Understanding and Improving Linear Fixed Effects Regression Models for Causal Inference*

Kosuke Imai† In Song Kim‡

First Draft: July 1, 2011
This Draft: December 13, 2011

Abstract

Linear fixed effects regression models are a primary workhorse for causal inference among applied researchers. And yet, it has been shown that even when the treatment is exogenous within each unit, the linear regression models with unit-specific fixed effects do not consistently estimate the average treatment effect in the presence of heterogenous treatment effects and treatment assignment probabilities across units. In this paper, we offer a simple solution. Specifically, we show that weighted fixed effects regression models consistently estimate the average treatment effect under various identification strategies such as propensity score weighting, first differencing, stratified randomization, post-treatment stratification, and difference-in-differences. We prove the results by establishing various finite sample equivalence relationships between fixed effects and matching estimators. At the basic level, the results suggest that these estimators do not fundamentally differ in their ability to cope with unobserved heterogeneity. More importantly, our analysis identifies the information implicitly used by fixed effects models to estimate counterfactual outcomes necessary for causal inference, highlighting the potential sources of their bias and inefficiency. In addition, the proposed framework offers simple, model-based standard errors for various matching estimators. Finally, we illustrate the proposed methodology by revisiting the controversy concerning the effects of the General Agreement on Tariffs and Trade (GATT) membership on international trade. Open-source software is available for fitting the proposed weighted linear fixed effects estimators.

Key Words: difference-in-differences, first differencing, matching, panel data, propensity score, stratified randomization

*A previous version of the paper was circulated under the title of "Equivalence between Fixed Effects and Matching Estimators for Causal Inference." The proposed methods 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). We thank Christina Davis for raising the questions this paper answers. Thanks also to Alberto Abadie, Josh Angrist, Scott Ashworth, Chris Berry, Jake Bowers, Kei Hirano, Shigeo Hirano, Gary King, Dustin Tingley, Teppei Yamamoto, and seminar participants at Inter-American Development Bank, MIT, and Princeton for helpful comments. Financial support from the National Science Foundation (SES-0918968) is acknowledged.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

‡Ph.D. candidate, Department of Politics, Princeton University, Princeton NJ 08544.Email: insong@princeton.edu

1 Introduction

Linear fixed effects regression models are a primary workhorse for causal inference in applied panel data analysis (e.g., Angrist and Pischke, 2009). Such models are also commonly used to analyze experiments under stratified randomization and cross-over designs (e.g., Duflo *et al.*, 2007; Jones and Kenward, 2003). In addition, linear fixed effects models may be used to estimate the average treatment effect in cross-sectional observational studies with post-treatment stratification where the exogeneity of treatment is assumed after stratifying on the observed values of pre-treatment confounders (e.g., Cochran, 1968; Rosenbaum and Rubin, 1984; Hansen, 2004; Iacus *et al.*, 2011). Yet, it has been shown that even if the treatment variable is assumed to be exogenous within each unit (or stratum), linear regression models with unit-specific (or stratum-specific) fixed effects cannot consistently estimate the average treatment effects in the presence of heterogeneous treatment effects and treatment assignment probabilities across units (e.g., Angrist, 1998; Wooldridge, 2005; Freedman, 2008; Humphreys, 2009; Chernozhukov *et al.*, 2011).

Specifically, for the sake of simplicity, consider a balanced panel data set of N units and T time periods where for each unit i at time t, we observe the outcome variable Y_{it} and the binary treatment indicator variable $X_{it} \in \{0, 1\}$. We emphasize that our theoretical results are applicable to unbalanced panel data as well as non-panel data settings including stratified randomized experiments and cross-sectional observational studies with post-treatment stratification. Define the linear one-way (unit specific) fixed effects estimators as follows,

$$(\hat{\alpha}_{FE}, \hat{\beta}_{FE}) \quad = \quad \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - X_{it}\,\beta)^2 \tag{1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N)$.

Now, suppose that the stable unit treatment value assumption holds across units and time periods (Rubin, 1990) and assume that the treatment is exogenous within each unit. Then, the fixed effects estimator is biased for the average treatment effect $\mathbb{E}(Y_{it}(1) - Y_{it}(0))$ where $Y_{it}(x)$ represents the potential outcome for unit i at time t under the treatment status $X_{it} = x$. Indeed, as shown by Chernozhukov *et al.* (2011, Theorem 2), the least squares estimate of β based on this model, $\hat{\beta}_{FE}$, converges in probability to the following weighted average of unit-specific average treatment effects where the weights are proportional to the variance of the treatment assignment variable,

$$\hat{\beta}_{FE} \quad \xrightarrow{p} \quad \frac{\mathbb{E}\{\Delta(X_i)\ \sigma_i^2\}}{\mathbb{E}(\sigma_i^2)} \tag{2}$$

where $\Delta(X_i) = \mathbb{E}(Y_{it}(1) - Y_{it}(0) \mid X_i)$ is the average treatment effect condition on $X_i = (X_{i1}, \ldots, X_{iT})$ and $\sigma_i^2 = \sum_{t=1}^{T}(X_{it} - \overline{X}_i)^2/T$. The result implies that unless either the treatment effect or treatment assignment probability is homogenous across units, the one-way fixed effects estimator does not consistently estimate the average treatment effect. The same result applies to non-panel data settings including stratified randomized experiments and observational studies with stratification on observables where the standard fixed effects estimator is biased in general even if the treatment is assumed to be exogenous within each stratum (see e.g., Angrist, 1998; Freedman, 2008; Humphreys, 2009).

In this paper, we offer a simple solution to this problem. Specifically, we show that weighted fixed effects regression models can consistently estimate the average treatment effect under various identification strategies. For example, the estimates based on the within-unit inverse propensity score weighting and first differencing can be obtained by fitting weighted one-way fixed effects regressions with different sets of regression weights. In addition, the weighted linear fixed effects model can also yield an unbiased estimate when analyzing either stratified randomized experiments or observational studies with post-treatment stratification where the treatment assignment is assumed to be exogenous within each stratum. Another example we present is the difference-in-differences estimator in a general case of multiple time periods and repeated treatments, which we show is algebraically equivalent to weighted regression estimator with two-way fixed effects. What is surprising about these results is that various estimators with different causal assumptions are reduced to weighted fixed effects regression estimators with different regression weights.

Our main result shows how to construct regression weights from different identification strategies for fixed effects regressions, thereby providing a straightforward way to improve the use of fixed effects regression models for causal inference. Our results differ from those of Hahn (2001) and Wooldridge (2005) who identify the conditions under which linear fixed effects regression models yield consistent estimates of the average treatment effect. Rather, we show that weighted fixed effects regression models can consistently estimate the average treatment effect under various identification assumptions. Despite this critical difference, we share the spirit of these authors who advocate that one should focus on causal parameters rather than structural parameters (see also Angrist, 2001).

Our results are also related to White (1980b) who propose the use of weighted linear least squares estimators to detect possible misspecification of the ordinary least squares model. In fact, the proposed weighted fixed effects regression models can be used as specification tests for

the standard linear fixed effects models. Finally, the equivalence between weighted fixed effects regression models and various matching estimators implies that we can easily calculate the model-based standard error for these matching estimators using the standard robust variance estimator. This offers an alternative standard error calculation that is simpler than what is available in the literature (Abadie and Imbens, 2006, 2011).

We prove these results by establishing various finite sample (i.e., algebraic) equivalence relationships between fixed effects and matching estimators. To illustrate the idea, consider the following simple cross-section linear regression model without fixed effects,

$$(\hat{\alpha}, \hat{\beta}) \quad = \quad \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N} (Y_i - \alpha - X_i\,\beta)^2.$$

It is well known that β is algebraically equivalent to the difference-in-means estimator for the average treatment effect. Now, this simple regression estimator $\hat{\beta}$ can also be written as,

$$\hat{\beta} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y_i(1)} - \widehat{Y_i(0)} \right)$$

where

$$\widehat{Y_i(1)} \;=\; \begin{cases} Y_i & \text{if } X_i = 1 \\ \frac{\sum_{i'=1}^{N} X_{i'} Y_{i'}}{\sum_{i'=1}^{N} X_{i'}} & \text{if } X_i = 0 \end{cases} \quad \text{and} \quad \widehat{Y_i(0)} \;=\; \begin{cases} \frac{\sum_{i'=1}^{N} (1-X_{i'}) Y_{i'}}{\sum_{i'=1}^{N} (1-X_{i'})} & \text{if } X_i = 1 \\ Y_i & \text{if } X_i = 0 \end{cases}$$

Following Heckman *et al.* (1998b), we call this alternative representation as the "matching" representation because it shows that each treated (control) observation is matched with the average of control (treated) observations. Others such as Hahn (1998) calls it as the "imputation" representation because for each observation the conditional expectation of its counterfactual outcome is imputed. The advantage of such representation is that it clarifies the set of observations that are used to estimate the counterfactual for each unit.

We apply this matching representation to linear fixed effects regression estimators and derive various finite sample equivalence relationships between fixed effects and matching estimators. At the basic level, the results of this paper imply that the fixed effects and matching estimators do not fundamentally differ in their ability to cope with endogeneity in observational studies. A commonly held belief that fixed effects estimators can adjust for unobservables but matching estimators cannot is incorrect. In Section 2, we first establish that the one-way fixed effects estimator is algebraically equivalent to an *adjusted* matching estimator where each observation is matched with the average of all the other observations of the same unit, which may include those

with the identical treatment status. These "mismatches" are subsequently adjusted by rescaling the resulting matching estimate to reduce the attenuation bias. As noted above, however, this adjustment is not sufficient to eliminate bias unless treatment effect and treatment assignment probability are homogeneous across units.

We then prove that a more natural matching estimator, which matches each observation only with other observations of opposite treatment status within the same unit, is algebraically equivalent to a weighted one-way fixed effects regression estimator where the weights are inversely proportional to the proportion of treated/control observations within each unit. Unlike the standard one-way fixed effects estimator, this estimator is unbiased for the average treatment effect under the assumption that the treatment assignment is exogenous within each unit. Our main theoretical result generalizes this finding and establishes the finite sample equivalence relationships between weighted fixed effects regression estimators and a broad class of matching estimators for various causal quantities of interest. We also show that weighted fixed effects regression estimators can accomodate various identification strategies including propensity score weighting and first differencing.

In Section 3, we extend these results to the linear two-way (i.e., unit and time) fixed effects estimator (Wallace and Hussain, 1969). Like the one-way case, we establish that the standard two-way fixed effects estimator is equivalent to an adjusted matching estimator where an adjustment is made to account for mismatches. What is different about the two-way fixed effects estimator, however, is that another adjustment is made to address the consequences of simultaneously including unit and time specific effects in the regression model. While it is in general impossible to eliminate mismatches using weighted two-way fixed effects regression, we prove that the difference-in-differences estimator in a general setting of multiple time periods and repeated treatments is algebraically equivalent to weighted two-way fixed effects estimator. This suggests that while the standard two-way fixed effects regression is difficult to justify from a causal inference point of view, the weighted two-way fixed effects estimator avoids mismatches and compares treated units with control units (and vice versa) based on the difference-in-differences design.

In Section 4, we illustrate the proposed matching framework of fixed effects regression models by revisiting the controversy about whether GATT membership increases international trade. We show that in some cases the proposed weighted fixed effects regression estimator gives substantively different results from those based on the standard fixed effects model. This finding underscores the importance of causal assumptions regarding the information used to estimate counterfactual

outcomes when using fixed effects regression models. Finally, in Section 5, we briefly discuss the implications of the findings of this paper for applied panel data analysis. The open-source software, `wfe: Weighted Linear Fixed Effects Estimators for Causal Inference`, for fitting the proposed weighted fixed effects estimators is available through the Comprehensive R Archive Network (http://cran.r-project.org/package=wfe).

