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

Kosuke Imai† In Song Kim‡

First Draft: July 1, 2014
This Draft: December 19, 2017

Abstract

Many researchers use unit fixed effects regression models as their default methods for causal inference with longitudinal data. We show that the ability of these models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationships, which are allowed to exist under an alternative selection-on-observables approach. Using the nonparametric directed acyclic graph, we highlight the two key causal identification assumptions of fixed effects models: past treatments do not directly influence current outcome, and past outcomes do not affect current treatment. Furthermore, we introduce a new nonparametric matching framework that elucidates how various fixed effects models implicitly compare treated and control observations to draw causal inference. By establishing the equivalence between matching and weighted fixed effects estimators, this framework enables a diverse set of identification strategies to adjust for unobservables provided that the treatment and outcome variables do not influence each other over time. We illustrate the proposed methodology through its application to the estimation of GATT membership effects on dyadic trade volume.

Key Words: before-and-after design, directed acyclic graph, matching, panel data, unobserved confounders, 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

Unit 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 adjust for unobserved time-invariant confounders when estimating causal effects from observational data. In spite of this widespread practice, much methodological discussion of unit fixed effects models in political science has taken place from model-based perspectives (often assuming linearity) with little attention to the causal identification assumptions (e.g., Beck, 2001; Wilson and Butler, 2007; Bell and Jones, 2015; Clark and Linzer, 2015). In contrast, 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).

Specifically, we show that the ability of unit fixed effects regression models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationships between treatment and outcome variables, which are allowed to exist under an alternative selection-on-observables approach (e.g., Robins *et al.*, 2000). Our analysis highlights the two key causal identification assumptions that are required under fixed effects models and yet are often overlooked by applied researchers: (1) past treatments do not directly influence current outcome, and (2) past outcomes do not affect current treatment. Unlike most of the existing discussions of unit fixed effects regression models that assume linearity, we use the directed acyclic graph (DAG) framework (Pearl, 2009) that can represent a wide class of nonparametric structural equation models, encompassing linear and other fixed effects models as special cases.

In addition, we propose a new analytical framework that directly connects fixed effects models to matching methods (e.g., Rubin, 2006; Ho *et al.*, 2007; Stuart, 2010). The framework makes explicit how counterfactual outcomes are estimated under unit fixed effects regression models. A simple but important insight is that the comparison of treated and control observations must occur within the same unit and across time periods in order to adjust for unobserved unit-specific time-invariant confounders. We establish this fact by proving the equivalence between within-unit matching estimators and weighted linear fixed effects regression estimators. The result implies, for example, that the counterfactual outcome for a treated observation in a given time period is estimated using the observed outcomes of different time periods of the same unit.

Our matching framework incorporates a diverse set of identification strategies to adjust for unobservables provided that the treatment and outcome variables do not influence each other over time. We show how to derive the weighted linear fixed effects regression estimator that is equivalent to a given within-unit matching estimator. This equivalence allows us to construct simple model-based standard errors instead of more complex and computationally intensive standard errors proposed in the literature (e.g., Abadie and Imbens, 2006, 2012; Otsu and Rai, 2017). In addition, we can use the model-based specification test to assess the appropriateness of linearity assumption in fixed effects regression models (White, 1980a). 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). The proposed methodology is freely available as an R package, `wfe: Weighted Linear Fixed Effects Estimators for Causal Inference`, at the Comprehensive R Archive Network (http://cran.r-project.org/package=wfe).

Finally, we illustrate the proposed methodology by applying it to the controversy regarding the causal effects of GATT membership on dyadic trade (Rose, 2004; Tomz *et al.*, 2007). Despite the substantive disagreement, there exists a remarkable methodological consensus among researchers in the literature, all of whom endorse the use of linear fixed effects regression models. We critically examine the causal identification assumptions of the models used in previous studies and also consider an alternative identification strategy. We show that the empirical conclusions are highly dependent on the choice of causal identification assumptions.

2 Causal Assumptions

We study the causal assumptions of regression models with unit fixed effects. While we begin our discussion 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). We show that the ability of unit fixed effects models to adjust for unobserved time-invariant confounders comes at the expense of dynamic causal relationship between treatment and outcome variables.

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 from a population 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 specification:

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

for each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, 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 with $\mathbb{E}(\epsilon_{it}) = 0$. In this model, the unit fixed effect α_i captures a vector of unobserved time-invariant confounders in a flexible manner. That is, 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. Formally, this assumption can be written as,

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \alpha_i) \;=\; 0 \tag{2}$$

for each $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$, where $\mathbf{X}_i$ is a $T \times 1$ vector of treatment variables for unit i. Since α_i can be any function of $\mathbf{U}_i$, this assumption is equivalent to $\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{U}_i) = 0$.

We refer to this model based on equations (1) and (2) as $\mathsf{LIN\text{-}FE}$. 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,

$$\hat{\beta}_{\mathsf{LIN\text{-}FE}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\{(Y_{it} - \overline{Y}_i) - \beta(X_{it} - \overline{X}_i)\}^2 \quad (3)$$

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 $\mathsf{LIN\text{-}FE}$, then $\hat{\beta}_{\mathsf{LIN\text{-}FE}}$ 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 (3) shows that units with no variation in the treatment variable do not contribute to the estimation of β. Thus, under $\mathsf{LIN\text{-}FE}$, the causal estimand is the following average treatment effect among the units with some variation in the treatment status.

$$\tau = \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid C_i = 1) \quad (4)$$

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

2.2 Nonparametric Causal Identification Analysis

To understand the fundamental causal assumptions of unit fixed effects models, we conduct a nonparametric identification analysis that avoids any parametric restriction. Specifically, we relax the linearity assumption of $\mathsf{LIN\text{-}FE}$ (i.e., equation (1)). We also generalize mean independence (i.e., equation (2)) to statistical independence. The resulting model is the following nonparametric fixed effects model ($\mathsf{NP\text{-}FE}$).

Assumption 1 (Nonparametric Fixed Effects Model) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$,*

$$Y_{it} = g(X_{it}, \mathbf{U}_i, \epsilon_{it}) \quad (5)$$

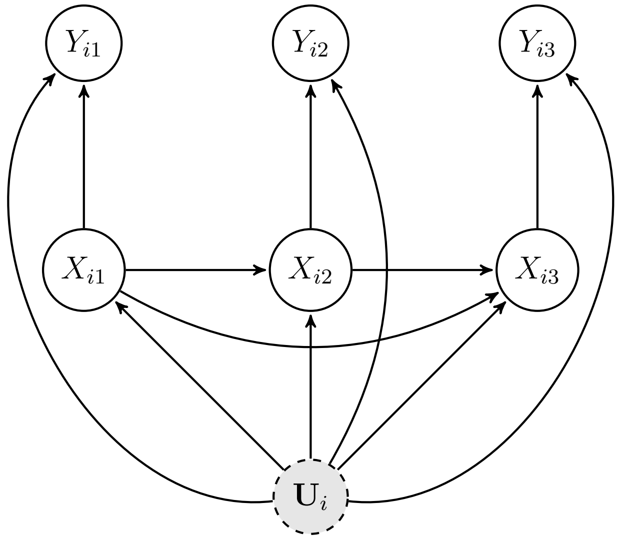


Figure 1: **Directed Acyclic Graph for Regression Models with Unit Fixed Effects based on Three Time Periods.** Solid circles represent observed outcome Y_{it} and treatment X_{it} variables whereas a grey dashed circle represents a vector of unobserved time-invariant confounders $\mathbf{U}_i$. The 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.

$$\epsilon_{it} \perp\!\!\!\perp \{\mathbf{X}_i, \mathbf{U}_i\} \tag{6}$$

where $g(\cdot)$ can be any function.

Note that NP-FE includes LIN-FE as a special case. Unlike LIN-FE, NP-FE does not assume a functional form and enables all effects to vary across observations.

We examine causal assumptions of NP-FE using the directed acyclic graphs (DAGs). Pearl (2009) shows that a DAG can formally represent a nonparametric structural equation model (NPSEM) avoiding functional-form and distributional assumptions while allowing for general forms of effect heterogeneity.[1] The DAG in Figure 1 graphically represents the NPSEM that corresponds to NP-FE. 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., $T > 3$. 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,

[1] More precisely, NP-FE implies $X_{it} = f(X_{i1}, \ldots, X_{i,t-1}, \mathbf{U}_i, \eta_{it})$ where η_{it} is an exogenous disturbance term and $f(\cdot)$ is any function.

$\mathbf{U}_i$. 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.

