysis since they enjoy similar "rights and obligations."

5.2 Models and Assumptions

We analyze the data set from Tomz *et al.* (2007) which updates and corrects some minor errors in the data used by Rose (2004). Unlike Rose (2004), however, our analysis is restricted to the period between 1948 and 1994 so that we focus on the effects of GATT and avoid conflating them with the effects of the WTO. As shown below, this restriction does not significantly change the original conclusion of Rose (2004), but it leads to a conceptually cleaner analysis. This yields a dyadic data set of bilateral international trade where a total number of dyads is 10,289 and a total number of (dyad-year) observations is 196,207.

While, in this literature, scholars almost universally relied upon linear fixed effects regression models, our earlier discussion underscores the importance of causal assumptions when applying fixed effects regression models. In particular, all of fixed effects regression models used in the literature assume that past GATT/WTO membership status does not affect current trade volume and past trade volume does not affect current membership status. Moreover, it is also assumed that one dyad's trade volume is not directly affected by other dyad's membership status. In the world of economic interdependence, such an assumption may be problematic. However, as we explained in this paper, relaxing these assumptions is difficult within the fixed effects regression framework.

For the empirical analysis of this paper, we maintain these assumptions and focus on improving the linear fixed effects regression estimators used in the literature. We use two different definitions of GATT membership: formal membership as used by Rose and participants as adopted by Tomz *et al.*. For each membership definition, we estimate its average effects on bilateral trade. We consider the treatment variable X_{it} indicating whether both countries in a dyad i are members of GATT or not in a given year t (mix of dyads with one member and no member). This analysis focuses on the reciprocity hypothesis that the GATT can impact bilateral trade only when countries mutually agree on reducing trade barriers. We begin with the following model with dyadic fixed effects,

$$\log Y_{it} \quad = \quad \alpha_i + \beta X_{it} + \delta^\top \mathbf{Z}_{it} + \epsilon_{it} \tag{40}$$

where Y_{it} is the bilateral trade volume for dyad i in year t, and $\mathbf{Z}_{it}$ represents a vector of time-varying

confounders including 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. As shown in Proposition 1, even if no carryover effect and strict exogeneity assumptions hold, the linear fixed effects regression estimator may be subject to potential biases when there exists heterogeneous treatment effect and/or treatment assignment probability across dyads. Indeed, Subramanian and Wei (2007) find substantial heterogeneity in the effects of GATT/WTO on trade.

Next, we consider the linear time fixed effects model, which relies on the comparison across dyads within each year. This model is given by,

$$\log Y_{it} \quad = \quad \gamma_t + \beta X_{it} + \delta^\top \widetilde{\mathbf{Z}}_{it} + \epsilon_{it} \tag{41}$$

where in addition to the aforementioned variables for $\mathbf{Z}_{it}$, $\widetilde{\mathbf{Z}}_{it}$ also includes variables that do not vary within a dyad such as log distance between the two countries and log product of land area (see Rose (2004) for the full list of covariates and their description). Similar to the linear model with dyadic fixed effects, this model might suffer from the bias due to the linearity assumption, unless treatment effect is constant and/or treatment assignment probability is identical over time. However, both Rose (2004) and Tomz *et al.* (2007) find substantial heterogeneity in the effects of GATT/WTO on trade in different time periods. Furthermore, the variability in institutional membership suggests that the model may not provide a consistent estimate for the average treatment effect. As Tomz *et al.* (2007) note, most countries have become a member such that "participation was nearly ubiquitous" (p. 2014) in recent periods.

Finally, some scholars have used the linear regression models with both dyadic and year fixed effects (e.g., Tomz *et al.*, 2007). Such model is given by,

$$\log Y_{it} \quad = \quad \alpha_i + \gamma_t + \beta X_{it} + \delta^\top \mathbf{Z}_{it} + \epsilon_{it}. \tag{42}$$

Our analysis in Section 4 shows that this estimator heavily relies upon the linearity assumption and uses all observations to estimate each counterfactual outcome regardless of their treatment status. The problem stems from the difficulty to nonparametrically adjust for unobserved time-invariant and dyad-invariant confounders at the same time.

We examine the robustness of findings when counterfactual outcomes are estimated based on different causal assumptions. We first compare the linear one-way dyad (time) fixed effects regression estimators with the within-unit (within-time) nonparametric matching estimators by relaxing the linearity assumption. Second, we further restrict the matched set of the within-unit matching estimator by focusing on the observations in two subsequent time periods with treatment status change (i.e., the before-and-after design). The assumption that the average outcome stays constant over the two periods might not be credible given that there is an increasing time trend in bilateral trade volumes. Thus, we also employ the DiD design. We estimate the local ATE for the treated assuming parallel time trend between the treated (dyads with GATT membership) and control (dyads with no membership) observations. Finally, as described at the end of Section 4.4, we adjust for the past outcome by focusing exclusively on the control units whose previous outcome values are within 0.35 times the standard deviation of the outcome variable away from that of a treated unit.

5.3 Empirical Results

Figure 7 summarizes the results. We implement various matching methods with weighted linear fixed effects regression to adjust for both observed time-varying confounders and unobserved heterogeneity within unit (or time or both). We incorporate model-based covariate adjustment as described in equation (21). Throughout our analysis, we report the robust standard errors that allow for the presence of serial correlation and heteroskedasticity (Arellano, 1987; Hansen, 2007). Table 1 in Appendix A.10 contains results from specification tests as discussed in Section 3.3 and estimates from alternative definitions of dyadic membership.