2 The One-Way Fixed Effects Estimator

2.1 Adjusted and Unadjusted Within-unit Matching Estimators

Recall that $\hat{\alpha}_{FE}$ and $\hat{\beta}_{FE}$ denote the least squares estimate of $\alpha = (\alpha_1, \ldots, \alpha_N)$ and β from the linear one-way fixed effects regression model (equation (1)). The following proposition establishes that this one-way fixed effects estimator can be written as a particular *adjusted* within-unit matching estimator.

Proposition 1 (Finite Sample Equivalence between the One-way Fixed Effects and Adjusted Within-unit Matching Estimators) *The following algebraic equality holds,*

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

where for $x = 0, 1$,

$$\widehat{Y_{it}(x)} \;=\; \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{T-1}\sum_{t' \neq t} Y_{it'} & \text{if } X_{it} = 1 - x \end{cases}$$

$$K \;=\; \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\cdot\frac{1}{T-1}\sum_{t'\neq t}(1 - X_{it'}) + (1 - X_{it})\cdot\frac{1}{T-1}\sum_{t'\neq t} X_{it'}\right\}.$$

Proof is in Appendix A.1.

The proposition provides a clear interpretation of the fixed effects estimator as an adjusted within-unit matching estimator. Specifically, the observation for unit i at time t is matched with the average of all the other observations for the same unit, and then the estimated treatment effect for this observation is calculated by computing the difference between the observed outcome and the average outcome of the matched units. In this matching procedure, the matched-set of a given observation contains other observations of the same treatment status unless there are only two time periods and the treatment is given to each unit exactly once. This means that, for example, a treated observation may be matched with a set of observations, some of which are treated. When

such "mismatches" occur, the treatment effect estimate for that observation may be attenuated. Proposition 1 shows that the one-way fixed effects estimator adjusts for this attenuation bias by dividing the average of observation-specific treatment effect estimates by the average proportion of properly matched units across all observations, i.e., K. Thus, a greater number of mismatches leads to a smaller value of K, which results in larger adjustment.

In general, this adjustment is not sufficient for yielding a consistent estimate of the average treatment effect even when the treatment assignment is assumed to be randomized within each unit. Given this problem, we consider the following more natural within-unit matching estimator which is unbiased for the average treatment effect so long as the treatment assignment is exogeneous within each unit,

$$\hat{\beta}^M \;=\; \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right) \tag{3}$$

where

$$\widehat{Y_{it}(1)} \;=\; \begin{cases} Y_{it} & \text{if } X_{it}=1 \\ \frac{\sum_{t'=1}^{T} X_{it'}Y_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it}=0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}(0)} \;=\; \begin{cases} \frac{\sum_{t'=1}^{T}(1-X_{it'})Y_{it'}}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if } X_{it}=1 \\ Y_{it} & \text{if } X_{it}=0 \end{cases} \tag{4}$$

This estimator has no mismatch because the treated (control) observation is compared with the average of control (treated) observations within each unit. This set of matched observations is given by,

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

where (i,t) denotes the observation for unit i at time t. We can relate this matching estimator with the adjusted within-unit matching estimator given in Proposition 1 by noting that this matching estimator does not require any adjustment due to mismatches.

The next proposition establishes that this within-unit matching estimator is algebraically equivalent to a weighted one-way fixed effects estimator where the weight is inversely proportional to the proportion of treated/control observations within each unit.

Proposition 2 (Finite Sample Equivalence between the Weighted One-Way Fixed Effects and Within-unit Matching Estimators) *Assume that the treatment varies within each unit, i.e., $0 < \sum_{t=1}^{T} X_{it} < T$ for all i. Then, the following algebraic equality always holds,*

$$(\hat{\alpha}^M, \hat{\beta}^M) \;=\; \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - X_{it}\,\beta)^2$$

where $\hat{\alpha}_i^M = \frac{1}{2}\left(\frac{1}{T}\sum_{t=1}^{T} W_{it}Y_{it} - \hat{\beta}^M\right)$, $\hat{\beta}^M$ *is given in equation* (3), *and*

$$W_{it} \equiv \begin{cases} \frac{T}{\sum_{t'=1}^{T} X_{it'}} & \text{if} \quad X_{it} = 1, \\ \frac{T}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if} \quad X_{it} = 0. \end{cases}$$

Proof is omitted since this proposition is a special case of a more general result given below as Theorem 1. The proposition shows that for a treated (control) observation, the weight is inversely proportional to the proportion of treated (control) observations for that unit. Proposition 2 also implies that the standard (i.e., unweighted) one-way fixed effects estimator is algebraically equivalent to the matching estimator given in equations (3) and (4) if and only if for every unit the number of treatment time periods is the same as that of control time periods. While such a situation is unlikely in observational studies, in cross-over experiments researchers can balance treatment assignment both within each unit and across units so that the unweighted one-way fixed effects estimator is algebraically equivalent to the unadjusted matching estimator.

2.2 Stratification in Randomized Experiments and Observational Studies

The results given above are also relevant for matching and stratification in randomized experiments and observational studies. For example, under the stratified design, experimental units are grouped into several strata based on the similarity of observed pre-treatment covariates. Then, the complete randomization of the treatment assignment is conducted within each strata. This design yields a more efficient estimate than the complete randomization without stratification (see Appendix A Imai *et al.*, 2008). Under this stratified design, the use of one-way fixed effects regression with strata specific effects is quite common among applied researchers (see e.g., Duflo *et al.*, 2007). Similarly, one-way fixed effecs regression model is applicable to observational studies with post-treatment stratification (or subclassification) where strata are constructed based upon the similarity of observed pre-treatment covariates and the exogeneity of treatment assignment is assumed within each stratum (e.g., Cochran, 1968; Rosenbaum and Rubin, 1984; Hansen, 2004; Iacus *et al.*, 2011).

Unfortunately, as discussed in Section 1, it has been shown that such approach does not yield an unbiased estimate of the average treatment effect if the treatment assignment probabilities and average treatment effects differ across strata (e.g., Angrist, 1998; Freedman, 2008; Humphreys, 2009). In fact, the linear one-way fixed effects estimator yields the weighted average of average treatment effects across strata where the weights are proportional to the variance of treatment assignment variable within each stratum. Proposition 2 shows that in these situations the weighted

linear one-way fixed effects regression model with stratum specific effects offers a simple way to estimate the average treatment effect without bias. We also emphasize that our main result is more general than Proposition 2 and can incorporate strata of different sizes.

2.3 Adjusting for Additional Covariates

When there exists a vector of additional covariates Z_{it} (e.g., time-varying confounders in panel data settings), we can follow the suggestion of Heckman *et al.* (1997, 1998a) and use the regression-adjusted matching procedure. Specifically, we assume $\mathbb{E}(Y_{it}(0) \mid Z_{it} = z) = g(z)$ and estimate this regression function using the control group observations alone, $g(z) = \mathbb{E}(Y_{it} \mid X_{it} = 0, Z_{it} = z)$. Once the regression function is estimated, then one can obtain the fitted value for each observation based on its covariate value Z_{it} and subtract them from the observed outcome variable, i.e., $Y_{it} - \hat{g}(Z_{it})$. Finally, this regression-adjusted outcome variable can be used to perform a matching procedure via the weighted one-way fixed effects regression.

Alternatively, we may consider the following weighted one-way fixed effects estimators with these confounders,

$$(\hat{\alpha}_{WFE}, \hat{\beta}_{WFE}, \hat{\delta}_{WFE}) \;=\; \underset{(\alpha,\beta,\delta)}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - X_{it}\,\beta - Z_{it}^{\top}\delta)^2, \tag{6}$$

where the expression for W_{it} is given in Proposition 2. This model can be justified *ex post* because it is algebraically equivalent to the matching estimator for the covariate adjusted outcome, i.e., $Y_{it} - Z_{it}^{\top}\hat{\delta}_{WFE}$.

Another way to adjust for time-varying confounders is to use propensity score weighting. Such an approach may be appealing because as shown by Hirano, Imbens, and Ridder (2003) weighting by the inverse of a nonparametric estimate of the propensity score leads to a semiparametrically efficient estimate of the average treatment effect (see also Hahn, 1998). Suppose that we estimate the propensity score as a function of these confounders, $\pi(Z_{it}) = \Pr(X_{it} = 1 \mid Z_{it})$, for each observation. Then, an efficient estimator of the average treatment effect based on the within-unit propensity score weighting with normalized weights (though this normalization is not required for our result as mentioned below) is given by,

$$\hat{\beta}^{W} \;=\; \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{t=1}^{T}\frac{X_{it}Y_{it}}{\hat{\pi}(Z_{it})}\Big/\sum_{t=1}^{T}\frac{X_{it}}{\hat{\pi}(Z_{it})} - \sum_{t=1}^{T}\frac{(1-X_{it})Y_{it}}{1-\hat{\pi}(Z_{it})}\Big/\sum_{t=1}^{T}\frac{(1-X_{it})}{1-\hat{\pi}(Z_{it})}\right\} \tag{7}$$

where $\hat{\pi}(Z_{it})$ is the estimated propensity score. An intuitive interpretation of this estimator is that for each unit the average treatment effect is estimated using the inverse-propensity score weighting

and then this estimate is averaged across units. The next proposition shows that this estimator is also algebraically equivalent to a weighted one-way fixed effects estimator with the transformed outcome variable.

PROPOSITION 3 (FINITE SAMPLE EQUIVALENCE BETWEEN THE TRANSFORMED WEIGHTED ONE-WAY FIXED EFFECTS AND WITHIN-UNIT PROPENSITY SCORE WEIGHTING ESTIMATORS) *Assume that the treatment varies within each unit, i.e.,* $0 < \sum_{t=1}^{T} X_{it} < T$ *for all* i*. Let* $\hat{\pi}(Z_{it})$ *be the estimated propensity score where the propensity score is defined as* $\pi(Z_{it}) \equiv \Pr(X_{it} = 1 \mid Z_{it})$ *and is assumed to be bounded away from* 0 *and* 1*. Then, the following algebraic equality always holds,*

$$(\hat{\alpha}^W, \hat{\beta}^W) \quad = \quad \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (Y_{it}^* - \alpha_i - X_{it}\,\beta)^2$$

where $\hat{\alpha}_i^W = \frac{1}{2}\left(\frac{1}{T}\sum_{t=1}^{T} W_{it} Y_{it}^* - \hat{\beta}^W\right)$*,* $\hat{\beta}^W$ *is given in equation* (7)*,* W_{it} *is defined in Proposition 2, and the transformed outcome* Y_{it}^* *is given by,*

$$Y_{it}^* \quad = \quad \begin{cases} \frac{\left(\sum_{t'=1}^{T} X_{it'}\right) Y_{it}}{\hat{\pi}(Z_{it})} \Big/ \sum_{t'=1}^{T} \frac{X_{it'}}{\hat{\pi}(Z_{it'})} & \text{if} \quad X_{it} = 1 \\ \frac{\left\{\sum_{t'=1}^{T} (1 - X_{it'})\right\} Y_{it}}{1 - \hat{\pi}(Z_{it})} \Big/ \sum_{t'=1}^{T} \frac{(1 - X_{it'})}{1 - \hat{\pi}(Z_{it'})} & \text{if} \quad X_{it} = 0 \end{cases}$$

Proof is given in Appendix A.2. The proposition can also be modified to accomodate different weighting schemes. For example, the propensity score weighting estimator without normalization, i.e., $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\{X_{it}Y_{it}/\hat{\pi}(Z_{it}) - (1-X_{it})Y_{it}/(1-\hat{\pi}(Z_{it}))\}$, can be computed by fitting the weighted one-way fixed effects regression with the transformed outcome, $Y_{it}^* = \sum_{t'=1}^{T} X_{it'} Y_{it}/\hat{\pi}(Z_{it})$ if $X_{it} = 1$ and $Y_{it}^* = \{\sum_{t'=1}^{T}(1 - X_{it'})\} Y_{it}/(1 - \hat{\pi}(Z_{it}))$ if $X_{it} = 0$.