The DAG in Figure 1 shows that Assumption 1 of NP-FE can be understood as the following set of statements, each of which is represented by the absence of corresponding nodes or arrows:

Assumption (a) no unobserved time-varying confounder exists

Assumption (b) past outcomes do not directly affect current outcome

Assumption (c) past outcomes do not directly affect current treatment

Assumption (d) past treatments do not directly affect current outcome

No additional arrows can be added to the DAG without making it inconsistent with NP-FE. In particular, no additional arrows that point to X_{it} can be included in the DAG without violating the strict exogeneity assumption of ϵ_{it} under NP-FE. 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}.[2]

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 (5) 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.

[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.

Assumption 2 (No carryover effect) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, the potential outcome is given by,*

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

To better understand the assumed treatment assignment mechanism, we consider a randomized experiment, to which NP-FE is applicable. This experiment can be described as follows: 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}, but without conditioning on the previous outcome Y_{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 can formalize this treatment assignment mechanism as follows.

Assumption 3 (Sequential Ignorability with Unobserved Time-invariant Confounders) *For each $i = 1, 2, \ldots, N$,*

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

Thus, Assumption 2 corresponds to Assumption (d) of NP-FE and is implied by equation (1) of LIN-FE. In addition, Assumption 3 corresponds to Assumptions (a) and (c) of NP-FE and the strict exogeneity assumption of LIN-FE given in equation (2) (see Appendix A.1 for a proof).

2.3 Which Causal Identification Assumptions Can be Relaxed?

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

First, note that 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 causal identification. To see this, suppose that past outcomes directly affect current outcome as in Figure 2(a). Even under 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] The existence of such a relationship, however, may necessitates the adjustment of standard errors, for example, via cluster robust standard errors.

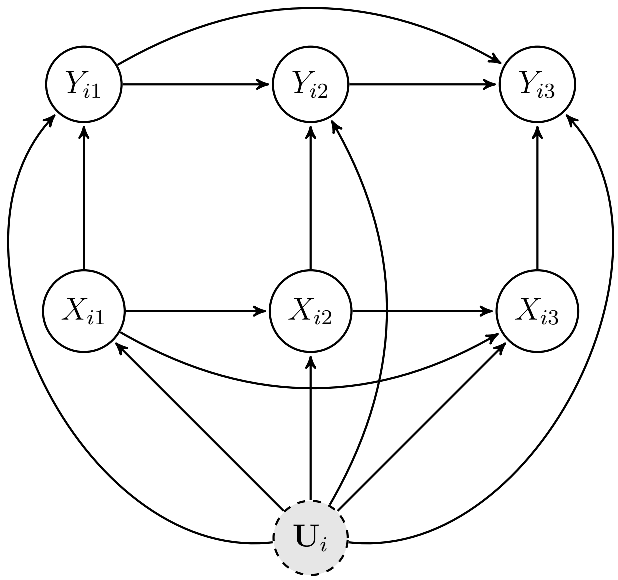
Next, we entertain the scenario in which past treatments directly affect current outcome (i.e., relaxing Assumption (d)). Typically, applied researchers address this possibility by including lagged treatment variables in LIN-FE. Here, we consider the following model with one period lag.

$$Y_{it} \quad = \quad \alpha_i + \beta_1 X_{it} + \beta_2 X_{i,t-1} + \epsilon_{it} \tag{7}$$

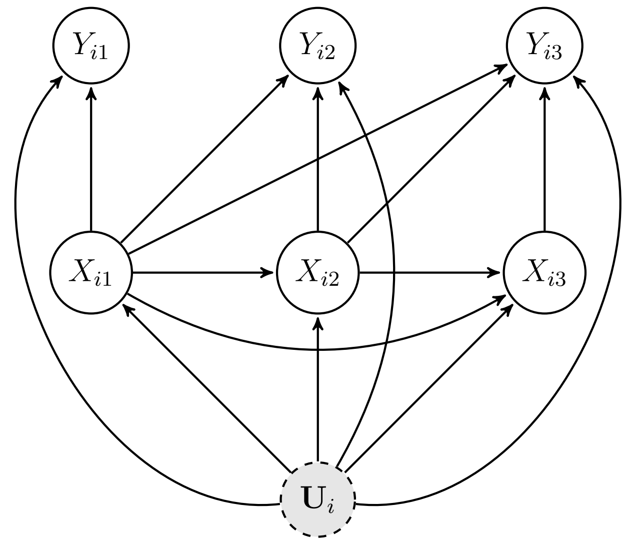
The model implies that the potential outcome can be written as a function of the contemporaneous and previous treatments, i.e., $Y_{it}(X_{i,t-1}, X_{it})$, rather than the contemporaneous treatment alone, partially relaxing Assumption 2.

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

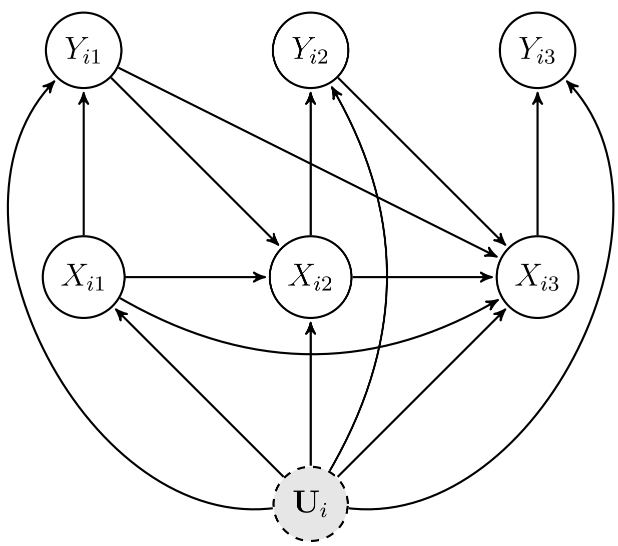
[4]The application of the adjustment criteria (Shpitser *et al.*, 2010) implies that these additional causal relationships do not violate Assumption 3 since every non-causal path between the treatment X_{it} and any outcome $Y_{it'}$ is blocked where $t \neq t'$.



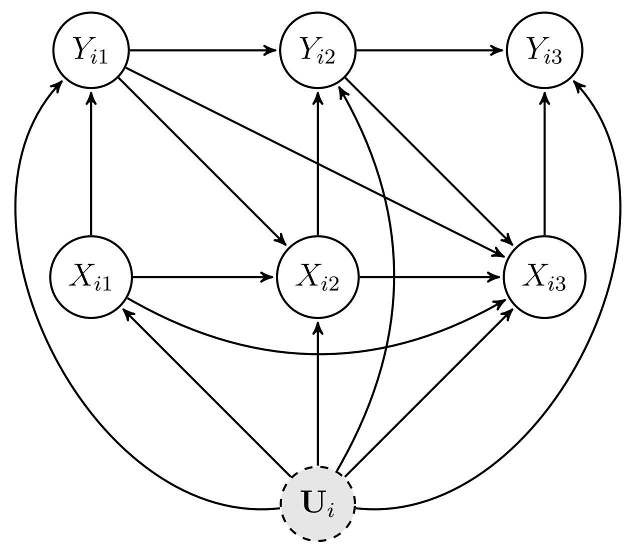
(a) past outcome affects current outcome



(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).

NP-FE replacing equation (5) with the following alternative model for the outcome,

$$Y_{it} \quad = \quad g(X_{i1}, \ldots, X_{it}, \mathbf{U}_i, \epsilon_{it}) \tag{8}$$

It can be shown that under this NPSEM Assumption 3 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 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. Indeed, a frequent practice is to adjust for one time period lag. 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 (7). However, typically the number of lagged treatments to be included is arbitrarily chosen and is rarely justified on substantive grounds.[6]

[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}. This result also holds even if past outcomes affect current outcomes (i.e., without Assumption (b)).

[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.

Finally, we consider relaxing the assumption that past outcomes do not directly affect current treatment (Assumption (c)). This scenario is depicted as Figure 2(c). It is immediate that Assumptions 3 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 the unit fixed effects models.

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

$$Y_{it} \quad = \quad \alpha_i + \beta X_{it} + \rho Y_{i,t-1} + \epsilon_{it} \tag{9}$$

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). The results of Brito and Pearl (2002) imply that we can identify the average causal effect of X_{i3} on Y_{i3} by using X_{i1}, X_{i2}, and Y_{i1} as instrumental variables while conditioning 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, in practice, these assumptions are often made without a substantive justification.