We begin by estimating the dyad fixed effects estimator as done in Rose (2007). We compare this result with those based on the weighted dyad fixed effects estimator given in equation (21) and first differences where the within-unit matched set is constrained by equation (17). We find that relaxing the linearity assumption by eliminating mismatches within each dyad leads to some substantive differences. For example, large and positive effects of participants in the standard dyad fixed effects models disappear

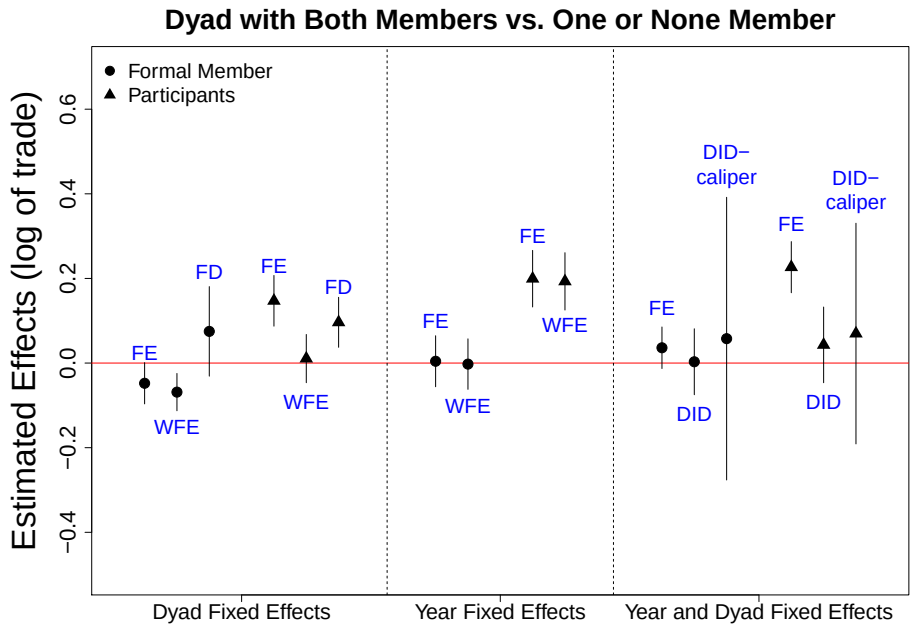


Figure 7: **Estimated Effects of GATT on the Logarithm of Bilateral Trade based on Various Fixed Effects Estimators**: The plot presents point estimates and 95% confidence intervals for the estimated effects of "Both vs. Mix" on bilateral trade, i.e., the comparison between dyads of two GATT members and those consisting of either one GATT member. Formal Member includes only formal GATT members as done in Rose (2004), whereas Participants includes nonmember participants as defined in Tomz *et al.* (2007). "WFE" are the estimates based on the regression weights given in equation (20), which yield the estimated average treatment effects based on equation (21). Results based on other weighted fixed effects estimators, i.e., first differences and difference-in-differences, are also presented. The results from an alternative way to adjust for past outcomes when employing the multi-period DiD design are presented in "DiD-caliper" where matched units' pre-treatment outcomes are within 0.35 times the standard deviation of the outcome variable away from that of a treated unit. Robust standard errors, allowing for the presence of serial correlation as well as heteroskedasticity are used.

when relaxing the linearity assumption, whereas slightly negative effects of formal membership appear to be stronger. Furthermore, under the before-and-after design, the point estimates are all positive for both formal members and participants.

Next, we use the year fixed effects estimator as done in Rose (2004) and compare its results with those from the weighted year fixed effects estimator given in equation (21) (except that in the current case i represents year and t represents dyad). Consistent with Rose (2004)'s original finding, we find little effect of GATT formal membership on trade using the standard year fixed effects estimator. The results from the weighted year fixed effects estimator generally agree with his finding. Likewise, the positive effect of GATT when membership variable includes nonmember participants remain unchanged with the

weighted year fixed effects estimator. When compared with the dyad fixed effects estimators, the year fixed effects estimators exhibit a less degree of sensitivity to underlying causal assumptions (i.e., different weighting schemes and different ways of adjusting for covariates). This happens because the membership variable does not vary much within each dyad and heterogeneity across dyads is greater than across years. Different causal assumptions yield different sets of observations for estimating counterfactual outcomes.

Finally, we estimate the two-way fixed effects estimators with both dyad and year fixed effects as done in Tomz *et al.* (2007). We compare these results with those based on the weighted two-way fixed effects estimator that corresponds to the DiD estimator. For both formal members and participants the standard two-way fixed effects models may overestimate the effect when compared to the weighted regression. In fact, the DiD estimator consistently give smaller estimates than the standard two-way fixed effects model. Using a caliper by choosing the maximum deviation allowed in the pre-treatment outcome for matched units yields similar differences (equation (39)). Note that standard errors for these estimators tend to be larger. This is because high collinearity between covariates arises in this data set when a small number of observations with non-zero weights are actually used. For example, we observe very little variation in most year-varying covariates such as GSP and Regional FTA across dyads.

In sum, our analysis suggests that the empirical results based on linear fixed effects regression models can critically depend on their underlying causal assumptions. We have shown that the effects of GATT membership can vary substantially depending on how counterfactual outcomes are estimated.

6 Concluding Remarks

The title of this paper asks the question of when researchers should use linear fixed effects regression models for causal inference with longitudinal data. According to our analysis, the answer to this question depends on the tradeoff between unobserved time-invariant confounders and dynamic causal relationships between outcome and treatment variables. In particular, if the treatment assignment mechanism critically depends on past outcomes, then researchers are likely to be better off investing their efforts in measuring and adjusting for confounders rather than adjusting for unobserved time-invariant con-

founders through fixed effects models under unrealistic assumptions. In such situations, methods based on the selection-on-observables such as matching and weighting are more appropriate. This conclusion also applies to the before-and-after and difference-in-differences designs that are closely related to linear fixed effects regression models.