2.4 The Main Result

We now prove the main result of this paper. The following theorem states that there exists an algebraically equivalent weighted one-way fixed effects estimator for *any* matching estimator that uses a weighted average of other observations from the same unit but with the opposite treatment status to estimate counterfactual outcome.

THEOREM 1 (GENERAL FINITE SAMPLE EQUIVALENCE BETWEEN THE WEIGHTED ONE-WAY FIXED EFFECTS AND MATCHING ESTIMATORS) *Assume that the treatment varies within each unit, i.e.,* $0 < \sum_{t=1}^{T} X_{it} < T$ *for all* i*. Define a general class of matching estimators as follows,*

$$\tilde{\beta}^M \quad = \quad \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} C_{it} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where $0 \leq C_{it} < \infty$, $\sum_{t=1}^{T}\sum_{i=1}^{N} C_{it} > 0$,

$$\widehat{Y_{it}(1)} = \begin{cases} Y_{it} & \text{if } X_{it}=1 \\ \sum_{t'=1}^{T} v_{it}^{it'} X_{it'} Y_{it'} & \text{if } X_{it}=0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}(0)} = \begin{cases} \sum_{t'=1}^{T} v_{it}^{it'}(1-X_{it'})Y_{it'} & \text{if } X_{it}=1 \\ Y_{it} & \text{if } X_{it}=0 \end{cases}$$

and $v_{it}^{it'}$ *is the normalized pre-defined non-negative weight for observation* (i,t') *which is matched with observation* (i,t) *such that* $\sum_{t'=1}^{T} v_{it}^{it'} X_{it'} = \sum_{t'=1}^{T} v_{it}^{it'}(1-X_{it'}) = 1$. *Then, the following algebraic equality holds,*

$$(\tilde{\alpha}^M, \tilde{\beta}^M) = \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - X_{it}\beta)^2$$

where

$$\tilde{\alpha}_i^M = \frac{1}{2}\left(\frac{1}{T}\sum_{t=1}^{T} W_{it}Y_{it} - \tilde{\beta}^M\right)$$

$$W_{it} = \sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{and} \quad w_{it}^{i't'} = \begin{cases} C_{it} & \text{if } (i,t)=(i',t') \\ v_{it}^{it'} C_{i't'} & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ 0 & \text{otherwise.} \end{cases}$$

Proof is given in Appendix A.3. Theorem 1 is applicable to an unbalanced panel. If the observation for unit i at time period t is missing, we set $C_{it} = 0$ and $v_{i't'}^{it} = 0$ for all (i',t').

The theorem shows how to derive the regression weights for the weighted one-way fixed effects regression estimator that is equivalent to a general class of matching estimators. For example, Proposition 2 is a special case of this theorem where for any observation (i,t') that is matched with observation (i,t), we have,

$$v_{it}^{it'} = \begin{cases} \frac{1}{\sum_{t^*=1}^{T}(1-X_{it^*})} & \text{if } X_{it}=1 \\ \frac{1}{\sum_{t^*=1}^{T} X_{it^*}} & \text{if } X_{it}=0 \end{cases} \tag{8}$$

and $C_{it} = 1$ for all observations. Since Theorem 1 allows for arbitrary matching weights $v_{it}^{it'}$, it opens up the possibility of implementing various matching methods within the framework of the weighted one-way fixed effects regression model. The theorem also allows researchers to define various causal quantities of interest by specifying non-negative values for C_{it}. For example, setting $C_{it} = X_{it}$ yields the matching estimator of the average treatment effect for the treated, and using the survey weights for C_{it} enables the estimation of the average treatment effect for the target population.

In addition, the theorem shows that our results are applicable to unbalanced panel data, experimental data with stratified randomization, post-treatment stratified data in observational studies

(see Section 2.2). For example, for unbalanced panel data, one can set $C_{it} = 0$ and $v_{i't'}^{it} = 0$ for all (i', t') if the data for unit i at time t are missing. In sum, Theorem 1 shows that a broad class of matching estimators for various causal quantities of interest is algebraically equivalent to a particular weighted one-way fixed effects estimator.

2.5 First Differencing

To further illustrate the applicability of Theorem 1, consider the following first difference model derived from the one-way fixed effects model,

$$\Delta Y_{it} \quad = \quad \Delta X_{it}\,\beta + \Delta\epsilon_{it} \tag{9}$$

for $t = 2, \ldots, T$ where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\Delta X_{it} = X_{it} - X_{i,t-1}$, and $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{i,t-1}$. Let β denote the least squares estimate of β,

$$\hat{\beta}^{FD} \quad \equiv \quad \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=2}^{T} (\Delta Y_{it} - \Delta X_{it}\,\beta)^2. \tag{10}$$

We show that this first difference estimator is algebraically equivalent to a matching estimator where each observation is matched with the observation of the same unit from the previous time period. If the treatment status changes from one time period to another, then the change in the outcome is used as an estimate of the treatment effect for this observation. If the treatment status remains identical between the two periods, then this observation does not contribute to the estimation of the average treatment effect. Then, because of this equivalence between the first difference and matching estimators, Theorem 1 implies that the first difference estimator is also equivalent to a weighted one-way fixed effects estimator. This result is formalized as the following proposition.

Proposition 4 (Finite Sample Equivalence between the First Difference, Matching, and Weighted One-Way Fixed Effects Estimators) *Consider the following matching estimator,*

$$\ddot{\beta}^{M} \quad = \quad \frac{1}{\sum_{i=1}^{N} \sum_{t=2}^{T} D_{it}} \sum_{i=1}^{N} \sum_{t=2}^{T} D_{it} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)} \right)$$

where $D_{it} = |X_{it} - X_{i,t-1}|$ *and*

$$\widehat{Y_{it}(1)} \quad = \quad \begin{cases} Y_{it} & \text{if } X_{it} = 1 \\ Y_{i,t-1} & \text{if } X_{it} = 0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}(0)} \quad = \quad \begin{cases} Y_{i,t-1} & \text{if } X_{it} = 1 \\ Y_{it} & \text{if } X_{it} = 0 \end{cases}$$

Then, the first difference estimator is algebraically equivalent to this matching estimator, i.e., $\hat{\beta}^{FD} = \ddot{\beta}^{M}$. *The estimator is also algebraically equivalent to the weighted fixed effects estimator*

given in Theorem 1 where $C_{it} = D_{it}$ and when $C_{it} = 1$, we have $v_{it}^{it'} = 1$ for $t' = t-1$ and $v_{it}^{it'} = 0$ otherwise.

Proof is given in Appendix A.4. In this proposition, for the observations with $C_{it} = 0$, the definition of $v_{it}^{it'}$ is not given because they do not contribute to the regression weights.

2.6 Efficient Computation of the Weighted One-Way Fixed Effects Estimator

When N is large, the computation of equation (6) can become demanding. Here, we show that the weighted least squares estimates, $(\hat{\beta}_{WFE}, \hat{\delta}_{WFE})$, can be computed by first subtracting its within-unit weighted average from each of the variables and then running the weighted regression using these "weighted demeaned" variables. This computation strategy is useful when the number of fixed effects is large because it avoids the inversion of a large matrix. To see why this procedure works, define the following vector and matricies,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}, \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} X_1 & \mathbf{Z}_1 \\ X_2 & \mathbf{Z}_2 \\ \vdots & \vdots \\ X_N & \mathbf{Z}_N \end{bmatrix},$$

where Y is an NT-dimensional column vector of outcome variable with $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$, $\mathbf{U}$ is an $(NT \times N)$ matrix of unit indicator variables with $\mathbf{0}$ and $\mathbf{1}$ being T-dimensional column vectors of zeros and ones, $\mathbf{V}$ is an $(NT \times (M+1))$ matrix of covariates with X_1 and Z_1 with $X_i = (X_{i1}, X_{i2}, \ldots, X_{iT})^\top$ and $\mathbf{Z}_i = [Z_{i1}\ Z_{i2}\ \ldots\ Z_{iT}]^\top$ being a T-dimensional vector of the treatment indicator variable and a $(T \times M)$ matrix of confounders, respectively. Finally, $\mathbf{W} = \text{diag}(W_1^\top\ W_2^\top\ \ldots\ W_N^\top)$ is the diagonal weight matrix with $W_i = (W_{i1}, W_{i2}, \ldots W_{iT})$.

This weighted one-way fixed effects regression is equivalent to the (unweighted) regression of $Y^* = W^{1/2}Y$ on $\mathbf{U}^* = W^{1/2}\mathbf{U}$ and $\mathbf{V}^* = W^{1/2}\mathbf{V}$. Thus, we use the well-known partitioned regression formula. First, we obtain the residuals from the regression of each column of $\mathbf{V}^*$ on $\mathbf{U}^*$, i.e., $\widetilde{X}_{it} = \sqrt{W_{it}}\left(X_{it} - \sum_{t'=1}^{T} \frac{W_{it'}X_{it'}}{\sum_{t^*=1}^{T} W_{it^*}}\right)$ and $\widetilde{Z}_{it} = \sqrt{W_{it}}\left(Z_{it} - \sum_{t'=1}^{T} \frac{W_{it'}Z_{it'}}{\sum_{t^*=1}^{T} W_{it^*}}\right)$. Similarly, the regression of Y^* on $\mathbf{U}^*$ yields the following residuals, i.e., $\widetilde{Y}_{it} = \sqrt{W_{it}}\left(Y_{it} - \sum_{t'=1}^{T} \frac{W_{it'}Y_{it'}}{\sum_{t^*=1}^{T} W_{it^*}}\right)$. Finally, the regression of the latter residuals on the other residuals yields the least squares estimates (β, δ) of equation (6). Thus, successive applications of weighted least squares allow one to efficiently fit weighted one-way fixed effects models.

2.7 Calculation of Standard Error and Specification Test

The proposed formulation of various matching estimators as weighted fixed effects regression estimators implies the model-based standard error calculation that is much simpler than the existing standard error calculation for matching estimators (Abadie and Imbens, 2006, 2011). In particular, the well-known Huber-White sandwich formula for robust standard errors can be directly applied (Huber, 1967; White, 1980a). Consider the weighted one-way fixed effects model given in equation (6). Let $\widetilde{V}_i$ be an $(T \times (M+1))$ matrix of "weighted demeaned" explanatory variables V_i, i.e., $\widetilde{V}_{it} = [\widetilde{X}_{it} \ \widetilde{Z}_{it}]$ where $\widetilde{X}_{it}$ and $\widetilde{Z}_{it}$ are defined above. Then, the asymptotic variance for the weighted fixed effects estimators $\hat{\theta}_{WFE} = (\hat{\beta}_{WFE}, \hat{\delta}_{WFE})$ is given by,

$$\widehat{\Psi}_{WFE} \quad = \quad \left(\frac{1}{N} \sum_{i=1}^{N} \widetilde{V}_i^\top \widetilde{V}_i \right)^{-1} \widehat{\Omega}_{WFE} \left(\frac{1}{N} \sum_{i=1}^{N} \widetilde{V}_i^\top \widetilde{V}_i \right)^{-1} \tag{11}$$

where the expression for $\widehat{\Omega}_{WFE}$ depends on the assumption one invokes.