In sum, three key causal identification assumptions are required for LIN-FE and its nonparametric generalization NP-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 the 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. The use of instrumental variables is a popular approach to relax the latter assumption, but under this

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

[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.

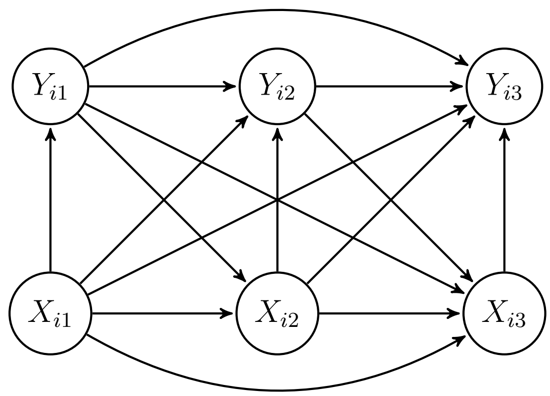


Figure 3: Directed Acyclic Graph for Marginal Structural Models (MSMs). When compared to regression models with unit fixed effects (Figure 1), MSMs assume the absence of unobserved time-invariant confounders $\mathbf{U}_i$ but relax the other assumptions by allowing the past treatments to affect the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

approach we must instead assume that lagged outcome variables do not directly affect current outcome. Unfortunately, researchers rarely justify these alternative identification strategies on substantive grounds.

2.4 Trade-off between Causal Dynamics and Time-invariant Unobservables

The above causal identification analysis shows that the validity of unit fixed effects models critically hinges upon the assumed absence of dynamic causal relationships between treatment and outcome variables. Indeed, there exists a key trade-off between causal dynamics and time-invariant unobservables. We show below that an alternative selection-on-observable approach allows for the existence of such dynamic causal relationships even though it assumes the absence of time-invariant unobservables. The key decision for applied researchers is then whether, in a given substantive problem, they believe causal dynamics is more important than time-invariant unobserved confounders.

A prominent causal model based on this alternative selection-on-observables approach is the marginal structural models (MSMs) developed in epidemiology (Robins *et al.*, 2000), which was recently introduced to political science (Blackwell, 2013). Figure 3 presents a DAG for the MSMs.

First, unlike NP-FE, the MSMs assume the absence of unobserved time-invariant confounders $\mathbf{U}_i$ (both the MSMs and NP-FE assume the absence of unobserved time-varying confounders). Second, the MSMs relax Assumption 2, 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 $(X_{i1}, \ldots, X_{it}) = (x_1, \ldots, x_t)$.

Third, under the MSMs, past outcomes can affect current treatment as well as current outcome. In 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 4 (Sequential Ignorability with Past Outcomes (Robins *et al.*, 2000)) *For* $i = 1, 2, \ldots, N$ *and* $t = 1, 2, \ldots, T$,

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

Unlike Assumption 3, Assumption 4 conditions on past outcomes as well as past treatments. However, under Assumption 4, we cannot adjust for unobserved time-invariant confounders $\mathbf{U}_i$.

In summary, the ability of LIN-FE and its nonparametric generalization NP-FE to adjust for unobserved time-invariant confounders comes at the cost: we must assume that past treatments

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

do not directly affect current outcome and past outcomes do not directly affect current treatment. In contrast, the selection-on-observables approach such as the MSMs, 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 fixed effects regression models. Since fixed effects models can only adjust for unobserved confounders that are time-invariant, applied researchers often measure a vector of observed time-varying confounders $\mathbf{Z}_{it}$ to improve the credibility of assumptions.[10] We show here that the main conclusion of the above identification analysis remains unchanged even if we include these additional observed time-varying confounders as covariates of the fixed effects regression models. In fact, in this case, we must make an additional assumption that there exists no dynamic causal relationship between outcome and these time-varying confounders, leading to potentially even less credible causal identification in many applications.

Consider the following NP-FE, which now includes observed time-varying confounders $\mathbf{Z}_{it}$.

Assumption 5 (Nonparametric Fixed Effects Model with Observed Time-varying Confounders) *For each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$,*

$$\begin{aligned} Y_{it} &= g(X_{i1}, \ldots, X_{it}, \mathbf{U}_i, Z_{i1}, \ldots, Z_{it}, \epsilon_{it}) && (10) \\ \epsilon_{it} &\perp\!\!\!\perp \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i\} && (11) \end{aligned}$$

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

Figure 4 presents a DAG that corresponds to a nonparametric structural equation model consistent with this NP-FE. 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}, the 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 $\mathbf{U}_i$ can directly affect these observed time-varying confounders. Under

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

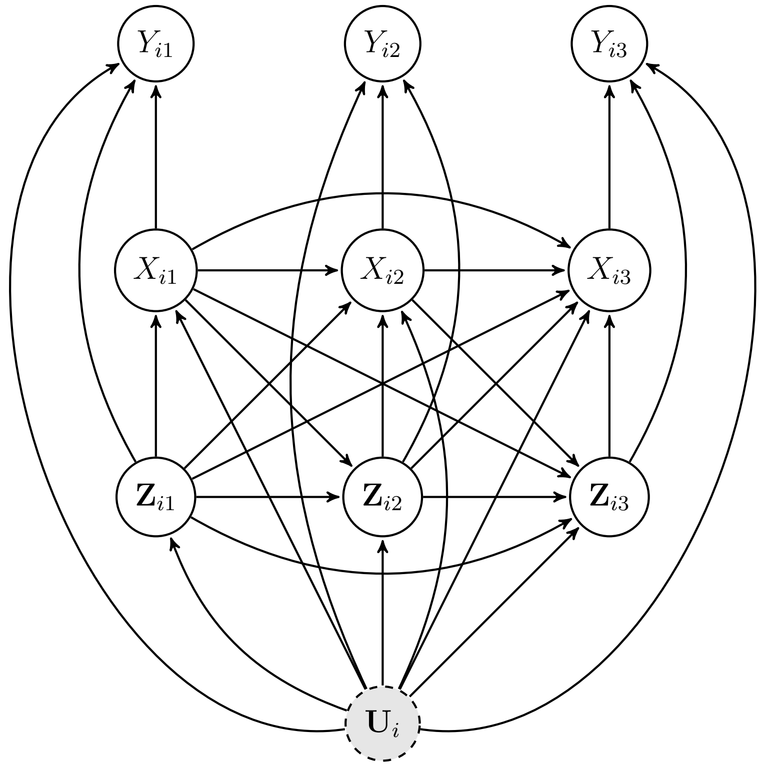


Figure 4: Directed Acyclic Graph for Regression Models with Unit Fixed Effects and Observed Time-varying Confounders based on Three Time Periods.

this model, 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 need to be adjusted because they do not directly affect current outcome Y_{it}.

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, i.e., $\mathbf{Z}_{it'}$ with $t' < t$, in fixed effects regression models. 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. 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 NP-FE without time-varying confounders, the average contemporaneous treatment effect of 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 5. 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 outcome and treatment variables. 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 present. 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 selection-on-observables approach, such as the MSMs discussed in Section 2.4, can relax these assumptions so long as there exists no unobserved time-invariant confounder. Under this alternative approach, past treatments can directly affect current outcome and past outcomes can either directly or indirectly affect current treatment (through time-varying confounders).

3 A New Matching Framework

Causal inference is all about the question of how to credibly estimate the counterfactual outcomes through the comparison of treated and control observations. For a treated observation, we observe the outcome under the treatment condition but must infer its counterfactual outcome under the control condition using the observed outcomes of control observations. Matching is a class of nonparametric methods where we solve this fundamental problem of causal inference by finding a set of control observations similar to each treated observation (e.g., Rubin, 2006; Ho *et al.*, 2007;

Stuart, 2010).

In this section, we propose a new matching framework to shed new light on the causal identification assumptions of fixed effects regression models. We show that causal inference based on unit fixed effects regression models relies upon within-unit comparison where a treated observation is matched with the control observations of the same unit at different time periods. We establish this fact by proving that a within-unit matching estimator is equivalent to a weighted linear unit fixed effects regression model. Like the DAGs introduced above, this new matching framework is completely nonparametric and can accommodate a variety of identification strategies based on within-unit comparison.

3.1 The Within-Unit Matching Estimator

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

Proposition 1 (Inconsistency of the Linear Fixed Effects Regression Estimator (Chernozhukov *et al.*, 2013, Theorem 1)) *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 2 and 3 as well as simple random sampling of units with* T *fixed, the linear fixed effects regression estimator given in equation* (3) *is inconsistent for the average treatment effect* τ *defined in equation* (4),

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

Thus, in general, under Assumptions 2 and 3, the linear unit fixed effects estimator fails 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 LIN-FE in a cross-sectional context. For example, LIN-FE is often used to analyze stratified randomized experiments (Duflo *et al.*, 2007). Even in this case, if the proportion of treated observations 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 2 and 3 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.