If, on the other hand, researchers are concerned about time-invariant confounders and are willing to assume the absence of dynamic causal relationships, then fixed effects regression models are effective tools to adjust for unobserved time-invariant confounders. In this paper, we propose a new matching framework that improves linear fixed effects regression models by relaxing the linearity assumption. Under this framework, we show how to incorporate various identification strategies and implement them as weighted linear fixed effects regression estimators. For example, we extend the difference-in-differences estimator to the general case of multiple periods and repeated treatments and show its equivalence to a weighted linear regression with unit and time fixed effects. Our framework also facilitates the incorporation of additional covariates, model-based inference, and specification tests.

Unfortunately, researchers must choose either to adjust for unobserved time-invariant confounders under the fixed effects modeling framework or model dynamic causal relationships between treatment and outcome under a selection-on-observables approach. No existing method can achieve both objectives without additional assumptions. Finally, while we limit our discussion to the case of a binary treatment, our nonparametric identification analysis based on DAGs is applicable to the case where the treatment is non-binary. Thus, researchers who are analyzing a non-binary treatment must face the same tradeoff described in 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.

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.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 2, 277–297.

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

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics* **67**, 4, 648–660.

Beck, N. (2001). Time-series-cross-section data: What have we learned in the past few years. *Annual Review Political Science* **4**, 271–293.

Bell, A. and Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods* **3**, 1, 133–153.

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.

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science* **57**, 2, 504–520.

Brito, C. and Pearl, J. (2002). Generalized instrumental variables. In *Proceedings of the 18th Conference of Uncertainty in Artificial Intelligence*, 85–93.

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

Clark, T. S. and Linzer, D. A. (2015). Should i use fixed or random effects? *Political Science Research and Methods* **3**, 2, 399–408.

Duflo, E., Glennerster, R., and Kremer, M. (2007). *Handbook of Development Economics*, vol. 4, chap. Using Randomization in Development Economics Research: A Toolkit, 3895–3962. Elsevier.

Gibbons, C. E., Suárez Serrato, J. C., and Urbancic, M. B. (2011). Broken or fixed effects? Tech. rep., Department of Economics, University of California, Berkeley.

Gowa, J. and Kim, S. Y. (2005). An exclusive country club: The effects of the GATT on trade, 1950-94. *World Politics* **57**, 4, 453–478.

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.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**, 3, 199–236.

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.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics* **29**, 121–130.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association* **96**, 455, 870–882.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edn.

Plümper, T. and Troeger, V. E. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis* **15**, 2, 124–139.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 5, 550–560.

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.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth COnference on Uncertainty in Artificial Intelligence*.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* **101**, 476, 1398–1407.

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

Stock, J. H. and Watson, M. W. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* **76**, 1, 155–174.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25**, 1, 1–21.

Subramanian, A. and Wei, S.-J. (2007). The WTO promotes trade, strongly but unevenly. *Journal of International Economics* **72**, 1, 151–175.

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. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 4, 817–838.

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

Wilson, S. E. and Butler, D. M. (2007). A lot more to do: The sensitivity of time-series cross-section analyses to simple alternative specifications. *Political Analysis* **15**, 2, 101–123.

Wooldridge, J. M. (2005a). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics* **87**, 2, 385–390.

Wooldridge, J. M. (2005b). *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg (eds. D. Andrews and J. Stock)*, chap. Unobserved Heterogeneity and Estimation of Average Partial Effects. Cambridge University Press, Cambridge.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA, 2nd edn.

A Supplementary Appendix

A.1 Equivalence between Assumptions 2 and 4 under LM-FE

First, under LM-FE, $Y_{it}(x) = \alpha_i + \beta x + \epsilon_{it}$. This implies that Assumption 2 is equivalent to the conditional independence between $\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T$ and $\mathbf{X}_i$ given $\mathbf{U}_i$. Thus, Assumption 2 implies Assumption 4. Next, we show that Assumption 4 implies Assumption 2. It is sufficient to prove the equivalence when $T = 3$ as the same argument can be repeatedly applied to the case with $T > 3$.

$$
\begin{aligned}
& p(\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3, X_{i1}, X_{i2}, X_{i3} \mid \mathbf{U}_i) \\
= \; & p(\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3 \mid X_{i1}, X_{i2}, X_{i3}, \mathbf{U}_i) p(X_{i1}, X_{i2}, X_{i3} \mid \mathbf{U}_i) \\
= \; & p(\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3 \mid X_{i1}, X_{i2}, \mathbf{U}_i) p(X_{i1}, X_{i2}, X_{i3} \mid \mathbf{U}_i) \\
= \; & p(\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3 \mid X_{i1}, \mathbf{U}_i) p(X_{i1}, X_{i2}, X_{i3} \mid \mathbf{U}_i) \\
= \; & p(\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3 \mid \mathbf{U}_i) p(X_{i1}, X_{i2}, X_{i3} \mid \mathbf{U}_i)
\end{aligned}
$$

which shows that $\{Y_{it}(1), Y_{it}(0)\}_{t=1}^3$ are conditionally independent of $\mathbf{X}_i$ given $\mathbf{U}_i$.