For example, in the case of stratified randomized experiments or observational studies with stratification described in Section 2.2, we may assume the independence across observations but account for heteroskedasticity, which results in the following expression for $\widehat{\Omega}$,

$$\widehat{\Omega}_{WFE} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \widetilde{V}_i^\top \, \mathrm{diag}(\tilde{u}_{i1}^2 \ \ldots \ \tilde{u}_{iT}^2) \, \widetilde{V}_i \tag{12}$$

where $\tilde{u}_{it} = \sqrt{W_{it}}(\widetilde{Y}_{it} - \widetilde{X}_{it}\hat{\beta}_{WFE} - \widetilde{Z}_{it}^\top \hat{\delta}_{WFE})$. Similarly, in the panel data, we may allow for an arbitrary autocorrelation as well as heteroskedasticity,

$$\widehat{\Omega}_{WFE} \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \widetilde{V}_i^\top \, \tilde{u}_i \tilde{u}_i^\top \, \widetilde{V}_i \tag{13}$$

where $\tilde{u}_i = (\tilde{u}_{i1}, \ldots, \tilde{u}_{iT})^\top$.

Finally, our framework also implies a specification test to determine whether one should use the proposed weighted fixed effects model or stick with the standard fixed effects model. In particular, we follow White (1980b) who shows that the difference between the ordinary least squares and weighted least squares estimates can be used as a specification test. Under the null hypothesis that the assumptions of ordinary least squares are correct, the weighted least squares estimates should asymptotically converge to those of the ordinary least squares. In contrast, when misspecified (i.e., the linearity assumption does not hold), the ordinary least squares estimates asymptotically converge to the weighted least squares estimates with unknown weights that minimize the mean

squared prediction error. Thus, our test statistic and its asymptotic distribution are given by,

$$N(\hat{\theta}_{FE} - \hat{\theta}_{WFE})^\top \, \widehat{\Phi}^{-1} \, (\hat{\theta}_{FE} - \hat{\theta}_{WFE}) \quad \overset{A}{\sim} \quad \chi^2_{M+1} \tag{14}$$

where $\hat{\theta}_{FE} = (\hat{\beta}_{FE}, \hat{\delta}_{FE})$ and $\hat{\theta}_{WFE} = (\hat{\beta}_{WFE}, \hat{\delta}_{WFE})$ are the standard and weighted fixed effects estimators of $\theta = (\beta, \delta)$, respectively. If we account for both autocorrelation and heteroskedasticity, for example, the variance term is given by,

$$\begin{aligned}
\widehat{\Phi} &= \widehat{\Psi}_{WFE} + \widehat{\Psi}_{FE} - \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{V}_i^\top \widehat{V}_i\right)^{-1} \widehat{\Lambda} \left(\frac{1}{N}\sum_{i=1}^{N} \widetilde{V}_i^\top \widetilde{V}_i\right)^{-1} - \left(\frac{1}{N}\sum_{i=1}^{N} \widetilde{V}_i^\top \widetilde{V}_i\right)^{-1} \widehat{\Lambda} \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{V}_i^\top \widehat{V}_i\right)^{-1} \\
\widehat{\Psi}_{FE} &= \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{V}_i^\top \widehat{V}_i\right)^{-1} \widehat{\Omega}_{FE} \left(\frac{1}{N}\sum_{i=1}^{N} \widehat{V}_i^\top \widehat{V}_i\right)^{-1} \\
\widehat{\Lambda} &= \frac{1}{N}\sum_{i=1}^{N} \widetilde{V}_i^\top \, \tilde{u}_i \hat{u}_i^\top \, \widehat{V}_i
\end{aligned}$$

where $\widehat{\Psi}_{WFE}$ is given in equation (11) with the definition of Ω_{WFE} from equation (13) (The definition of $\widehat{\Psi}_{FE}$ is analogous to that of $\widehat{\Psi}_{WFE}$). And $\widehat{V}_i = [\widehat{X}_i \; \widehat{Z}_i]$ is the "demeaned" explanatory variables, e.g., $\widehat{X}_{it} = X_{it} - \frac{1}{T}\sum_{t'=1}^{T} X_{it}$, and $\hat{u}_i = (\hat{u}_{i1}, \ldots, \hat{u}_{iT})$ with $\hat{u}_{it} = \widehat{Y}_{it} - \widehat{X}_{it}\hat{\beta}_{FE} - \widehat{Z}_{it}^\top \hat{\delta}_{FE}$.

3 The Two-way Fixed Effects Estimator

3.1 Adjusted Matching Estimator

Next, we extend the analysis of the previous section to the two-way fixed effects estimator with unit and time specific dummy variables,

$$Y_{it} \quad = \quad \alpha_i + \gamma_t + X_{it}\,\beta + \epsilon_{it}, \tag{15}$$

where $\mathbb{E}(\epsilon_{it}) = 0$. For the purpose of identification, a restriction such as $\sum_{t=1}^{T} \gamma_t = 0$ must be applied. Without loss of generality, we assume $\gamma_1 = 0$. Let $\hat{\alpha}^{FE^*} = (\hat{\alpha}_1^{FE^*}, \ldots, \hat{\alpha}_N^{FE^*})$, $\hat{\gamma}^{FE^*} = (\hat{\gamma}_1^{FE^*}, \ldots, \hat{\gamma}_T^{FE^*})$, and $\hat{\beta}^{FE^*}$ represent the least squares estimate of $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\gamma = (\gamma_1, \ldots, \gamma_T)$, and β of this model, respectively, i.e.,

$$(\hat{\alpha}^{FE^*}, \hat{\gamma}^{FE^*}, \hat{\beta}^{FE^*}) \quad \equiv \quad \underset{(\alpha,\gamma,\beta)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} (Y_{it} - \alpha_i - \gamma_t - X_{it}\,\beta)^2. \tag{16}$$

The following proposition establishes that this two-way fixed effects estimator can be written as a particular *adjusted* matching estimator.

Proposition 5 (Finite Sample Equivalence between the Two-way Fixed Effects and Adjusted Matching Estimators) *The two-way fixed effects estimator defined in equation* (16) *is equivalent to the following adjusted matching estimator,*

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

where for $x = 0, 1$,

$$\widehat{Y_{it}(x)} = \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{T-1}\sum_{t'\neq t} Y_{it'} + \frac{1}{N-1}\sum_{i'\neq i} Y_{i't} - \frac{1}{(T-1)(N-1)}\sum_{i'\neq i}\sum_{t'\neq t} Y_{i't'} & \text{if } X_{it} = 1-x \end{cases}$$

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

Proof is given in Appendix A.5.

The proposition shows that the estimated counterfactual outcome of a given unit is a function of three averages. First, the average of all the other observations from the same unit and the average of all the other observations from the same time period are added together. Then, the average of the rest of the observations is subtracted from this sum in order to adjust for the fact that two averages rather than one are combined. Since all of these three averages include units of the same treatment status, as in the one-way case (see Proposition 1), the two-way fixed effects estimator adjusts for the attenuation bias due to these "mismatches." This is done via the factor K which is equal to the net proportion of proper matches between the observations of opposite treatment status. However, Proposition 5 shows that to estimate the counterfactual outcome of each unit, all the other observations are used, including the observations from different units and different years. This makes the causal interpretation of the standard two-way fixed effects estimator difficult.

Given this result, we improve the two-way fixed effects estimator in the same manner as done in Section 2 by only matching each observation with other observations of the opposite treatment status to estimate the counterfactual outcome. That is, we use the within-unit matched set defined in equation (5) and the following within-time matched set, both of which do not contain observations of the same treatment status,

$$\mathcal{N}_{it} = \{(i', t') : t' = t, X_{i't'} = 1 - X_{it}\}. \tag{17}$$

The next proposition establishes the finite sample equivalence between a weighted two-way fixed effects estimator and this adjusted matching estimator based on $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$. Unlike the one-way case, however, another adjustment is required because other observations that are used to adjust for time and fixed effects, i.e., $\mathcal{A}_{it}$ defined in the proposition, may contain those of the same treatment status. In fact, there exists no weighted two-way fixed effects estimator that can remove all mismatches.

Proposition 6 (Finite Sample Equivalence between the Weighted Two-way Fixed Effects and Adjusted Matching Estimators) *Assume that the treatment varies within each unit as well as within each time period, i.e.,* $0 < \sum_{t=1}^{T} X_{it} < T$ *for each* i *and* $0 < \sum_{i=1}^{N} X_{it} < N$ *for each* t*. Consider the following adjusted matching estimator,*

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

where $0 \leq C_{it} < \infty$, $\sum_{t=1}^{T}\sum_{i=1}^{N} C_{it} > 0$, *and for* $x = 0, 1$,

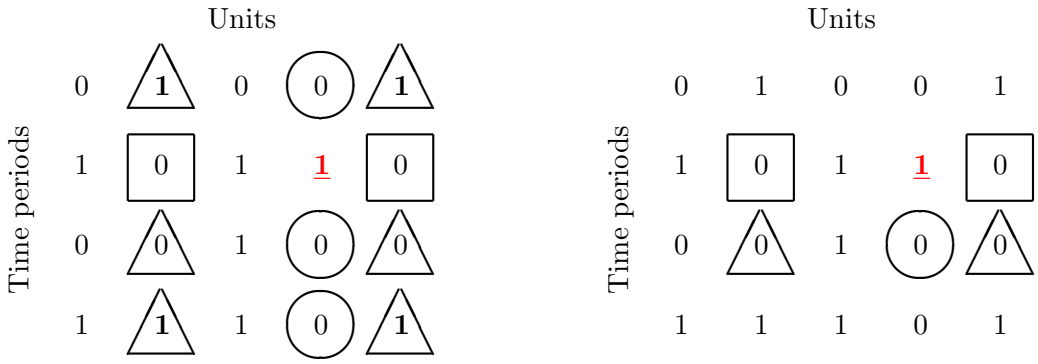
$$\begin{aligned}
\widehat{Y_{it}(x)} &= \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{m_{it}} \sum_{(i,t') \in \mathcal{M}_{it}} Y_{it'} + \frac{1}{n_{it}} \sum_{(i',t) \in \mathcal{N}_{it}} Y_{i't} - \frac{1}{m_{it} n_{it}} \sum_{(i',t') \in \mathcal{A}_{it}} Y_{i't'} & \text{if } X_{it}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