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

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

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* (12) *is consistent for the average treatment effect defined in equation* (4).

Proof is in Appendix A.2.

To further generalize this idea, 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,

$$\mathcal{M}_{it} = \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}\} \tag{13}$$

Thus far, we focused on the average treatment effect as a parameter of interest given that researchers often interpret the parameter β of LIN-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,

$$\hat{\tau} \quad = \quad \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}) \tag{14}$$

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 $X_{it} = 1 - x$,

$$\widehat{Y_{it}(x)} \quad = \quad \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{|\mathcal{M}_{it}|}\sum_{(i',t')\in\mathcal{M}_{it}} Y_{i't'} & \text{if } X_{it} = 1 - x \end{cases} \tag{15}$$

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 the matched set defined in equation (13), we have $D_{it} = C_i$ for any t.

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 show how to 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,

$$\mathcal{M}_{it} \quad = \quad \{(i', t') : i' = i, X_{i't'} = 1 - X_{it}, \mathcal{D}(\mathbf{Z}_{it}, \mathbf{Z}_{i't'}) = J_{it}\} \tag{16}$$

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

$$J_{it} \quad = \quad \min_{(i,t')\in\mathcal{M}_{it}} \mathcal{D}(\mathbf{Z}_{it}, \mathbf{Z}_{it'}) \tag{17}$$

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 (14). 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 of no time trend can be written as,

Assumption 6 (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 2 and 6, 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,

$$\mathcal{M}_{it} \quad = \quad \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 1 - X_{it}\} \tag{18}$$

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

$$\hat{\beta}_{\mathsf{FD}} \quad = \quad \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=2}^{T} \{(Y_{it} - Y_{i,t-1}) - \beta(X_{it} - X_{i,t-1})\}^2 \tag{19}$$

In standard econometrics textbooks, the FD estimator defined in equation (19) is presented as an alternative estimation method for the LIN-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 LIN-FE and FD estimators match observations within the same unit. However, the FD estimator matches observations only from subsequent time periods whereas the LIN-FE estimator matches observations from all time periods regardless of the treatment status.

There exists an important limitation of the BA design. While Assumption 6 is written in terms of time trend of potential outcomes, the assumption is violated if past outcomes affect current treatment. Since the first-difference estimator can be thought of as a special case of linear unit fixed effects regression estimators, the critical assumption of no dynamic causal relationships between the outcome and treatment variables remains relevant under the BA design. 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.

Finally, we can generalize the BA design so that we use a larger number of lags to estimate the counterfactual outcome. If we let L to represent the number of lags, then the matched set becomes,

$$\mathcal{M}_{it} \quad = \quad \{(i', t') : i' = i, t' \in \{t-1, \ldots, t-L\}, X_{i't'} = 1 - X_{it}\} \tag{20}$$

Furthermore, the causal quantity of interest can also be generalized to a longer-term average treatment effect, i.e.,

$$\mathbb{E}\{Y_{i,t+F}(1) - Y_{i,t+F}(0) \mid X_{it} \neq X_{i,t-1}\} \tag{21}$$

where F is a non-negative integer representing the number of leads. Under this setting, we estimate the potential outcome under $X_{it} = x$ at time $t+F$ using the following matching estimator,

$$\hat{\tau} = \frac{1}{\sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}} \sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}(\widehat{Y_{i,t+F}(1)} - \widehat{Y_{i,t+F}(0)}) \tag{22}$$

where

$$\widehat{Y_{i,t+F}(x)} = \begin{cases} Y_{i,t+F} & \text{if } X_{it} = 1 - X_{i,t-1} = x \\ \frac{1}{|\mathcal{M}_{it}|}\sum_{(i',t')\in\mathcal{M}_{it}} Y_{i't'} & \text{if } X_{it} = 1 - x \end{cases}, \quad D_{it} = \begin{cases} 1 & \text{if } X_{it} = 1 - X_{i,t-1} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

3.3 Estimation, Inference, and Specification Test

As the 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 Unit 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,*

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

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,

$$W_{it} = D_{it}\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{where} \quad w_{it}^{i't'} = \begin{cases} 1 & \text{if } (i,t) = (i',t') \\ 1/|\mathcal{M}_{i't'}| & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

Proof is in Appendix A.3. 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.

For example, it can be shown that when the matched set is given by equation (13), the regression weights equal to the inverse of propensity score computed within each unit. Thus, along with Proposition 2, this implies that, if the data generating process is given by the linear unit fixed effects model defined in equation (1) with Assumptions 2 and 3, 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).

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 LIN-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, Theorem 1 can be used to improve the credibility of the BA design. Recall that the BA design makes the assumption that there is no time trend (Assumption 6). That is, the outcome from the previous time period can be used to estimate the average counterfactual outcome in the next time period when the treatment status changes. In practice, however, this identification assumption may be questionable especially when estimating a long-term average treatment effect defined in equation (21). One possibile way to address this problem is to exploit the equivalence between matching and weighted least squares estimators and model a time trend using observations prior to the administration of treatment.

For example, researchers may use the following weighted least squares estimator of the average treatment effect in the F time periods ahead,

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i) - \beta(X_{it} - \overline{X}_i) - f_\gamma(t)\}^2 \tag{26}$$

where $f_\gamma(t)$ is a parametric time trend model (e.g., $f_\gamma(t) = \gamma_1(t - \bar{t}) + \gamma_2(t^2 - \bar{t^2})$) with $\bar{t}$ and $\bar{t^2}$ representing the average year and average squared year respectively,[11] and W_{it} is the weight computed according to Theorem 1 based on the matched set defined in equation (20),

$$W_{it} \;=\; D_{it} \sum_{i'=1}^{N} \sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{where} \quad w_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t) = (i', t' + F) \\ 1/|\mathcal{M}_{i't'}| & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

where $\mathcal{M}_{it}$ is defined in equation (20). Although the above model assumes a common time trend across units, we can also estimate a separate time trend for each unit if there is a sufficient number of time periods. This is done by replacing $f_\gamma(t)$ with $f_{\gamma_i}(t)$ (e.g., $f_{\gamma_i}(t) = \gamma_{i1}(t - \bar{t}) + \gamma_{i2}(t^2 - \bar{t^2})$). While this strategy enables flexible modeling of time trend, care must be taken especially for a large value of F since we are extrapolating into the future.

4 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 and comparing our results with various fixed effects models. We show that different causal assumptions can yield substantively different results, but our methodology using the before-and-after design generally suggests that joint GATT membership slightly increases bilateral trade volume on average.

4.1 Effects of GATT Membership on Bilateral Trade

Does GATT membership increase international trade? Rose (2004) finds that the answer to this question is negative. Based on the standard gravity model applied to dyadic trade data, his

[11] For an unbalanced time-series cross section data set, $\bar{t}$ and $\bar{t^2}$ will vary across units.

analysis yields 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 positive 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."

Despite the substantive controversy regarding the effects of GATT on trade, researchers appear to have reached a methodological consensus that the LIN-FE is the correct model to be used (Feenstra, 2003; Anderson and Wincoop, 2003). In another article, Rose (2005) expresses his belief in the usefulness of LIN-FE by stating "In terms of confidence, I follow the profession in placing most confidence in the fixed effects estimators; I have no clear ranking between country-specific and country pair-specific effects" (p. 10). Tomz *et al.* (2007) agree with this assessment and write, "We, too, prefer FE estimates over OLS on both theoretical and statistical grounds" (p. 2013). Below, we critically examine the assumptions of the LIN-FE in the context of this specific application.

4.2 Data

We analyze the data set from Tomz *et al.* (2007) which updates and corrects some minor errors in the data set used by Rose (2004). Unlike Rose (2004) and Tomz *et al.* (2007), however, we restrict our analysis to the period between 1948 and 1994 so that we focus on the effects of GATT and avoid conflating them with the effects of WTO. As shown below, this restriction does not significantly change the conclusions of the studies, 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.

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 two treatment variables: whether both countries in a dyad i are members (formal or participants) of GATT or not in a given year t (mix of dyads with one GATT member and no GATT member). This analysis focuses on the reciprocity hypothesis that GATT can impact bilateral trade only when countries mutually agree on reducing trade barriers (Bagwell and Staiger, 1999).[12]

Figure 5 shows the distributions of these two treatment variables across dyads (vertical axis) and over time (horizontal axis). For any dyad, the treatment status changes at most once in only one direction from the control condition to the treatment condition. This observation holds true for both formal membership (left plot) and participant status (right plot). Given this distribution of treatment variables, we next consider different identification strategies.