A.2 Sequential Ignorability with Unobserved Time-invariant and Observed Time-varying Confounders

It is helpful to adopt the potential outcomes framework and think about a randomized experiment, in which Assumption 6 holds. Consider the sequential randomization of treatment at each time period conditional on all past treatments and the values of all previous and current time-varying covariates. As before, we assume that the treatment assignment mechanism does not depend on the realized outcome from the previous time periods but allow treatment assignment probabilities to be a function of unobserved time-invariant confounders. This treatment assignment mechanism is formalized as the following sequential ignorability assumption with unobserved time-invariant and observed time-varying confounders.

ASSUMPTION 10 (SEQUENTIAL IGNORABILITY WITH UNOBSERVED TIME-INVARIANT AND OBSERVED TIME-VARYING CONFOUNDERS) *For each* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

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

This assumption implies the following conditional independence, $\{Y_{it}(1), Y_{it}(0)\}_{t=1}^T \perp\!\!\!\perp \mathbf{X}_i \mid \mathbf{Z}_i, \mathbf{U}_i$.

A.3 Proof of Proposition 2

We begin by rewriting the within-unit matching estimator as,

$$
\hat{\tau}_{\mathsf{match}} = \frac{1}{\frac{1}{N}\sum_{i=1}^N C_i} \cdot \frac{1}{N} \sum_{i=1}^N C_i \left(\frac{\sum_{t=1}^T X_{it} Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1 - X_{it}) Y_{it}}{\sum_{t=1}^T (1 - X_{it})} \right) \tag{43}
$$

By law of large numbers, the first term converges in probability to $1/\Pr(C_i = 1)$. To derive the limit of the second term, first note that Assumption 4 implies the following conditional independence,

$$
\{Y_{it}(1), Y_{it}(0)\} \perp\!\!\!\perp \mathbf{X}_i \mid \mathbf{U}_i \tag{44}
$$

The law of iterated expectation implies,

$$\mathbb{E}(Y_{it}(x) \mid C_i = 1) \;=\; \mathbb{E}\{\mathbb{E}(Y_{it}(x) \mid \mathbf{U}_i, C_i = 1) \mid C_i = 1\} \tag{45}$$

for $x = 0, 1$. We show that the difference of means over time within unit i estimates the inner expectation of equation (45) without bias because it adjusts for $\mathbf{U}_i$. Consider the case with $x = 1$.

$$\begin{aligned}
\mathbb{E}\left(C_i \frac{\sum_{t=1}^T X_{it} Y_{it}}{\sum_{t=1}^T X_{it}}\right) &= \mathbb{E}\left\{\frac{1}{\sum_{t=1}^T X_{it}} \sum_{t=1}^T X_{it} \mathbb{E}\left(Y_{it}(1) \mid \mathbf{X}_i, \mathbf{U}_i\right) \;\middle|\; C_i = 1\right\} \Pr(C_i = 1) \\
&= \mathbb{E}\left\{\frac{1}{\sum_{t=1}^T X_{it}} \sum_{t=1}^T X_{it} \mathbb{E}\left(Y_{it}(1) \mid C_i = 1, \mathbf{U}_i\right) \;\middle|\; C_i = 1\right\} \Pr(C_i = 1) \\
&= \mathbb{E}(Y_{it}(1) \mid C_i = 1) \Pr(C_i = 1)
\end{aligned}$$

where the second equality follows from equation (44). We note that $\hat{\tau}_{\mathsf{match}}$ can be more precisely defined as $1/|\mathcal{I}| \sum_{i \in \mathcal{I}} \left\{\sum_{t=1}^T X_{it} Y_{it} / \sum_{t=1}^T X_{it} - \sum_{t=1}^T (1 - X_{it}) Y_{it} / \sum_{t=1}^T (1 - X_{it})\right\}$ where $\mathcal{I} := \{i \in \{1, \ldots, N\} \mid C_i = 1\}$ to avoid the division by zero. A similar argument can be made to show $\mathbb{E}\left(C_i \frac{\sum_{t=1}^T (1 - X_{it}) Y_{it}}{\sum_{t=1}^T (1 - X_{it})}\right) = \mathbb{E}(Y_{it}(0) \mid C_i = 1) \Pr(C_i = 1)$. Thus, the second term of equation (43) converges to $\mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid C_i = 1) \Pr(C_i = 1)$. The result then follows from the continuous mapping theorem. □

A.4 Linear Unit Fixed Effects Regression Estimator as a Naive Matching Estimator

The following proposition establishes that the linear fixed effects regression estimator given in equation (2) can be written as this naive within-unit matching estimator with adjustment. The adjustment is made in order to correct the possible attenuation bias due to the "mismatches" that occur because two observations with the identical treatment status are matched with each other.

PROPOSITION 5 (THE UNIT FIXED EFFECTS ESTIMATOR AS A NAIVE WITHIN-UNIT MATCHING ESTIMATOR WITH ADJUSTMENT) *The following algebraic equality holds,*