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

$$(\hat{\alpha}^{M^*}, \hat{\gamma}^{M^*}, \hat{\beta}^{M^*}) = \underset{(\alpha,\beta,\gamma)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} (Y_{it} - \alpha_i - \gamma_t - X_{it}\,\beta)^2$$

where

$$\begin{aligned}
\hat{\alpha}_i^{M^*} &= \frac{\sum_{t=1}^{T} W_{it} Y_{it} - \sum_{t=1}^{T} W_{it} \hat{\gamma}_t^{M^*}}{\sum_{t=1}^{T} W_{it}} - \frac{1}{2} \hat{\beta}^{M^*} \\
\hat{\gamma}_t^{M^*} &= \begin{cases} 0 & \text{if } t = 1 \\ \frac{\sum_{i=1}^{N} W_{it} Y_{it} - \sum_{i=1}^{N} W_{it} \hat{\alpha}_i^{M^*}}{\sum_{i=1}^{N} W_{it}} - \frac{1}{2} \hat{\beta}^{M^*} & \text{otherwise.} \end{cases} \\
W_{it} &= \sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'} \quad \text{and} \quad w_{it}^{i't'} = \begin{cases} \frac{C_{i't'} m_{i't'} n_{i't'}}{m_{i't'} n_{i't'} + a_{i't'}} & \text{if } (i,t) = (i',t') \\ \frac{C_{i't'} n_{i't'}}{m_{i't'} n_{i't'} + a_{i't'}} & \text{if } (i,t) \in \mathcal{M}_{i't'} \\ \frac{C_{i't'} m_{i't'}}{m_{i't'} n_{i't'} + a_{i't'}} & \text{if } (i,t) \in \mathcal{N}_{i't'} \\ \frac{C_{i't'} (2X_{it} - 1)(2X_{i't'} - 1)}{m_{i't'} n_{i't'} + a_{i't'}} & \text{if } (i,t) \in \mathcal{A}_{i't'} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$



(a) Adjusted matching estimator (b) Difference-in-differences

Figure 1: An Example of the Binary Treatment Matrix with Five Units and Four Time Periods for the Weighted Two-way Fixed Effects Regressions. Panels (a) and (b) illustrate how observations are used to estimate counterfactual outcomes for the adjusted matching estimator (Proposition 6) and the difference-in-differences estimator (Proposition 7), respectively. In the figures, the red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated. Circles indicate the set of matched observations that are from the same unit, $\mathcal{M}_{it}$ for Panel (a) and $Y_{i,t-1}$ for Panel (b), whereas squares indicate those from the same time period, $\mathcal{N}_{it}$ for Panel (a) and $\mathcal{N}_{it}^*$ for Panel (b). Finally, triangles represent the set of observations that are used to make adjustment for unit and time effects, $\mathcal{A}_{it}$ for Panel (a) and $\mathcal{A}_{it}^*$ for Panel (b). For Panel (a), $\mathcal{A}_{it}$ may include the observations of the same treatment status (bold **1** entries in triangles), leading to an adjustment in the matching estimator. For Panel (b), however, $\mathcal{A}_{it}^*$ only contains control observations and hence no adjustment is required.

Proof is in Appendix A.6.

To illustrate the problem of two-way fixed effects estimators, Panel (a) of Figure 1 presents an example of the binary treatment matrix with five units and four time periods, i.e., $N = 5$ and $T = 4$. In the figure, the red underlined **1** entry represents the treated observation, for which the counterfactual outcome $Y_{it}(0)$ needs to be estimated using other observations. According to Proposition 6, this is estimated as the average of control observations from the same unit $\mathcal{M}_{it}$ (circles in the figure), plus the average of control observations from the same time period $\mathcal{N}_{it}$ (squares), minus the average of adjustment observations, $\mathcal{A}_{it}$ (triangles).

There are two underlying adjustments in this process. First, the adjustment factor K_{it} attempts to address the attenuation bias due to the fact that some of the adjustment observations in $\mathcal{A}_{it}$ contain observations of the same treatment status. The adjustment is done by inflating the estimated treatment effect. In the example of Panel (b) of Figure 1, $\mathcal{A}_{it}$ contains four such observations (bold **1** entries in triangles), i.e., $a_{it} = 4$, and hence the adjustment factor is $0.6 = 6/(6 + 4)$ where the numerator represents the total number of adjustments observations.

The second adjustment occurs due to the fact that the average value of the adjustment observations in $\mathcal{A}_{it}$ is being subtracted rather than added when computing $\widehat{Y_{it}(x)}$. For this reason, $w_{it}^{i't'}$ is defined to take a negative value for the adjustment observations (i,t) in $\mathcal{A}_{i't'}$ which have the treatment status opposite to that of $X_{i't'}$. Consequently, for a small number of observations, the regression weight W_{it} may take a negative value if disproportionately many observations have the same treatment status across both time and unit dimensions.

Proposition 6 establishes the equivalence between this adjusted matching estimator and the weighted two-way fixed effects estimator. Unlike the one-way case, however, without restricting the matched sets, $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$, in certain ways, there exists no weighted two-way fixed effects estimator that is in general equivalent to an *unadjusted* matching estimator. This can be seen from the fact that one cannot eliminate mismatches from $\mathcal{A}_{it}$ for some (i,t) so long as there is variation in the treatment within each unit and time period. This makes it difficult to justify the general use of two-way fixed effects estimators for causal inference.

3.2 Difference-in-Differences

Nevertheless, we present an important (unadjusted) matching estimator with restricted matched sets, $\mathcal{M}_{it}$ and $\mathcal{N}_{it}$, that can be shown to be algebraically equivalent to a weighted two-way fixed effects estimator. Specifically, we consider the application of a multi-period difference-in-differences estimator to the panel data, which is often motivated by the linear two-way fixed effects regression model (e.g., Bertrand *et al.*, 2004; Angrist and Pischke, 2009). Unlike Athey and Imbens (2006), we focus on the linear difference-in-differences model but avoid the assumptions that no unit receives the treatment in the initial period, i.e., $X_{i1} = 0$ for all i, and that after receiving the treatment a unit continues to receive the treatment, i.e., $X_{it} \geq X_{i,t-1}$ for all i and $t = 2, \ldots, T$. Our setting also differs from Abadie (2005) who focuses on the simple case of two time periods with the treatment given only in the second period but allow for semiparametric covariate adjustment via propensity score weighting.

Define the following matching representation of multi-period difference-in-differences estimator,

$$\hat{\beta}^{DID} \;=\; \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right) \tag{18}$$

where for $t = 2, \ldots, T$,

$$\begin{aligned}
\mathcal{N}_{it}^* &= \{(i', t') : t' = t, X_{i't'} = X_{i',t'-1} = 1 - X_{it}\} \\
\mathcal{A}_{it}^* &= \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = X_{i't} = 1 - X_{it}\} \\
n_{it}^* &= |\mathcal{N}_{it}^*| \;=\; |\mathcal{A}_{it}^*| \\
D_{it} &= \mathbf{1}\{n_{it}^* \cdot |X_{it} - X_{i,t-1}| > 0\},
\end{aligned}$$

$D_{i1} = n_{i1}^* = 0$, and for $D_{it} = 1$, we define

$$\widehat{Y_{it}(x)} \;=\; \begin{cases} Y_{it} & \text{if } X_{it} = x \\ Y_{i,t-1} + \frac{1}{n_{it}^*} \sum_{(i',t) \in \mathcal{N}_{it}^*} Y_{i't} - \frac{1}{n_{it}^*} \sum_{(i',t') \in \mathcal{A}_{it}^*} Y_{i't'} & \text{if } X_{it} = 1 - x \end{cases} \tag{19}$$

When the treatment status of a unit changes from one period to the next, i.e., $D_{it} = 1$, this estimator estimates the counterfactual outcome by subtracting from its observed outcome of the first period $Y_{i,t-1}$ the average difference in outcomes between two periods among the other units whose treatment status remains unchanged between these periods and is opposite to the new treatment status of this unit X_{it}.

This estimator is illustrated in Panel (b) of Figure 1. In this example, the counterfactual outcome for the treated unit (represented by the red underlined 1) is estimated as the difference between the outcome under the control condition from the previous period (circle) and the average difference (square minus triangle) between the current and previous periods for the two units which receive the control condition in both time periods. Thus, the difference-in-differences estimator, by construction, avoids mis-matches and therefore eliminates the need for adjustment.

The next proposition establishes the algebraic equivalence between this difference-in-differences, matching, and weighted two-way fixed effects estimators. Heckman *et al.* (1997, 1998a) show that the difference-in-differences estimator can be interpreted as a matching estimator. What we prove is that the estimator is also algebraically equivalent to weighted two-way fixed effects estimator. Following these authors, the proposition allows for the weighted average of control (treated) outcome differences to be an estimate for the counterfactual of the treated outcome difference rather than the simple average as formulated above.

Proposition 7 (Algebraic Equivalence between the Difference-in-Differences, Matching, and Weighted Two-way Fixed Effects Estimators) *Assume that there is at least one treated and control unit, i.e.,* $0 < \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} < NT$, *and that there is at least one unit with* $D_{it} = 1$, *i.e.,* $0 < \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}$. *Consider the following matching estimator that is slightly more*

general than $\hat{\beta}^{DID}$ *in equation* (18),

$$\ddot{\beta}^{M^*} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}D_{it}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where $0 \leq C_{it} < \infty$, $\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it} > 0$,

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

and $v_{it}^{i't}$ *is the normalized pre-defined non-negative weight for observation* (i',t) *which is matched with observation* (i,t) *such that* $\sum_{(i',t)\in\mathcal{N}_{it}^*} v_{it}^{i't} = \sum_{(i',t')\in\mathcal{A}_{it}^*} v_{it}^{i't} = 1$. *Then, this estimator is equivalent to the following weighted two-way fixed effects estimator,*

$$(\ddot{\alpha}^{M^*}, \ddot{\gamma}^{M^*}, \ddot{\beta}^{M^*}) = \underset{(\alpha,\beta,\gamma)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \alpha_i - \gamma_t - X_{it}\,\beta)^2$$

where

$$\ddot{\alpha}_i^{M^*} = \frac{\sum_{t=1}^{T} W_{it}Y_{it} - \sum_{t=1}^{T} W_{it}\ddot{\gamma}_t^{M^*}}{\sum_{t=1}^{T} W_{it}} - \frac{1}{2}\ddot{\beta}^{M^*}$$

$$\ddot{\gamma}_t^{M^*} = \begin{cases} 0 & \text{if } t = 1 \\ \frac{\sum_{i=1}^{N} W_{it}Y_{it} - \sum_{i=1}^{N} W_{it}\ddot{\alpha}_i^{M^*}}{\sum_{i=1}^{N} W_{it}} - \frac{1}{2}\ddot{\beta}^{M^*} & \text{otherwise.} \end{cases}$$

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

Proof is in Appendix A.7. The use of C_{it} means that the proposition applies to the average treatment effect for a subpopulation of interest.

Thus, the difference-in-differences estimate of the average treatment effect for the treated can be obtained by fitting the weighted two-way fixed effects regression model. In general, the resulting regression weights, W_{it}, are different across observations and hence the (unweighted) standard two-way fixed effects estimator differs from the difference-in-differences estimators. It is interesting to note that in the case of two time periods with the treatment given only in the second period the regression weights can be different across observations and yet the estimated average treatment effect from the weighted two-way fixed effects model is identical to the one based on the standard two-way fixed effects regression.

4 Empirical Illustration

In this section, we illustrate the proposed matching framework to understand and improve fixed effects regression models by estimating the effects of GATT (General Agreement on Tariffs and Trade) membership on trade between countries. We show that in some cases the proposed weighted fixed effects estimators give substantively different results from the standard fixed effects model.

4.1 Effects of GATT on Trade and Fixed Effects Regression

Does GATT membership increase international trade? Bagwell and Staiger (1999) provide a theoretical reason why GATT may facilitate trade among its members. They argue that each state's incentive to manipulate terms-of-trade entails inefficient equilibrium, and therefore GATT's principles of reciprocity and non-discrimination help them reach a more efficient multilateral outcomes. In addition to these reciprocal terms-of-trade incentives, Maggi and Rodrguez-Clare (2007) offer a political explanation of why governments may want to sign a trade agreement even *unilaterally*. Governments can make credible commitments to their domestic audiences so that they suffer less from the political pressure to protect domestic industries.