4.3 Models and Assumptions

We begin with the following linear regression model with dyadic fixed effects used by Rose (2007),

$$\log Y_{it} \quad = \quad \alpha_i + \beta X_{it} + \delta^\top \mathbf{Z}_{it} + \epsilon_{it} \tag{28}$$

where X_{it} is one of the binary treatment indicators for dyad i in year t described above, Y_{it} is the bilateral trade volume, 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 discussed earlier, the advantage of this standard dyad fixed effects model is its ability to adjust for unobserved time-invariant confounders.

Next, we progressively improve this LIN-FE. First, we relax the linearity assumption, which, as shown in Proposition 1, leads to bias when there exists heterogeneity across dyads in treat-

[12] We also conduct analyses based on alternative definitions of treatment, such as whether only one country or no country in a dyad is member, see Appendix A.4.

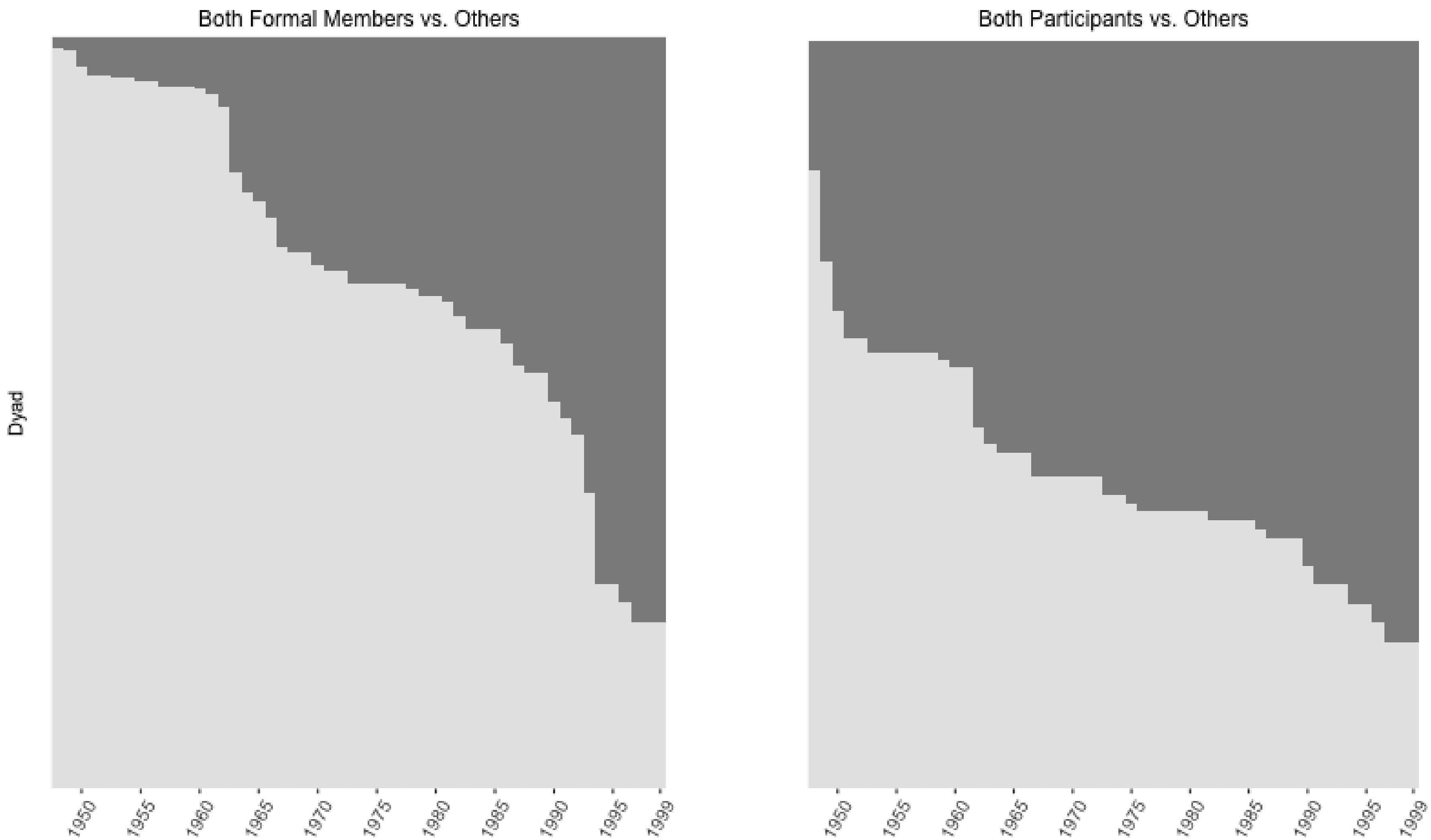


Figure 5: **Distributions of the Treatment Variables across Dyads and over Time.** This figure displays the distribution of treatment for 9,180 dyads from 1948 to 1999 based on 136 countries whose existence and membership status were identified by Tomz *et al.* (2007) during the entire period. In the left (right) panel, the binary treatment variable is colored in dark grey if both countries in the dyad are formal members (participants) and in light grey otherwise. In both cases, the visualization shows that the treatment status does not revert back and forth over time because countries do not exit the institution once they join.

ment effect and/or treatment assignment probability. Indeed, Subramanian and Wei (2007) find substantial heterogeneity in the effects of GATT/WTO on trade. We instead apply the proposed nonparametric matching estimator, given by equation (14), which compares each treated observation with the average outcome among *all* control observations within the same unit. In the current application, this implies the within-dyad comparison between the control observations in an earlier period and the treatment observations in a later period because the treatment status changes at most once for any given dyad. However, this comparison is potentially problematic since the data set spans for a relatively large number of years and hence some of the matched control observations may be too far apart in time from the treated observation to be comparable. For example, it would not be credible if one estimates the counter-factual bilateral trade volume in 1994 using the observed trade volume in 1950 since various other factors have changed between those two years.

Second, to improve the comparability of treated and control observations, we employ the first-difference (FD) estimator by restricting the matched set and compare the observations within only two subsequent time periods with treatment status change (see equation (19)). Under this special case of the before-and-after design, we require the assumption of no time trend because the control observation immediately before the change in treatment status is used to form an estimate of the counterfactual outcome under the treatment condition. While this identification strategy is more credible, the assumption of no time trend may be too stringent given that trade volume in general has increased over time. This problem is particularly pronounced if researchers are interested in estimating the long-term effects (see equation (21)) rather than the contemporaneous effect.

Thus, we generalize the first-difference estimator by including longer lagged control observations in the matched set prior to the change in the treatment status (see equation (20)). This allows us to parametrically adjust for the time trend in bilateral trade volume by exploiting the equivalence between matching and weighted fixed effects estimators. Under this general design, we can also estimate both contemporaneous and longer-term effects by setting various values of leads, $F \geq 0$. Specifically, we fit the following weighted fixed effects estimator with a quadratic time trend,

$$\hat{\beta}_{\mathsf{BA}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \{(Y_{it} - \overline{Y}_i) - \beta(X_{it} - \overline{X}_i) - \gamma_1(t - \bar{t}_i) - \gamma_2(t^2 - \bar{t}_i^2)\}^2 \qquad (29)$$

where W_{it} is given by equation (27), and $\bar{t}_i$ and $\bar{t}_i^2$ are the mean values of year and squared year variables, respectively (unlike the estimator in equation (26), these values may differ across dyads because some dyads do not cover the entire period in this application). In this application, we chose at most $F = 5$ as our maximum value of this lead variable since the identification of longer-term effects requires the extrapolation of time trends into the future based on the observed control observations from the past.