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

where $D_{it} = \mathbf{1}\{X_{it} \neq \overline{X}_i\}$, *and for* $x = 0, 1$,

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

Proof

$$\hat{\beta}_{\mathsf{FE}} \;=\; \frac{\sum_{i=1}^N \sum_{t=1}^T (X_{it} - \overline{X}_i)(Y_{it} - \overline{Y}_i)}{\sum_{i=1}^N \sum_{t=1}^T (X_{it} - \overline{X}_i)^2}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(X_{it}-\overline{X}_i)(Y_{it}-\overline{Y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i)^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left\{X_{it}Y_{it} - T\sum_{i=1}^{N}\overline{X}_i\overline{Y}_i\right\}}{NT\overline{X} - T\sum_{i=1}^{N}\overline{X}_i^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(2X_{it}-1)(Y_{it}-\overline{Y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(2X_{it}-1)(X_{it}-\overline{X}_i)} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left\{X_{it}(Y_{it}-\overline{Y}_i) + (1-X_{it})(\overline{Y}_i - Y_{it})\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left\{X_{it}(X_{it}-\overline{X}_i) + (1-X_{it})(\overline{X}_i - X_{it})\right\}} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left\{X_{it}\left(Y_{it} - \frac{1}{T}\sum_{t'\neq t} Y_{it'} - \frac{1}{T}Y_{it}\right) + (1-X_{it})\left(\frac{1}{T}\sum_{t'\neq t} Y_{it'} + \frac{1}{T}Y_{it} - Y_{it}\right)\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left\{X_{it}\left(X_{it} - \frac{1}{T}\sum_{t'\neq t} X_{it'} - \frac{1}{T}X_{it}\right) + (1-X_{it})\left(\frac{1}{T}\sum_{t'\neq t} X_{it'} + \frac{1}{T}X_{it} - Y_{it}\right)\right\}} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \frac{(T-1)}{T} D_{it}\left\{X_{it}\left(Y_{it} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right) + (1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t} Y_{it'} - Y_{it}\right)\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} \frac{(T-1)}{T} D_{it}\left\{X_{it}\left(X_{it} - \frac{1}{T-1}\sum_{t'\neq t} X_{it'}\right) + (1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t} X_{it'} - X_{it}\right)\right\}} \\
&= \frac{\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} - \frac{1}{T-1}\sum_{t'\neq t} Y_{it'}\right) + (1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t} Y_{it'} - 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_{it}\left\{X_{it}\cdot\frac{1}{T-1}\sum_{t'\neq t}(1-X_{it'}) + (1-X_{it})\cdot\frac{1}{T-1}\sum_{t'\neq t} X_{it'}\right\}} \\
&= \frac{1}{K}\left\{\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)\right\}
\end{aligned}
$$

where $\overline{X} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}$. The second and fourth equalities follow from the fact that $X_{it} - \overline{X}_i = D_{it} = 0$ if there is no variation in treatment status for unit i. □

Note that the matched set for each unit consists of all the other unit-unit observations regardless of the treatment status, i.e., $\mathcal{M}_{it}^{\mathsf{LM\text{-}FE}} = \{(i', t') : i' = i, t' \neq t\}$. The proposition shows that the adjustment factor $1/K$ represents the inverse of the proportion of matching weights coming from properly matched observations that have the opposite treatment status. Thus, a greater number of mismatches leads to a smaller value of K, which results in a larger adjustment. As shown earlier, however, in general, this adjustment is not sufficient for yielding a consistent estimate of the ATE under Assumptions 2 and 3.

A.5 Proof of Theorem 1

We begin this proof by establishing two algebraic equalities. First, we prove that for any constants $(\alpha_1^*, \ldots, \alpha_N^*)$, the following equality holds,

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

$$+(1-X_{i't'})\left(\sum_{(i,t)\in\mathcal{M}_{i',t'}}\frac{1}{\#\mathcal{M}_{i't'}}X_{it}\alpha_i^*-\alpha_i^*\right)\Bigg\}$$

$$= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left\{X_{i't'}\left(\alpha_i^*-\alpha_i^*\right)+\left(1-X_{i't'}\right)\left(\alpha_i^*-\alpha_i^*\right)\right\} = 0 \tag{46}$$

where the last equality follows from $\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{\#\mathcal{M}_{i't'}}(1-X_{it})=1$ if $X_{i't'}=1$, and $\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{\#\mathcal{M}_{i't'}}X_{it}=1$ if $X_{i't'}=0$.

Similarly, the second algebraic equality we prove is the following,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}+(1-X_{i't'})w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left[X_{i't'}\left(1+\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{\#\mathcal{M}_{i't'}}(1-X_{it})\right)+(1-X_{i't'})\left(1+\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{\#\mathcal{M}_{i't'}}X_{it}\right)\right] \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left\{X_{i't'}\left(1+1\right)+\left(1-X_{i't'}\right)\left(1+1\right)\right\} = 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}
\end{aligned} \tag{47}$$

Third, we show that $\overline{X}_i^*=1/2$.

$$\begin{aligned}
\overline{X}_i^* &= \frac{\sum_{t=1}^{T} W_{it}X_{it}}{\sum_{t=1}^{T} W_{it}} \\
&= \frac{\sum_{t=1}^{T} D_{it}\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}X_{it}}{\sum_{t=1}^{T} D_{it}\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}} \\
&= \frac{\sum_{t=1}^{T} D_{it}\sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(X_{i't'}w_{it}^{i't'}X_{it}+(1-X_{i't'})w_{it}^{i't'}X_{it}\right)}{\sum_{t=1}^{T} D_{it}\sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(X_{i't'}w_{it}^{i't'}+(1-X_{i't'})w_{it}^{i't'}\right)} \\
&= \frac{\sum_{t=1}^{T} D_{it}\cdot(1+0)}{\sum_{t=1}^{T} D_{it}\cdot(1+1)} \\
&= \frac{1}{2}
\end{aligned}$$

where the fourth equality follows from the fact that (1) $w_{it}^{i't'}=1$ when $X_{it}=X_{i't'}$ and 0 otherwise, and (2) $(1-X_{i't'})X_{it}=0$ if $(i,t)\in\mathcal{M}_{i't'}$ because only the years with opposite treatment status are in the matched set. This implies,

$$X_{it}-\overline{X}_i^* = \begin{cases} \frac{1}{2} & \text{if } X_{it}=1 \\ -\frac{1}{2} & \text{if } X_{it}=0 \end{cases} \tag{48}$$

Using the above algebraic equalities, we can derive the desired result.