Despite this theoretical development, Rose (2004) finds little empirical support for the existing theories. Based on the standard "gravity" model with year fixed effects applied to dyadic trade data, he consistently finds economically and statistically insignificant effect of GATT membership (and its successor World Trade Organization or WTO) on trade volume between countries. This finding led to subsequent debates among empirical researchers as to whether or not GATT actually promotes trade (e.g., Gowa and Kim, 2005; Subramanian and Wei, 2007; Goldstein *et al.*, 2007; Tomz *et al.*, 2007; Rose, 2007).

As shown in Section 2, the standard one-way fixed effects model such as the one used by Rose (2004) is equivalent to an adjusted matching estimator, where each observation is matched with the average of all the other observations of the same unit. In the case of the fixed effects regression model given in equation (20), this implies that a treated dyad (e.g., a dyad with both GATT members) in year t is matched with all other dyads in the same year t regardless of their treatment status. Our theoretical results imply that this estimation strategy may be subject to potential biases when there exists heterogeneous treatment effect and/or treatment assignment probability. Indeed, Subramanian and Wei (2007) find substantial heterogeneity in the effects of GATT/WTO on trade. Thus, we examine the robustness of findings when counterfactual outcomes are estimated based on the observations with opposite treatment status. We also consider several other substantively

reasonable ways to estimate counterfactual outcomes and investigate the sensitivity of the results to different causal assumptions underlying fixed effects models.

Finally, some scholars have used the two-way fixed effects model with both dyadic and year fixed effects to estimate the effects of GATT/WTO on trade (e.g., Tomz *et al.*, 2007). However, our analysis in Section 3 suggests that this model uses all observations to estimate the counterfactual outcome for any given dyad in a particular year and this shortcoming cannot be corrected within the fixed effects modeling framework. Thus, in our analysis, we do not consider the application of two-way fixed effects models.

4.2 Data and Methods

We use the data set from Tomz *et al.* (2007) which updates and corrects some minor errors in the original data used by Rose (2004). Unlike Rose (2004), however, our analysis is restricted to the period between 1948 and 1994 so that we focus on the effects of GATT and avoid conflating them with the effects of the World Trade Organization (WTO), which replaced GATT in 1995 after the conclusion of the Uruguay Round. WTO has functionalities that are different from GATT. In particular, it possesses a dispute settlement mechanism that became more legalized than GATT with end of defendant veto and addition of appellate body. In addition, a number of developing countries joined the WTO, which makes it difficult to justify pooling of the effects of GATT and WTO (Subramanian and Wei, 2007). As shown below, this restriction does not significantly change the original conclusion of Rose (2004), but it leads to a conceptually cleaner analysis. This yields a dyadic data set of bilateral international trade where a total number of countries is 162 and a total number of (dyad-year) observations is 196,207.

We begin by estimating the following standard fixed effects regression model (with year fixed effects),

$$\ln Y_{it} \quad = \quad \alpha_t + X_{it}\beta + Z_{it}^{\top}\delta + \epsilon_{it} \tag{20}$$

where i and t denote a dyad and year, respectively. In this model, Y_{it} represents the trade volume for dyad i in year t, whereas X_{it} is a binary treatment variable which equals one (zero) when dyad i receives (does not receive) the "treatment" at time t. In addition, following the original analysis, we include a vector of 15 dyad-varying covariates in year t as Z_{it}. These variables include log distance between the two countries, log product real GDP, and log product real GDP per capita (Rose, 2004, see Column 1 of Table 1 for the full list of covariates).

The key difference between this model and the one used by Rose is that the latter includes two

binary treatment variables at the same time: one indicating whether both countries in a dyad are members of GATT/WTO in a given year and the other indicating whether only one country is a member.[1] In contrast, we fit a separate model for each treatment variable. First, we estimate the model given in equation (20) with the treatment variable X_{it} indicating whether both countries in dyad i are members of GATT or only one is a member. For this analysis, we exclude dyad-years that do not have a GATT member, thereby estimating the effect of having two GATT members versus having only one GATT member. Second, we estimate the same model but this time we estimate the effect of only one country being a GATT member versus no GATT member at all. This analysis therefore excludes those dyad-years which have two GATT members.

For each of these two models, we conduct three separate analyses. First, we fit the standard one-way fixed effects model with year fixed effects as done in Rose (2004). Second, we fit the weighted one-way fixed effects of the following specification,

$$\underset{(\alpha,\beta,\delta)}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (\ln Y_{it} - \alpha_t - X_{it}\beta - Z_{it}^{\top}\delta)^2 \tag{21}$$

where the expression for W_{it} is given in Theorem 1 (note that subscripts i and t must be switched because in this model we have year fixed effects rather than unit fixed effects). This model ensures that within each year treated dyads are compared with untreated dyads. Third, we use the propensity score weighting given in Proposition 3 to adjust the dyad-varying covariates Z_{it} (again, note that subscripts i and t must be switched). Within each year, we use the logistic regression to obtain the estimated propensity score $\hat{\pi}(Z_{it})$.

Finally, we repeat the above analysis based on the weighted one-way fixed effects and propensity score weighting by placing an additional restriction. Specifically, for each dyad in a given year, we restrict its matched set to be a group of dyads in the same year who have the opposite treatment status and share at least one country. For instance, suppose country A and country B in dyad i are both GATT members in year t, and thus dyad i is a treated observation. The restriction requires that the control observations for this dyad should only include the dyads in year t which contain either country A or country B along with a non-GATT member. Such a restriction may be appealing because it avoids the comparison between dyads which have completely different countries. Our weighted fixed effects framework can accomodate this and other similar restrictions to construct proper counterfactuals for each observation.

[1] We focus on the effects of "formal" membership. See Tomz *et al.* (2007) for a detailed discussion.

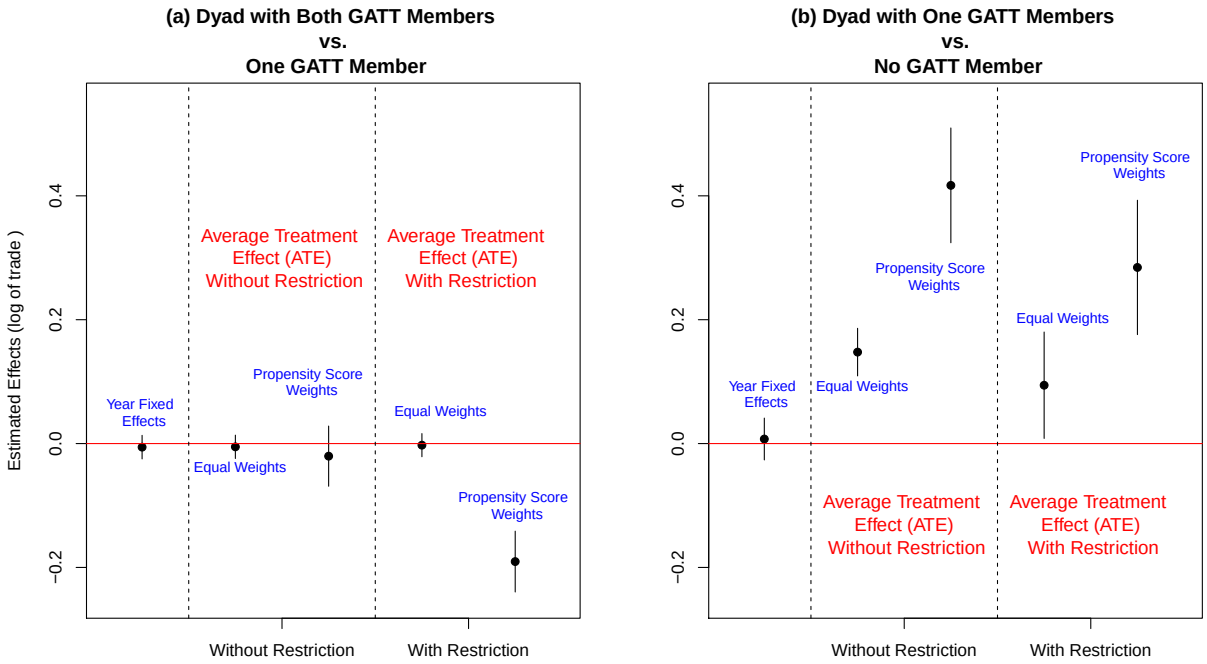


Figure 2: Estimated Average Treatment Effects of GATT Membership on Trade. Panel (a) shows estimated effects when treatment group includes dyads with both GATT members and control group includes dyads with only one GATT member. Panel (b) presents estimated effects when treated units are dyads with only one GATT member while controlled units are dyads with no GATT member. Left, middle and right part of each panel corresponds to the standard one-way year fixed effects model, the weighted one-way fixed effects models without restriction, and the weighted one-way fixed effects models with restriction, respectively. The restriction implies that in a given year a dyad is matched with other dyads in the same year that share one country. Vertical lines represent 95% confidence intervals.

4.3 Empirical Results

Figure 2 summarizes the results of our analysis. We estimate the effect of GATT membership on trade based on the two definitions of treatment. Panel (a) compares trade when both countries are members to the case where only one country is a formal member, whereas Panel (b) presents the results when treated units are dyads with one member and controlled units are dyads with no member. Within each panel, we present the results based on the standard one-way fixed effects model (first result), the weighted one-way fixed effects model with equal weights among matched observations (second result), and the same model with propensity score based weights (the third result). The last two analyses are repeated with the additional restriction (the last two results).

Consistent with Rose (2004)'s original finding, we find little effect of GATT membership on trade using the standard one-way fixed effects based model for both analyses. In contrast, our analysis that eliminates "mismatches" using the weighted one-way fixed effects model shows that

having one GATT member increases trade by 33% ($\approx \exp(0.284) - 1$) (with 95% confidence interval of [0.176, 0.393]) when compared with dyads having no GATT member whereas there is no statistically significant difference between dyads with two GATT members and those with one GATT member. The latter result changes when we place the additional restriction and limit our matched group to those dyads which share one country. According to this method, the effect of GATT membership is positive regardless of whether or not we impose the additional restriction in Panel (b).

In sum, our analysis suggests that the empirical results based on fixed effects models can depend on the causal assumptions underlying these models. We have shown that the effects of GATT membership can vary quite a bit depending on how counterfactual outcomes are estimated for each observation.

5 Concluding Remarks

In this paper, we propose a straightforward to improve standard linear fixed effects regression models, which have been shown to provide inconsistent estimates of the average treatment effect. We establish the algebraic equivalence relationships between weighted one-way fixed effects estimators and a broad class of matching estimators. The proposed weighted linear fixed effects estimators are consistent for the average treatment effect and can accomodate various identification strategies including propensity score weighting, first differencing, and difference-in-differences. Therefore, the proposed weighted fixed effects regression estimators can also be used as a specification test for the standard fixed effects regression models (White, 1980b). It is surprising that different causal assumptions that underlie each approach can be represented as different regression weights within the framework of linear fixed effects regression models.

To prove these theoretical results, we use the framework of matching methods and identify the information used implicitly by fixed effects regression models to estimate counterfactual outcomes necessary for causal inference. This analysis highlights the sources of potential bias and inefficiency of standard linear fixed effects estimators and suggests weighted linear fixed effects models as a simple way to address these problems. Our analysis also makes it clear that fixed effects and matching estimators are equivalent and neither method represents a magic bullet solution to endogeneity problems encountered in observational studies. A commonly held belief that fixed effects estimators can adjust for unobservables but matching estimators cannot is false. What is essential is to understand what information each method uses in order to estimate counterfactual outcomes

that are necessary for causal inference.