A word of caution is warranted about these estimators. As discussed in Section 2, these dyad linear fixed effects regression models and the before-and-after designs critically assume that there exist no time-varying unobserved confounders and past outcomes do not confound the causal relationship between current treatment and outcome.[13] These assumptions may be unrealistic. Studies have shown that economic interests and previous levels of engagement in bilateral trade affect countries' incentives to join the GATT/WTO (Davis and Wilf, 2017). That is, past outcomes (i.e., bilateral trade volumes) may affect current treatment (i.e., GATT membership). Furthermore, past treatments may also affect current outcome. Atkeson and Burstein (2010) find that changes in trade barriers, such as most favored nations (MFN) tariffs applied to GATT members, will determine forward-looking firms' reactions to exit, export, and product innovation, which will in turn affect future trade volumes. For the empirical analysis of this paper, however, we maintain these assumptions and focus on the aforementioned improvements of the linear fixed effects regression estimator used in the literature. In the concluding section, we briefly discuss potential extensions of the proposed methodology to address these fundamental identification assumptions

[13]Moreover, the model assumes no interference between units. That is, one dyad's treatment status does not affect the trade volumes of other dyads. Although this assumption may also be unrealistic in this interdependent world, relaxing it is difficult and is beyond the scope of this paper.

Membership	Dyad Fixed Effects			Before-and-After Design		
	Standard	Weighted	First Diff.	Lag = 3	Lag = 5	Lag = 7
Formal	−0.048	−0.069	0.075	−0.003	0.028	0.035
(N=196,207)	(0.025)	(0.023)	(0.054)	(0.028)	(0.022)	(0.020)
N (non-zero weights)	196,207	110,195	6,846	8,625	12,880	17,274
Participants	0.147	0.011	0.096	0.066	0.065	0.079
(N=196,207)	(0.031)	(0.029)	(0.030)	(0.035)	(0.028)	(0.026)
N (non-zero weights)	196,207	68,004	3,952	4,472	6,640	8,905
covariates	year-varying dyadic covariates			quadratic time trends		

Table 1: **Estimated Contemporaneous Effects of GATT on the Logarithm of Bilateral Trade based on Various Dyad Fixed Effects Estimators**. The "Standard" column presents the estimates based on the standard linear regression model with dyadic fixed effects given by equation (28). The "Weighted" column presents the estimates based on the nonparametric matching estimator given by equation (12). The "First Diff." column is based on the comparison between two subsequent time periods with treatment status change. Finally, the "Before-and-After Design" columns present the results based on three different lengths of lags with quadratic time trends (equation (29)). These estimators compare the dyads of two GATT members with those consisting of either one or no GATT member. "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 year-varying dyadic covariates include GSP (Generalized System of Preferences), log product real GDP, log product real GDP per capita, regional FTA (Free Trade Agreement), currency union, and currently colonized. Robust standard errors, allowing for the presence heteroskedasticity, are in parentheses. The results suggest that different causal assumptions, which imply different regression weights, can yield different results.

of unit fixed effects regression estimators.

4.4 Findings

We present the results based on each estimator discussed above. Table 1 summarizes the estimated contemporaneous effects of the GATT membership on bilateral trade volume. When we use the standard linear regression model with dyadic fixed effects (see the "Standard" column), we find that the formal membership does not increase trade volume on average (if anything, there appears to be a negative effect). In contrast, the estimated effect of participant is postive and statistically significant. As expected, these results are consistent with the orignal findings from both Rose

(2004) and Tomz *et al.* (2007) who also used standard fixed effects regression models.

However, the estimates based on the nonparametric matching estimator allowing for heterogeneity in treatment assignment and treatment effects suggest that the effect of participant membership is much smaller and statistically indistinguishable from zero (see the "Weighted" column). The nonparametric matching estimator uses a much fewer number of observations than the standard dyadic fixed effects because the dyads with no treatment status change are dropped. The standard dyadic fixed effects model still include these observations because of year-varying dyadic covariates.

In contrast, the analysis based on the comparison between two subsequent time periods with treatment status change, corresponding to the "First Diff." column, shows that the estimated effects for both treatment variables are positive, although the effect of formal membership is not statistically significant. The sample size for this analysis is further reduced because of its focus on the immediately before and after the change in treatment status. As a consequence, the standard error for the estimated effect of formal membership is substantially greater than those of standard and weighted analyses. On the other hand, the standard error for the estimated effect of participant does not increase as much, suggesting that the effect is relatively precisely estimated.

Furthermore, we implement the before-and-after design with three different lengths of lag in order to adjust for the time trends in trade volume for each dyad. Consistent with the above analyses, we find that the estimated contemporaneous effect of the formal membership is not distinguishable from zero with small point estimates, regardless of the lag length. However, while the standard error tends to be greater with a fewer number of lages, we consistently find positive effects of participant membership across models with various lags. In sum, our findings suggest that different causal assumptions can yield different results. Credible comparisons under the before-and-after design yield a robust finding that the contemporaneous effect of participant is positive while that of formal membership is not statistically distinguishable from zero.

Finally, we estimate the longer-term effects of the two treatment variables given in equa-

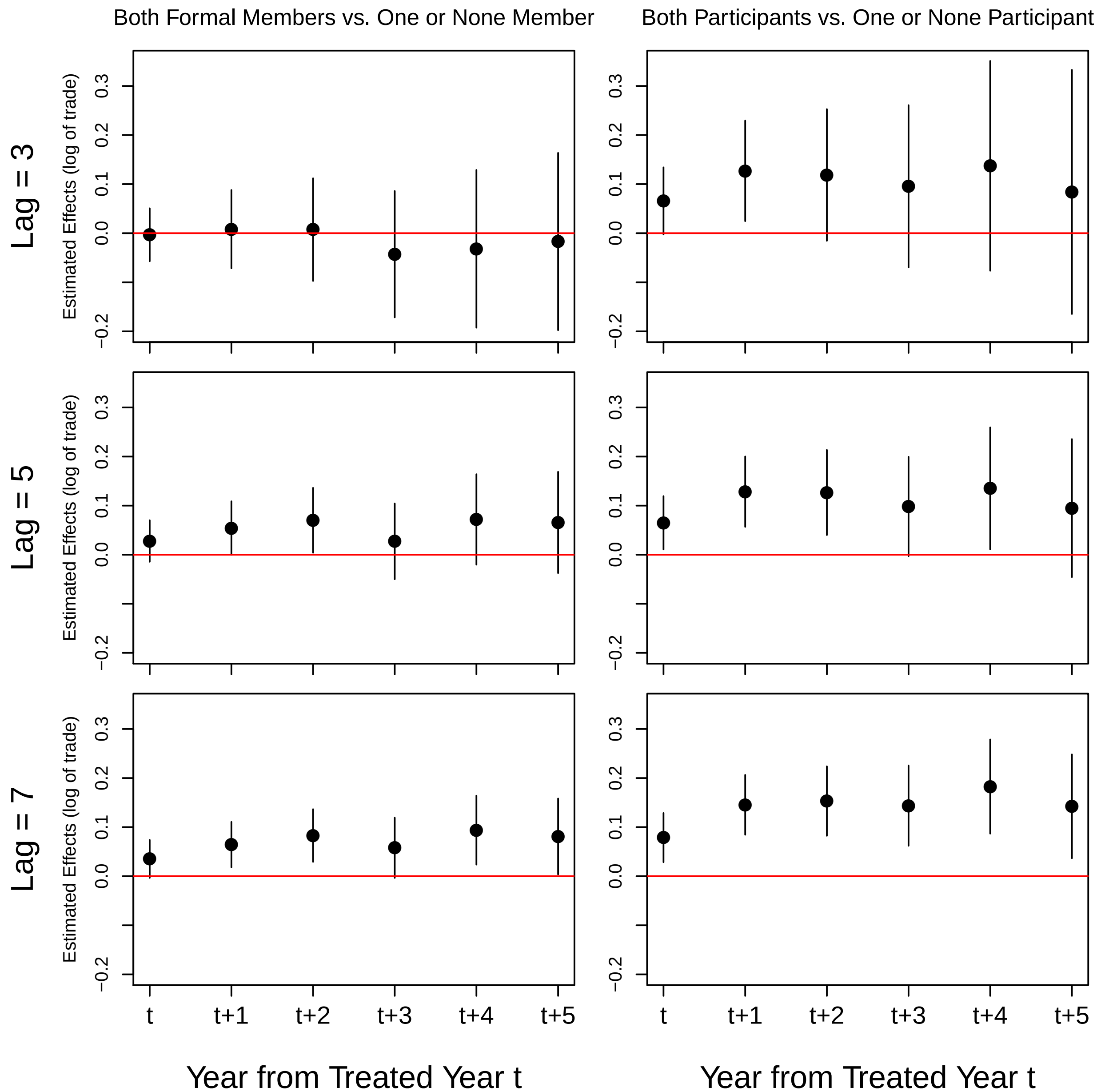


Figure 6: **Estimated Longer-term Effects of GATT on the Logarithm of Bilateral Trade based on Before-and-After Design**. This plot presents point estimates and 95% confidence intervals for the estimated effects of GATT membership on bilateral trade. The quantity of interest and the matched set are given by equations (21) and (20), respectively. That is, we compare the average bilateral trade volume across $L \in \{3, 5, 7\}$ years of lags before the treatment given at year t against the bilateral trade volume in $t + F$, where $F \in \{0, 1, 2, 3, 4, 5\}$ is the years since the treatment. We also include quadratic time trends. Overall, we find no evidence of positive effects of formal membership. We find some evidence of positive effects of participant membership with substantively much smaller effect size (e.g., 20% increase in 5 years) than the estimates from Tomz *et al.* (e.g., 71.6% contemporaneous effects). Robust standard errors allowing for the presence of heteroskedasticity are used.

tion (21). Figure 6 presents the point estimates and 95% confidence intervals for the estimated effects of formal membership (left column) and participant (right column) on trade volume from the year of treatment t to year $t+5$. We also use various lengths of lag to examine the robustness of our findings. The quadratic time trend is included in the model to account for time trend in trade volume.