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

where the second equalities follows from equation (47) and (48), and the fourth equality from equation (46). The last equality follows from applying the definition of $\widehat{Y_{it}(1)}$ and $\widehat{Y_{it}(0)}$ given in equation (13) □

As an example, we show that when the matched set is given by equation (11), the regression weights equal to the inverse of propensity score computed within each unit. Suppose $X_{it} = 1$. The case for $X_{it} = 0$ is similar.

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

Thus, along with Proposition 2, this implies that, if the data generating process is given by the linear fixed

effects model defined in equation (1) with Assumptions 3 and 4, then the weighted linear unit fixed effects regression estimator with weights inversely proportional to the propensity score is consistent for the average treatment effect β. We note that this weighted linear fixed effects estimator is numerically equivalent to the sample weighted treatment effect estimator of Wooldridge (2005b), which was further studied by Gibbons *et al.* (2011).

A.6 Proof of Proposition 3

Define:

$$\begin{aligned}
\overline{Y_{1i}} &= \sum_{t=1}^{T} X_{it}Y_{it} / \sum_{t=1}^{T} X_{it}, \quad \overline{Y_{0i}} = \sum_{t=1}^{T}(1-X_{it})Y_{it} / \sum_{t=1}^{T}(1-X_{it}) \\
\overline{Y_{1t}} &= \sum_{i=1}^{N} X_{it}Y_{it} / \sum_{i=1}^{N} X_{it}, \quad \overline{Y_{0t}} = \sum_{i=1}^{N}(1-X_{it})Y_{it} / \sum_{i=1}^{N}(1-X_{it}) \\
\overline{Y_1} &= \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}Y_{it} / \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}, \quad \overline{Y_0} = \sum_{i=1}^{N}\sum_{t=1}^{T}(1-X_{it})Y_{it} / \sum_{i=1}^{N}\sum_{t=1}^{T}(1-X_{it}).
\end{aligned}$$

Then, we have,

$$\begin{aligned}
\hat{\beta}_{\mathsf{FE2}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\{(X_{it}-\overline{X}_i)(Y_{it}-\overline{Y}_i)+(X_{it}-\overline{X}_t)(Y_{it}-\overline{Y}_t)-(X_{it}-\overline{X})(Y_{it}-\overline{Y})\}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\{(X_{it}-\overline{X}_i)^2+(X_{it}-\overline{X}_t)^2-(X_{it}-\overline{X})^2\}} \\
&= \frac{(T-1)\sum_{i=1}^{N} S_i^2(\overline{Y_{1i}}-\overline{Y_{0i}})+(N-1)\sum_{t=1}^{T} S_t^2(\overline{Y_{1t}}-\overline{Y_{0t}})-(NT-1)S^2(\overline{Y_1}-\overline{Y_0})}{(T-1)\sum_{i=1}^{N} S_i^2+(N-1)\sum_{t=1}^{T} S_t^2-(N-1)(T-1)S^2} \\
&= \frac{\omega_{\mathsf{FE}}\times\hat{\beta}_{\mathsf{FE}}+\omega_{\mathsf{FEtime}}\times\hat{\beta}_{\mathsf{FEtime}}-\omega_{\mathsf{pool}}\times\hat{\beta}_{\mathsf{pool}}}{w_{\mathsf{FE}}+w_{\mathsf{FEtime}}-w_{\mathsf{pool}}}
\end{aligned}$$

where

$$\begin{aligned}
\omega_{\mathsf{FE}} &= \frac{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{T}\right)\left(1-\frac{1}{N}\right)S^2} \\
\omega_{\mathsf{FEtime}} &= \frac{\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{T}\right)\left(1-\frac{1}{N}\right)S^2} \\
\omega_{\mathsf{pool}} &= \frac{\left(1-\frac{1}{NT}\right)S^2}{\left(1-\frac{1}{T}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2+\left(1-\frac{1}{N}\right)\frac{1}{T}\sum_{t=1}^{T} S_t^2-\left(1-\frac{1}{T}\right)\left(1-\frac{1}{N}\right)S^2}
\end{aligned}$$