References

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

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. (2011). A martingale representation for matching estimators. Tech. rep., Department of Economics, Harvard University.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* **66**, 2, 249–288.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics* **19**, 1, 2–16.

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

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74**, 2, 431–497.

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

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

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2011). Average and quantile effects in nonseparable panel models. Tech. rep., Department of Economics, Massachusetts Institute of Technology.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–313.

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.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics* **40**, 2, 180–193.

Goldstein, J. L., Rivers, D., and Tomz, M. (2007). Institutions in international relations: Understanding the effects of the gatt and the wto on world trade. *International Organization* **61**, 1, 37–67.

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.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hahn, J. (2001). Comment: Binary regressors in nonlinear panel-data models with fixed effects. *Journal of Business & Economic Statistics* **19**, 1, 16–17.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* **99**, 467, 609–618.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998a). Characterizing selection bias using experimental data. *Econometrica* **66**, 1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* **64**, 4, 605–654.

Heckman, J. J., Ichimura, H., and Todd, P. (1998b). Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**, 2, 261–294.

Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 4, 1307–1338.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. Le Cam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233, Berkeley. University of California Press.

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.

Iacus, S., King, G., and Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* **106**, 493, 345–361.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **171**, 2, 481–502.

Jones, B. and Kenward, M. G. (2003). *Design and Analysis of Cross-over Trials*. Chapman & Hall, London, 2nd edn.

Maggi, G. and Rodrguez-Clare, A. (2007). A political-economy theory of trade agreements. *The American Economic Review* **97**, 4, 1374–1406.

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

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 387, 516–524.

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.

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.

Wallace, T. D. and Hussain, A. (1969). The use of error component models in combining cross section with time series data. *Econometrica* **37**, 1, 55–72.

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.

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

A Mathematical Appendix

We define the following quantities, which will be used throughout this appendix.

$$\begin{aligned}
\overline{X}_i &= \frac{1}{T}\sum_{i=1}^{N} X_{it}, \quad \overline{X}_t = \frac{1}{N}\sum_{t=1}^{T} X_{it}, \quad \overline{X} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it} \\
\overline{Y}_i &= \frac{1}{T}\sum_{i=1}^{N} Y_{it}, \quad \overline{Y}_t = \frac{1}{N}\sum_{t=1}^{T} Y_{it}, \quad \overline{Y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}
\end{aligned}$$

A.1 Proof of Proposition 1

$$\begin{aligned}
\hat{\beta}^{FE} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}Y_{it} - T\sum_{i=1}^{N}\overline{X}_i\overline{Y}_i}{NT\overline{X} - T\sum_{i=1}^{N}\overline{X}_i^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(2X_{it}-1)(Y_{it}-\overline{Y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(2X_{it}-1)(X_{it}-\overline{X}_i)} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(Y_{it}-\overline{Y}_i)+(1-X_{it})(\overline{Y}_i-Y_{it})\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(X_{it}-\overline{X}_i)+(1-X_{it})(\overline{X}_i-X_{it})\right\}} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{1}{T}\sum_{t'\neq t}Y_{it'}-\frac{1}{T}Y_{it}\right)+(1-X_{it})\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}+\frac{1}{T}Y_{it}-Y_{it}\right)\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(X_{it}-\frac{1}{T}\sum_{t'\neq t}X_{it'}-\frac{1}{T}X_{it}\right)+(1-X_{it})\left(\frac{1}{T}\sum_{t'\neq t}X_{it'}+\frac{1}{T}X_{it}-Y_{it}\right)\right\}} \\
&= \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{1}{T-1}\sum_{t'\neq t}Y_{it'}\right)+(1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t}Y_{it'}-Y_{it}\right)\right\}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(X_{it}-\frac{1}{T-1}\sum_{t'\neq t}X_{it'}\right)+(1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t}X_{it'}-X_{it}\right)\right\}} \\
&= \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{1}{T-1}\sum_{t'\neq t}Y_{it'}\right)+(1-X_{it})\left(\frac{1}{T-1}\sum_{t'\neq t}Y_{it'}-Y_{it}\right)\right\}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\cdot\frac{1}{T-1}\sum_{t'\neq t}(1-X_{it'})+(1-X_{it})\cdot\frac{1}{T-1}\sum_{t'\neq t}X_{it'}\right\}} \\
&= \frac{1}{K}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right)\right\}
\end{aligned}$$

□

A.2 Proof of Proposition 3

We begin by defining a new matching estimator using the transformed outcome Y_{it}^* as follows,

$$\hat{\beta}^{M^*} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}^*(1)}-\widehat{Y_{it}^*(0)}\right)$$

where

$$\widehat{Y_{it}^*(1)} = \begin{cases} Y_{it}^* & \text{if } X_{it}=1 \\ \frac{\sum_{t'=1}^{T}X_{it'}Y_{it'}^*}{\sum_{t'=1}^{T}X_{it'}} & \text{if } X_{it}=0 \end{cases} \quad \text{and} \quad \widehat{Y_{it}^*(0)} = \begin{cases} \frac{\sum_{t'=1}^{T}(1-X_{it'})Y_{it'}^*}{\sum_{t'=1}^{T}(1-X_{it'})} & \text{if } X_{it}=1 \\ Y_{it}^* & \text{if } X_{it}=0 \end{cases}$$

Then, Proposition 2 implies that we only need to show $\hat{\beta}^{M^*} = \hat{\beta}^W$. This equality can be shown as follows.

$$
\begin{aligned}
& \hat{\beta}^{M^*} \\
= \; & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{ X_{it}\left(Y_{it}^* - \frac{\sum_{t'=1}^{T}(1-X_{it'})Y_{it'}^*}{\sum_{t'=1}^{T}(1-X_{it'})} \right) + (1-X_{it})\left(\frac{\sum_{t'=1}^{T} X_{it'}Y_{it'}^*}{\sum_{t'=1}^{T} X_{it'}} - Y_{it}^* \right) \right\} \\
= \; & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{ X_{it}\left(\frac{\sum_{t'=1}^{T} X_{it'} + \sum_{t'=1}^{T}(1-X_{it'})}{\sum_{t'=1}^{T} X_{it'}} \right) Y_{it}^* + (1-X_{it})\left(\frac{-\sum_{t'=1}^{T} X_{it'} - \sum_{t'=1}^{T}(1-X_{it'})}{\sum_{t'=1}^{T}(1-X_{it'})} \right) Y_{it}^* \right\} \\
= \; & \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{ X_{it}\left(\frac{1}{\sum_{t'=1}^{T} X_{it'}} \right) Y_{it}^* - (1-X_{it})\left(\frac{1}{\sum_{t'=1}^{T}(1-X_{it'})} \right) Y_{it}^* \right\} \\
= \; & \frac{1}{N}\sum_{i=1}^{N}\left\{ \sum_{t=1}^{T}\frac{X_{it}Y_{it}}{\hat{\pi}(Z_{it})} \Big/ \sum_{t=1}^{T}\frac{X_{it}}{\hat{\pi}(Z_{it})} - \sum_{t=1}^{T}\frac{(1-X_{it})Y_{it}}{1-\hat{\pi}(Z_{it})} \Big/ \sum_{t=1}^{T}\frac{(1-X_{it})}{1-\pi(Z_{it})} \right\}
\end{aligned}
$$

where the last equality follows from the definition of Y_{it}^*. □

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,

$$
\begin{aligned}
& \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)\alpha_i^* \\
= \; & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)\alpha_i^* \right) \\
= \; & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'} w_{it}^{i't'}(2X_{it}-1)\alpha_i^* + (1-X_{i't'}) w_{it}^{i't'}(2X_{it}-1)\alpha_i^* \right) \\
= \; & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{ X_{i't'}\left(C_{i't'}\alpha_i^* - \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'} C_{i't'}(1-X_{it})\alpha_i^* \right) \right. \\
& \qquad \left. +(1-X_{i't'})\left(-C_{i't'}\alpha_i^* + \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'} C_{i't'} X_{it}\alpha_i^* \right) \right. \\
= \; & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{ C_{i't'}X_{i't'}\left(\alpha_i^* - \alpha_i^*\right) + C_{i't'}(1-X_{i't'})\left(-\alpha_i^* + \alpha_i^*\right) \right\} \; = \; 0
\end{aligned}
\tag{22}
$$

The second algebraic equality we prove is the following,

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

$$
\begin{aligned}
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left[X_{i't'}\left(C_{i't'} + \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'}C_{i't'}(1-X_{it})\right) + (1-X_{i't'})\left(C_{i't'} + \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'}C_{i't'}X_{it}\right)\right] \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{C_{i't'}X_{i't'}\,(1+1) + C_{i't'}(1-X_{i't'})\,(1+1)\right\} \;=\; 2\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it} \qquad (23)
\end{aligned}
$$

Next, define $\delta_i = \alpha_i + \beta/2$ and $\xi = \beta/2$, and consider the following weighted least squares estimates,

$$
(\hat{\delta},\hat{\xi}) \;=\; \underset{(\delta,\xi)}{\arg\min}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \delta_i - (2X_{it}-1)\xi)^2
$$

where $\delta = (\delta_1,\ldots,\delta_N)$. Clearly, $\tilde{\alpha}_i^M = \hat{\delta}_i - \tilde{\beta}^M/2$ and $\tilde{\beta}^M = 2\hat{\xi}$ because (δ,ξ) is a linear one-to-one transformation of (α,β). Thus, using the above algebraic equalities, we can derive the desired result,

$$
\begin{aligned}
\tilde{\beta}^M &= \frac{2\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{W_{it}(2X_{it}-1)Y_{it} - W_{it}(2X_{it}-1)\hat{\delta}_i\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right)(2X_{it}-1)Y_{it} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)Y_{it}\right) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}(2X_{it}-1)Y_{it} + (1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)Y_{it}\right) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T}\left[X_{i't'}\left(C_{i't'}Y_{i't'} - \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'}C_{i't'}(1-X_{it})Y_{it}\right)\right. \\
&\qquad\qquad \left. +(1-X_{i't'})\left(-C_{i't'}Y_{it} + \sum_{(i,t)\in\mathcal{M}_{i't'}} v_{it}^{it'}C_{i't'}X_{it}Y_{it}\right)\right] \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}\left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)
\end{aligned}
$$

where the second equality follows from equations (22) and (23). □

A.4 Proof of Proposition 4

Proof of this proposition is straightforward. First, we note that the first difference estimator is given by,

$$
\hat{\beta}^{FD} \;=\; \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta Y_{it}\Delta X_{it}}{\sum_{i=1}^{N}\sum_{t=2}^{T}(\Delta X_{it})^2}.
$$

Then, the simple algebraic manipulation shows the equivalence between $\hat{\beta}^{FD}$ and $\ddot{\beta}^{M}$. Finally, the equivalence between $\hat{\beta}^{FD}$ and the weighted one-way fixed effects regression follows directly from Theorem 1 because of the equivalence between $\hat{\beta}^{FD}$ and the matching estimator. □