In general, we find little evidence for longer-term effects of formal membership. Notwithstanding the statistically significant positive effects estimated based on 7 years of lags, the effect sizes are modest: the highest estimate suggests 8 percent ($\approx \exp(0.081) - 1$) increase in trade volume over 5 years after joining the GATT. Similarly, we find some evidence of positive effects of participant membership with the model with the lag of 7 years. However, the estimated effect sizes are substantively much smaller (e.g., maximum 20% increase in 5 years) than the estimates reported in Tomz *et al.* (e.g., 71.6% contemporaneous effect).

We note that the estimated effect sizes are stable across various implementations of the before-and-after design with different values of the lags, although a longer lag yields smaller standard errors due to a larger sample size. To account for the correlations across dyads when they share a common member, we verify our findings with the cluster-robust standard errors proposed by Aronow *et al.* (2015) and find that the positive effects of participant membership remain statistically significant although the standard errors are about 20% larger on average. Our findings are also robust to including/excluding time-varying covariates as well as linear time trend (see Appendix A.5 for additional results). This is not surprising since matching is expected to reduce model dependence (Ho *et al.*, 2007).

5 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 confounders 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 designs that are closely related to linear unit 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 unit 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 unit fixed effects regression estimators. For example, we extend the first difference estimator to the general case of the before-and-after design and show its equivalence to a weighted linear regression with unit 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.

In this paper, we show that causal inference with unit fixed effects regression models is fundamentally based on the within-unit comparison between the treated and control observations.

The major limitation of this identification strategy is that one must assume a certain within-unit time trend for the average potential outcome. We used past control observations to model this time trend, but it is also possible to use the observations from other similar units to estimate this trend. For the space limitation, we do not explore such an approach, which includes the difference-in-differences designs. We are currently extending our matching framework to this alternative identification strategy in a separate paper.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 1, 235–267.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association* **107**, 498, 833–843.

Anderson, J. E. and Wincoop, E. v. (2003). Gravity with gravitas: A solution to the border puzzle. *The American Economic Review* **93**, 1, 170–192.

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.

Aronow, P. M., Samii, C., and Assenova, V. A. (2015). Cluster–robust variance estimation for dyadic data. *Political Analysis* **23**, 4, 564–577.

Atkeson, A. and Burstein, A. T. (2010). Innovation, firm dynamics, and international trade. *Journal of political economy* **118**, 3, 433–484.

Bagwell, K. and Staiger, R. W. (1999). An economic theory of GATT. *The American Economic Review* **89**, 1, 215–248.

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.

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.

Davis, C. L. and Wilf, M. (2017). Joining the club: Accession to the gatt/wto. *The Journal of Politics* **79**, 3, 964–978.

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.

Feenstra, R. C. (2003). *Advanced International Trade*. Princeton University Press.

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.

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.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association* Forthcoming.

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

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. (2005). Does the WTO make trade more stable? *Open Economies Review* **16**, 7–22.

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 Mathematical Appendix

A.1 Equivalence between Assumptions 1 and 3

Consider the potential outcome model $Y_{it}(x) = g(x, \mathbf{U}_i, \epsilon_{it})$, which is consistent with equation (5). It is obvious that under this model equation (5) of Assumption 1 implies Assumption 3. To prove the converse, we focus on the case with $T = 3$ 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 Proof of Proposition 2

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

$$\hat{\tau}_{\mathsf{match}} = \frac{1}{\frac{1}{N}\sum_{i=1}^N C_i} \cdot \frac{1}{N} \sum_{i=1}^N C_i \left(\frac{\sum_{t=1}^T X_{it} Y_{it}}{\sum_{t=1}^T X_{it}} - \frac{\sum_{t=1}^T (1 - X_{it}) Y_{it}}{\sum_{t=1}^T (1 - X_{it})} \right) \tag{30}$$

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 3 implies the following conditional independence,

$$\{Y_{it}(1), Y_{it}(0)\} \perp\!\!\!\perp \mathbf{X}_i \mid \mathbf{U}_i \tag{31}$$

The law of iterated expectation implies,

$$\mathbb{E}(Y_{it}(x) \mid C_i = 1) = \mathbb{E}\{\mathbb{E}(Y_{it}(x) \mid \mathbf{U}_i, C_i = 1) \mid C_i = 1\} \tag{32}$$

for $x = 0, 1$. We show that the difference of means over time within unit i estimates the inner expectation of equation (32) without bias because it adjusts for $\mathbf{U}_i$. Consider the case with $x = 1$.

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

$$= \mathbb{E}(Y_{it}(1) \mid C_i = 1)\Pr(C_i = 1)$$

where the second equality follows from equation (31). 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 (30) 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. $\square$

A.3 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,

$$\begin{aligned}
&\sum_{i=1}^N \sum_{t=1}^T W_{it}(2X_{it} - 1)\alpha_i^* \\
= & \sum_{i'=1}^N \sum_{t'=1}^T \left(\sum_{i=1}^N \sum_{t=1}^T D_{it} w_{it}^{i't'} (2X_{it} - 1)\alpha_i^* \right) \\
= & \sum_{i'=1}^N \sum_{t'=1}^T \left(\sum_{i=1}^N \sum_{t=1}^T D_{it} \left\{ X_{i't'} w_{it}^{i't'} (2X_{it} - 1)\alpha_i^* + (1 - X_{i't'}) w_{it}^{i't'} (2X_{it} - 1)\alpha_i^* \right\} \right) \\
= & \sum_{i'=1}^N \sum_{t'=1}^T D_{it} \Bigg\{ X_{i't'} \left(\alpha_i^* - \sum_{(i,t) \in \mathcal{M}_{i't'}} \frac{1}{|\mathcal{M}_{i't'}|} (1 - X_{it})\alpha_i^* \right) \\
& \qquad + (1 - X_{i't'}) \left(\sum_{(i,t) \in \mathcal{M}_{i',t'}} \frac{1}{|\mathcal{M}_{i't'}|} X_{it}\alpha_i^* - \alpha_i^* \right) \Bigg\} \\
= & \sum_{i'=1}^N \sum_{t'=1}^T D_{it} \left\{ X_{i't'} \left(\alpha_i^* - \alpha_i^* \right) + (1 - X_{i't'}) \left(\alpha_i^* - \alpha_i^* \right) \right\} = 0
\end{aligned} \tag{33}$$

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,

$$\begin{aligned}
&\sum_{i=1}^N \sum_{t=1}^T W_{it} \\
= & \sum_{i=1}^N \sum_{t=1}^T D_{it} \left(\sum_{i'=1}^N \sum_{t'=1}^T w_{it}^{i't'} \right) \\
= & \sum_{i'=1}^N \sum_{t'=1}^T D_{it} \left(\sum_{i=1}^N \sum_{t=1}^T w_{it}^{i't'} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left[X_{i't'}\left(1+\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{|\mathcal{M}_{i't'}|}(1-X_{it})\right) + (1-X_{i't'})\left(1+\sum_{(i,t)\in\mathcal{M}_{i't'}}\frac{1}{|\mathcal{M}_{i't'}|}X_{it}\right)\right] \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{it}\left\{X_{i't'}(1+1) + (1-X_{i't'})(1+1)\right\} \;=\; 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it} \qquad (34)
\end{aligned}
$$

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,

$$
X_{it} - \overline{X}_i^* \;=\; \begin{cases} \frac{1}{2} & \text{if } X_{it} = 1 \\ -\frac{1}{2} & \text{if } X_{it} = 0 \end{cases} \qquad (35)
$$