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

A.7 Proof of Proposition 4

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

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

$$
\begin{aligned}
&\qquad\qquad - \left(\frac{1}{N} \sum_{i' \neq i} Y_{i't} - \frac{1}{NT} \sum_{i' \neq i} Y_{i't} \right) + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'} \Bigg\} \\
&\quad -(1 - X_{it}) \left\{ Y_{it} \left(1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT} \right) - \left(\frac{1}{T} \sum_{t' \neq t} Y_{it'} - \frac{1}{NT} \sum_{t' \neq t} Y_{it'} \right) \right. \\
&\qquad\qquad \left. - \left(\frac{1}{N} \sum_{i' \neq i} Y_{i't} - \frac{1}{NT} \sum_{i' \neq i} Y_{i't} \right) + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'} \right\} \Bigg] \\
= \; & \sum_{i=1}^{N} \sum_{t=1}^{T} \left[X_{it} \left\{ \frac{(N-1)(T-1)}{NT} Y_{it} - \frac{N-1}{NT} \sum_{t' \neq t} Y_{it'} - \frac{T-1}{NT} \sum_{i' \neq i} Y_{i't} + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'} \right\} \right. \\
&\qquad \left. -(1 - X_{it}) \left\{ \frac{(N-1)(T-1)}{NT} Y_{it} - \frac{N-1}{NT} \sum_{t' \neq t} Y_{it'} - \frac{T-1}{NT} \sum_{i' \neq i} Y_{i't} + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i't'} \right\} \right] \\
= \; & \frac{(T-1)(N-1)}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ X_{it} \left(Y_{it} - \frac{\sum_{t'=1}^{T} Y_{it'}}{T-1} + \frac{\sum_{i'=1}^{N} Y_{i't}}{N-1} - \frac{\sum_{i' \neq i}^{N} \sum_{t' \neq t}^{T} Y_{i't'}}{(T-1)(N-1)} \right) \right. \\
&\qquad \left. - (1 - X_{it}) \left(\frac{\sum_{t'=1}^{T} Y_{it'}}{T-1} + \frac{\sum_{i'=1}^{N} Y_{i't}}{N-1} - \frac{\sum_{i' \neq i}^{N} \sum_{t' \neq t}^{T} Y_{i't'} - Y_{it}}{(T-1)(N-1)} \right) \right\}. \\
= \; & \frac{(T-1)(N-1)}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)} \right)
\end{aligned} \tag{49}
$$

The second algebraic equality we prove is the following,

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

$$
\begin{aligned}
&= \frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left\{X_{it}\left(\frac{\sum_{t'=1}^{T}(1-X_{it'})}{T-1}+\frac{\sum_{i'=1}^{N}(1-X_{i't})}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}(1-X_{i't'})}{(T-1)(N-1)}\right)\right.\right.\\
&\qquad\left.\left.+(1-X_{it})\left(\frac{\sum_{t'=1}^{T}X_{it'}}{T-1}+\frac{\sum_{i'=1}^{N}X_{i't}}{N-1}-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}X_{i't'}}{(T-1)(N-1)}\right)\right\}\right]\\
&= K(T-1)(N-1) \qquad (50)
\end{aligned}
$$

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

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

where the last equality follows from equation (49) and (50). □

A.8 The Adjusted Two-way Matching Estimator

We show 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. For a given observation (i,t) for which the counterfactual outcome needs to be estimated, the within-unit matched set $\mathcal{M}_{it}$ and the within-time matched set $\mathcal{N}_{it}$ consists only of observations with the opposite treatment status. $\mathcal{A}_{it}$ is then defined according to equation (29).

PROPOSITION 6 (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 $D_{it} = \mathbf{1}\{\#\mathcal{M}_{it}\cdot\#\mathcal{N}_{it}>0\}$*, and for* $x=0,1$*,*

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

and $a_{it} = \#\{(i',t') \in \mathcal{A}_{it} : X_{i't'} = X_{it}\}$*. 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*

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

Proof We first establish the following equality.

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
=\;& \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}\right) \\
=\;& \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
=\;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{M}_{i't'}\cdot\#\mathcal{N}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{N}_{i't'}\cdot\#\mathcal{M}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{(a_{i't'}-\#\mathcal{A}_{i't'}+a_{i't'})}{\#\mathcal{A}_{i't'}+a_{i't'}}\right)\right. \\
&\left. +(1-X_{i't'})\left(\frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{M}_{i't'}\cdot\#\mathcal{N}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{N}_{i't'}\cdot\#\mathcal{M}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} - \frac{(\#\mathcal{A}_{i't'}-a_{i't'}-a_{i't'})}{\#\mathcal{A}_{i't'}+a_{i't'}}\right)\right\} \\
=\;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left\{X_{i't'}\left(\frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{(a_{i't'}-\#\mathcal{A}_{i't'}+a_{i't'})}{\#\mathcal{A}_{i't'}+a_{i't'}}\right)\right. \\
&\left. +(1-X_{i't'})\left(\frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} + \frac{\#\mathcal{A}_{i't'}}{\#\mathcal{A}_{i't'}+a_{i't'}} - \frac{(\#\mathcal{A}_{i't'}-a_{i't'}-a_{i't'})}{\#\mathcal{A}_{i't'}+a_{i't'}}\right)\right\} \\
=\;& \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(2X_{i't'} + 2(1-X_{i't'})\right) = 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}. \qquad (51)
\end{aligned}$$

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 and the proof of Proposition 1, it is straightforward to show that $\overline{X}_i^* = \overline{X}_t^* = \overline{X}^* = \frac{1}{2}$, and thus

$$X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^* = \begin{cases} \frac{1}{2} & \text{if } X_{it} = 1 \\ -\frac{1}{2} & \text{if } X_{it} = 0 \end{cases} \tag{52}$$

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.

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

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

where the second and third equality follows from equation (51) and (52). 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 6. □

Unlike Proposition 4, the adjustment is done by deflating the estimated treatment effect by $1/K_{it}$. This is because the attenuation bias from $\mathcal{A}_{it}$ (the "pooled" part) is *subtracted* from the sum of two unbiased estimates from $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$, amplifying the estimated treatment effect for the observation. In the example of Panel (b) of Figure 5, $\mathcal{A}_{it}$ 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.