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

$$
\begin{aligned}
\hat{\beta}_{\mathsf{WFE}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(X_{it}-\overline{X}_i^*)(Y_{it}-\overline{Y}_i^*)}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(X_{it}-\overline{X}_i^*)^2} \\
&= \frac{2}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(X_{it}-\frac{1}{2})(Y_{it}-\overline{Y}_i^*) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{W_{it}(2X_{it}-1)Y_{it} - W_{it}(2X_{it}-1)\overline{Y}_i^*\right\} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)Y_{it} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{D_{it}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right)(2X_{it}-1)Y_{it}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} \Bigg[X_{i't'} D_{it} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'} (2X_{it}-1) Y_{it} \right) \\
&\qquad\qquad + (1 - X_{i't'}) D_{it} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'} (2X_{it}-1) Y_{it} \right) \Bigg] \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} \Bigg[X_{i't'} \left(D_{it} Y_{it} - \sum_{(i,t)\in\mathcal{M}_{i't'}} \frac{D_{it}}{|\mathcal{M}_{i't'}|} (1 - X_{it}) Y_{it} \right) \\
&\qquad\qquad + (1 - X_{i't'}) \left(\sum_{(i,t)\in\mathcal{M}_{i't'}} \frac{D_{it}}{|\mathcal{M}_{i't'}|} |X_{it} - D_{it} Y_{it} \right) \Bigg] \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it} (\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}) = \hat{\tau}
\end{aligned}
$$

where the second equalities follows from equation (34) and (35), and the fourth equality from equation (33). The last equality follows from applying the definition of $\widehat{Y_{it}(1)}$ and $\widehat{Y_{it}(0)}$ given in equation (15) $\square$

A.4 Estimated Contemporaneous Effects of GATT Membership on Bilateral Trade with Alternative Membership Definitions

Comparison	Membership	Dyad Fixed Effects Standard	Dyad Fixed Effects Weighted	Dyad Fixed Effects First Diff.
	Formal	−0.034	−0.061	0.076
	(N=175,814)	(0.025)	(0.023)	(0.055)
	White's p-value		0.031	0.000
Both vs. One	N (non-zero weights)	175,814	100,055	6,712
	Participants	0.161	0.020	0.099
	(N=187,651)	(0.031)	(0.029)	(0.030)
	White's p-value		0.000	0.086
	N (non-zero weights)	187,651	64,152	3,900
	Formal	−0.011	−0.094	0.031
	(N=109,702)	(0.041)	(0.041)	(0.067)
	White's p-value		0.058	0.000
One vs. None	N (non-zero weights)	109,702	36,115	2,670
	Participants	0.181	−0.034	0.053
	(N=70,298)	(0.062)	(0.058)	(0.063)
	White's p-value		0.004	0.000
	N (non-zero weights)	70,298	15,766	1,087
covariates		year-varying covariates		

Table 2: **Estimated Contemporaneous Effects of GATT on the Logarithm of Bilateral Trade based on Alternative Membership Definitions**: The "Weighted" columns present the estimates based on the non-parametric matching estimator given by equation (12). "First Diff." is based on the comparison between two subsequent time periods with treatment status change. "Both vs. One" represents the comparison between dyads of two GATT members and those consisting of 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 covariates include GSP (Generalized System of Preferences), log product real GDP, log product real GDP per capita, regional FTA (Free Trade Agreement), currency union, and currently colonized. 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.

A.5 Before-and-After Design: Effects of GATT Membership on Trade

Before	After	Both vs. Mix Formal Membership				Both vs. Mix Participants			
Lag = 3	t	−0.003	−0.003	0.008	0.006	0.066	0.066	0.071	0.071
		(0.027)	(0.028)	(0.027)	(0.027)	(0.035)	(0.035)	(0.035)	(0.035)
	t+1	0.008	0.008	0.021	0.015	0.127	0.127	0.128	0.120
		(0.041)	(0.041)	(0.040)	(0.040)	(0.052)	(0.052)	(0.052)	(0.052)
	t+2	0.005	0.007	0.013	0.005	0.115	0.119	0.112	0.103
		(0.053)	(0.053)	(0.053)	(0.053)	(0.068)	(0.068)	(0.068)	(0.069)
	t+3	−0.049	−0.043	−0.036	−0.045	0.095	0.096	0.076	0.054
		(0.065)	(0.066)	(0.065)	(0.065)	(0.084)	(0.084)	(0.084)	(0.084)
	t+4	−0.040	−0.032	−0.021	−0.022	0.141	0.137	0.153	0.137
		(0.082)	(0.082)	(0.082)	(0.082)	(0.109)	(0.109)	(0.108)	(0.109)
	t+5	−0.043	−0.017	−0.013	−0.004	0.087	0.084	0.095	0.078
		(0.092)	(0.092)	(0.092)	(0.092)	(0.126)	(0.127)	(0.127)	(0.126)
Lag = 5	t	0.028	0.028	0.037	0.006	0.066	0.065	0.072	0.071
		(0.022)	(0.022)	(0.021)	(0.027)	(0.028)	(0.028)	(0.028)	(0.035)
	t+1	0.052	0.054	0.060	0.015	0.129	0.128	0.123	0.120
		(0.028)	(0.028)	(0.028)	(0.040)	(0.036)	(0.037)	(0.036)	(0.052)
	t+2	0.065	0.070	0.066	0.005	0.123	0.127	0.109	0.103
		(0.034)	(0.034)	(0.033)	(0.053)	(0.044)	(0.044)	(0.044)	(0.069)
	t+3	0.017	0.027	0.025	−0.045	0.096	0.098	0.076	0.054
		(0.039)	(0.039)	(0.039)	(0.065)	(0.051)	(0.052)	(0.051)	(0.084)
	t+4	0.057	0.072	0.079	−0.022	0.136	0.135	0.161	0.137
		(0.047)	(0.047)	(0.047)	(0.082)	(0.063)	(0.063)	(0.063)	(0.109)
	t+5	0.030	0.066	0.059	−0.004	0.097	0.095	0.118	0.078
		(0.052)	(0.053)	(0.052)	(0.092)	(0.071)	(0.072)	(0.071)	(0.126)
Lag = 7	t	0.033	0.035	0.049	0.006	0.079	0.079	0.088	0.071
		(0.020)	(0.020)	(0.020)	(0.027)	(0.026)	(0.026)	(0.025)	(0.035)
	t+1	0.059	0.064	0.077	0.015	0.145	0.145	0.145	0.120
		(0.024)	(0.024)	(0.023)	(0.040)	(0.031)	(0.031)	(0.031)	(0.052)
	t+2	0.073	0.083	0.085	0.005	0.148	0.153	0.142	0.103
		(0.027)	(0.027)	(0.027)	(0.053)	(0.036)	(0.036)	(0.036)	(0.069)
	t+3	0.042	0.058	0.063	−0.045	0.141	0.144	0.132	0.054
		(0.031)	(0.031)	(0.031)	(0.065)	(0.041)	(0.042)	(0.041)	(0.084)
	t+4	0.073	0.094	0.107	−0.022	0.182	0.183	0.212	0.137
		(0.036)	(0.036)	(0.036)	(0.082)	(0.049)	(0.049)	(0.048)	(0.109)
	t+5	0.037	0.081	0.077	−0.004	0.145	0.142	0.172	0.078
		(0.039)	(0.039)	(0.039)	(0.092)	(0.053)	(0.054)	(0.053)	(0.126)
linear time trend		✓	✓	✓	✓	✓	✓	✓	✓
quadratic time trend			✓		✓		✓		✓
year-varying covariates				✓	✓			✓	✓

Table 3: **Estimated Effects of GATT Membership on the Logarithm of Bilateral Trade based on Before-and-After Design**: This table summarizes the estimated effects of GATT membership on bilateral trade, i.e., the comparison between dyads of two GATT members and those consisting of either one GATT member. To implement the before-and-after design, we compare the average bilateral trade volume across lags $= L \in \{3, 5, 7\}$ years against the bilateral trade volume in leads $= F \in \{0, 1, 2, 3, 4, 5\}$ years since the treated year, which yield the quantity of interest given in equation (22). We find no evidence of short-term and long-term effects of formal membership. Although we find some evidence of long-term positive effects of participant membership with the model with Lag =7, the estimated effects are about 2.6 to 7.8 times smaller than the estimate (0.56) from (Tomz *et al.*, 2007, p.g., 2012) Our findings are robust to including liner/quadratic time trends and with/without year-varying covariates. Robust standard errors, allowing for the presence of heteroskedasticity are used.