A.9 Proof of Theorem 2

The proof of this theorem follows directly from Proposition 6 as the within-unit and within-time matched sets are subsets of $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$. Specifically, $\mathcal{M}_{it}^{\mathsf{DiD}}$ 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 $\mathcal{N}_{it}^{\mathsf{DiD}}$ 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}_{i't'}^{\mathsf{DiD}}} Y_{it'}}{\#\mathcal{N}_{i't'}^{\mathsf{DiD}}} + \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}^{\mathsf{DiD}}} Y_{it}}{\#\mathcal{A}_{i't'}^{\mathsf{DiD}}} \right) \right. \\
&\qquad \left. + (1 - X_{i't'}) \left(Y_{i',t'-1} + \frac{\sum_{(i,t')\in\mathcal{N}_{i't'}^{\mathsf{DiD}}} Y_{it'}}{\#\mathcal{N}_{i't'}^{\mathsf{DiD}}} - \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}^{\mathsf{DiD}}} Y_{it}}{\#\mathcal{A}_{i't'}^{\mathsf{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}_{\mathsf{DID}}
\end{aligned}$$

where the seventh equality follows from the fact that, given $\mathcal{M}_{i't'}^{\mathsf{DiD}}$ and $\mathcal{N}_{i't'}^{\mathsf{DiD}}$, all the units in $\mathcal{A}_{i't'}^{\mathsf{DiD}}$ are under the opposite treatment status (i.e., $a_{i't'} = 0$), and thus $K_{i't'} = 1$ (see Proposition 6). □

A.10 Estimated Effects of GATT on Bilateral Trade with Various Membership Definitions

		Dyad Fixed Effects			Year Fixed Effects		Dyad and Year Fixed Effects		
Comparison	Membership	Standard	Weighted	First Diff.	Standard	Weighted	Standard	DiD	DiD-caliper
	Formal	−0.048	−0.069	0.075	0.004	−0.002	0.036	0.003	0.044
	(N=196,207)	(0.025)	(0.023)	(0.054)	(0.031)	(0.030)	(0.025)	(0.040)	(0.149)
	White's p-value		0.064	0.000		1.000		0.318	0.333
Both vs. Mix	N (non-zero weights)	196,207	110,195	6,846	196,207	196,207	196,207	92,043	88,565
	Participants	0.147	0.011	0.096	0.199	0.193	0.227	0.043	0.065
	(N=196,207)	(0.031)	(0.029)	(0.030)	(0.034)	(0.035)	(0.031)	(0.046)	(0.135)
	White's p-value		0.000	0.102		0.998		0.003	0.006
	N (non-zero weights)	196,207	68,004	3,952	196,207	196,207	196,207	59,732	53,843
	Formal	−0.034	−0.061	0.076	−0.006	−0.005	0.030	0.001	0.026
	(N=175,814)	(0.025)	(0.023)	(0.055)	(0.031)	(0.031)	(0.025)	(0.038)	(0.043)
	White's p-value		0.031	0.000		1.000		0.721	0.002
Both vs. One	N (non-zero weights)	175,814	100,055	6,712	175,814	175,814	175,814	75,873	72,983
	Participants	0.161	0.020	0.099	0.180	0.174	0.234	0.035	0.034
	(N=187,651)	(0.031)	(0.029)	(0.030)	(0.035)	(0.036)	(0.031)	(0.046)	(0.036)
	White's p-value		0.000	0.086		0.999		0.001	0.000
	N (non-zero weights)	187,651	64,152	3,900	187,651	187,651	187,651	54,364	48,541
	Formal	−0.011	−0.094	0.031	0.007	0.046	0.039	−0.041	−0.060
	(N=109,702)	(0.041)	(0.041)	(0.067)	(0.053)	(0.056)	(0.041)	(0.819)	(0.318)
	White's p-value		0.058	0.000		0.276		0.141	0.132
One vs. None	N (non-zero weights)	109,702	36,115	2,670	109,702	109,702	109,702	16,940	15,671
	Participants	0.181	−0.034	0.053	0.163	0.171	0.216	0.028	0.160
	(N=70,298)	(0.062)	(0.058)	(0.063)	(0.072)	(0.079)	(0.063)	(0.121)	(0.590)
	White's p-value		0.004	0.000		0.046		0.979	1.000
	N (non-zero weights)	70,298	15,766	1,087	70,298	70,298	70,298	13,194	8,032
covariates		year-varying covariates			dyad-varying covariates		year-varying covariate		

Table 1: **Estimated Effects of GATT on the Logarithm of Bilateral Trade based on Various Fixed Effects Estimators**: For estimators with either dyad or year fixed effects, the "Weighted" columns present the estimates based on the regression weights given in equation (20), which yield the estimated average treatment effects based on equation (21). Other weighted fixed effects estimators, i.e., first differences and difference-in-differences, are also presented. "Both vs. Mix" ("Both vs. One") represents the comparison between dyads of two GATT members and those consisting of either one or no (only one) GATT member. "One vs. None" refers to the comparison between dyads consisting of only one GATT member and those of two non-GATT members. "Formal" membership includes only formal GATT members as done in Rose (2004), whereas "Participants" includes nonmember participants as defined in Tomz *et al.* (2007). The results from an alternative way to adjust for past outcomes when employing the multi-period DiD design are presented in "DiD-caliper" column where matched units' pre-treatment outcomes are within two-tenth of the standard deviation of the outcome variable away from that of a treated unit. The covariates include GSP (Generalized System of Preferences), log product real GDP, log product real GDP per capita, regional FTA (Free Trade Agreement), currency union, currently colonized, log distance, common language, land border, number landlocked, number of islands, log product land area, common colonizer, past colonial relationship, and common country. White's p-value is based on the specification test with the null hypothesis that the corresponding standard fixed effects model is correct. Robust standard errors, allowing for the presence of serial correlation as well as heteroskedasticity (Arellano, 1987; Hansen, 2007), are in parentheses. The results suggest that different causal assumptions, which imply different regression weights, can yield different results.