A.5 Proof of Proposition 5

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

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(Y_{it}-\overline{Y_i}-\overline{Y_t}+\overline{Y})-(1-X_{it})(Y_{it}-\overline{Y_i}-\overline{Y_t}+\overline{Y})\right\} \\
= &\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}+\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right.\right. \\
&\qquad\left.-\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\} \\
&-(1-X_{it})\left\{Y_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}Y_{it'}+\frac{1}{NT}\sum_{t'\neq t}Y_{it'}\right)\right. \\
&\qquad\left.\left.-\left(\frac{1}{N}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}Y_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right] \\
= &\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right. \\
&\qquad\left.-(1-X_{it})\left\{\frac{(N-1)(T-1)}{NT}Y_{it}-\frac{N-1}{NT}\sum_{t'\neq t}Y_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}Y_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}Y_{i't'}\right\}\right] \\
= &\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}\left(Y_{it}-\frac{1}{T-1}\sum_{t'=1}^{T}Y_{it'}+\frac{1}{N-1}\sum_{i'=1}^{N}Y_{i't}-\frac{1}{(T-1)(N-1)}\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}\right)\right. \\
&\qquad\left.-(1-X_{it})\left(\frac{1}{T-1}\sum_{t'=1}^{T}Y_{it'}+\frac{1}{N-1}\sum_{i'=1}^{N}Y_{i't}-\frac{1}{(T-1)(N-1)}\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}Y_{i't'}-Y_{it}\right)\right\}. \\
= &\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{Y_{it}(1)}-\widehat{Y_{it}(0)}\right) \qquad (24)
\end{aligned}
$$

The second algebraic equality we prove is the following,

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{X_{it}(X_{it}-\overline{X_i}-\overline{X_t}+\overline{X})-(1-X_{it})(X_{it}-\overline{X_i}-\overline{X_t}+\overline{X})\right\}\\
=\;&\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{X_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}X_{it'}+\frac{1}{NT}\sum_{t'\neq t}X_{it'}\right)\right.\right.\\
&\qquad\left.-\left(\frac{1}{N}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}X_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\\
&\quad-(1-X_{it})\left\{X_{it}\left(1-\frac{1}{N}-\frac{1}{T}+\frac{1}{NT}\right)-\left(\frac{1}{T}\sum_{t'\neq t}X_{it'}+\frac{1}{NT}\sum_{t'\neq t}X_{it'}\right)\right.\\
&\qquad\left.\left.-\left(\frac{1}{N}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}X_{i't}\right)+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\right]\\
=\;&\sum_{i=1}^{N}\sum_{t=1}^{T}\left[X_{it}\left\{\frac{(N-1)(T-1)}{NT}X_{it}-\frac{N-1}{NT}\sum_{t'\neq t}X_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\right.\\
&\quad\left.-(1-X_{it})\left\{\frac{(N-1)(T-1)}{NT}X_{it}-\frac{N-1}{NT}\sum_{t'\neq t}X_{it'}-\frac{T-1}{NT}\sum_{i'\neq i}X_{i't}+\frac{1}{NT}\sum_{i'\neq i}\sum_{t'\neq t}X_{i't'}\right\}\right]\\
=\;&\frac{(T-1)(N-1)}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left\{X_{it}\left(\frac{1}{T-1}\sum_{t'=1}^{T}(1-X_{it'})+\frac{1}{N-1}\sum_{i'=1}^{N}(1-X_{i't})-\frac{\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}(1-X_{i't'})}{(T-1)(N-1)}\right)\right.\right.\\
&\qquad\left.\left.+(1-X_{it})\left(\frac{1}{T-1}\sum_{t'=1}^{T}X_{it'}+\frac{1}{N-1}\sum_{i'=1}^{N}X_{i't}-\frac{1}{(T-1)(N-1)}\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T}X_{i't'}\right)\right\}\right]\\
=\;&K(T-1)(N-1). \qquad (25)
\end{aligned}
$$

Next, we consider the following least squares estimates,

$$
\tilde{\xi} \;=\; \arg\min_{\xi}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{(y_{it}-\overline{y_i}-\overline{y_t}+\overline{y})-(X_{it}-\overline{X_i}-\overline{X_t}+\overline{X})\xi\right\}^2
$$

We know from Wallace and Hussain (1969) that $\hat{\beta}^{FE^*}=\tilde{\xi}$. Thus, using the above algebraic equalities, we can derive the desired result as follows,

$$
\begin{aligned}
&\hat{\beta}^{FE^*}\\
=\;&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})(Y_{it}-\overline{Y}_i-\overline{Y}_t+\overline{Y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\overline{X}_i-\overline{X}_t+\overline{X})^2}\\
=\;&\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}Y_{it}-T\sum_{i=1}^{N}\overline{X}_i\overline{Y}_i-N\sum_{t=1}^{T}\overline{X}_t\overline{Y}_t+NT\overline{XY}}{NT\overline{X}-T\sum_{i=1}^{N}\overline{X}_i^2-N\sum_{t=1}T\overline{X}_t^2+NT\overline{X}^2}
\end{aligned}
$$

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

where the last equality follows from equation (24) and (25). □

A.6 Proof of Proposition 6

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

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}W_{it}(2X_{it}-1)\alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}^{i't'}(2X_{it}-1)\alpha_i^*\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}X_{i't'}w_{it}^{i't'}(2X_{it}-1)\alpha_i^*+(1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)\alpha_i^*\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}\left(\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}\left(a_{i't'}+m_{i't'}n_{i't'}-a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\alpha_i^*\right. \\
&\quad \left.+(1-X_{i't'})\left(\frac{-C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}\left(m_{i't'}n_{i't'}-a_{i't'}+a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\alpha_i^*\right\} \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}C_{i't'}(0)\,\alpha_i^*+(1-X_{i't'})C_{i't'}(0)\,\alpha_i^*\right\} \;=\; 0. \qquad (26)
\end{aligned}
$$

Similarly, for any constants $(\gamma_1^*,\ldots,\gamma_T^*)$, the following equality holds,

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T}W_{it}(2X_{it}-1)\gamma_t^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}w_{it}^{i't'}(2X_{it}-1)\gamma_t^*\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}X_{i't'}w_{it}^{i't'}(2X_{it}-1)\gamma_t^*+(1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)\gamma_t^*\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}\left(\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}\left(a_{i't'}+m_{i't'}n_{i't'}-a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\gamma_t^*\right. \\
&\quad \left.+(1-X_{i't'})\left(\frac{-C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}+\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}}-\frac{C_{i't'}\left(m_{i't'}n_{i't'}-a_{i't'}+a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\gamma_t^*\right\} \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}C_{i't'}(0)\,\gamma_t^*+(1-X_{i't'})C_{i't'}(0)\,\gamma_t^*\right\} \;=\; 0. \qquad (27)
\end{aligned}
$$

The third algebraic equality we prove is the following,

$$
\begin{aligned}
& \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
= & \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}\left(\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} + \frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} + \frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} + \frac{C_{i't'}\left(a_{i't'}-m_{i't'}n_{i't'}+a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\right. \\
& \left. +(1-X_{i't'})\left(\frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} + \frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} + \frac{C_{i't'}m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} - \frac{C_{i't'}\left(m_{i't'}n_{i't'}-a_{i't'}-a_{i't'}\right)}{m_{i't'}n_{i't'}+a_{i't'}}\right)\right\} \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(2X_{i't'}C_{i't'} + 2(1-X_{i't'})C_{i't'}\right) = 2\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}. \qquad (28)
\end{aligned}
$$

Next, define $\delta_i = \alpha_i + \beta/2$ and $\xi = \beta/2$, and consider the following weighted least squares estimates,

$$
(\bar{\delta}, \bar{\gamma}, \bar{\xi}) \quad = \quad \underset{(\delta,\gamma,\xi)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \delta_i - \gamma_t - (2X_{it}-1)\xi)^2
$$

where $\delta = (\delta_1, \ldots, \delta_N)$. Clearly, $\bar{\alpha}_i^{M^*} = \bar{\delta}_i - \tilde{\beta}^{M^*}/2$ and $\tilde{\beta}^{M^*} = 2\bar{\xi}$ because (δ, ξ) is a linear one-to-one transformation of (α, β). Thus, using the above algebraic equalities, we can derive the desired result,

$$
\begin{aligned}
\tilde{\beta}^{M^*} &= \frac{2\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{W_{it}(2X_{it}-1)Y_{it} - W_{it}(2X_{it}-1)\bar{\delta}_i - W_{it}(2X_{it}-1)\bar{\gamma}_t\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right)(2X_{it}-1)Y_{it} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)Y_{it}\right) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}(2X_{it}-1)Y_{it} + (1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)Y_{it}\right) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} \frac{m_{i't'}n_{i't'}}{m_{i't'}n_{i't'}+a_{i't'}} \left\{X_{i't'}C_{i't'}\left(Y_{i't'} - \frac{\sum_{(i,t)\in\mathcal{M}_{i't'}} Y_{it}}{m_{i't'}} - \frac{\sum_{(i,t)\in\mathcal{N}_{i't'}} Y_{it}}{n_{i't'}} + \right.\right. \\
&\qquad \left.\left. \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}} Y_{it}}{m_{i't'}n_{i't'}}\right) + (1-X_{i't'})C_{i't'}\left(\frac{\sum_{(i,t)\in\mathcal{M}_{i't'}} Y_{it}}{m_{i't'}} + \frac{\sum_{(i,t)\in\mathcal{N}_{i't'}} Y_{it}}{n_{i't'}} - \frac{\sum_{(i,t)\in\mathcal{A}_{i't'}} Y_{it}}{m_{i't'}n_{i't'}} - Y_{i't'}\right)\right\}
\end{aligned}
$$

$$= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{C_{it}}{K_{it}} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)}\right)$$

where the second equality follows from equation (26),(27) and (28). □

A.7 Proof of Proposition 7

The proof of this proposition proceeds in a manner similar to that of Proposition 6 in Appendix A.6. To economize on notation, let $C^*_{i't'} \equiv C_{it}D_{it}$. We begin this proof by establishing three algebraic equalities. First, we prove that for any constants $(\alpha^*_1, \ldots, \alpha^*_N)$, the following equality holds,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)\alpha^*_i \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)\alpha^*_i\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}(2X_{it}-1)\alpha^*_i + (1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)\alpha^*_i\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left[X_{i't'}\left\{C^*_{i't'} - C^*_{i't'} - \sum_{(i,t)\in\mathcal{N}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'} + \sum_{(i,t)\in\mathcal{A}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'}\right\}\alpha^*_i\right. \\
&\qquad\qquad \left. +(1-X_{i't'})\left\{-C^*_{i't'} + C^*_{i't'} + \sum_{(i,t)\in\mathcal{N}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'} - \sum_{(i,t)\in\mathcal{A}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'}\right\}\alpha^*_i\right] \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}C^*_{i't'}(0)\,\alpha^*_i + (1-X_{i't'})C^*_{i't'}(0)\,\alpha^*_i\right\} \;=\; 0. \qquad (29)
\end{aligned}$$

Similarly, for any constants $(\gamma^*_1, \ldots, \gamma^*_T)$, the following equality holds,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(2X_{it}-1)\gamma^*_t \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)\gamma^*_t\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}(2X_{it}-1)\gamma^*_t + (1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)\gamma^*_t\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left[X_{i't'}\left\{C^*_{i't'} - C^*_{i't'} - \sum_{(i,t)\in\mathcal{N}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'} + \sum_{(i,t)\in\mathcal{A}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'}\right\}\gamma^*_t\right. \\
&\qquad\qquad \left. +(1-X_{i't'})\left\{-C^*_{i't'} + C^*_{i't'} + \sum_{(i,t)\in\mathcal{N}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'} - \sum_{(i,t)\in\mathcal{A}^*_{i't'}} v_{i't'}^{it'}C^*_{i't'}\right\}\gamma^*_t\right] \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left\{X_{i't'}C^*_{i't'}(0)\,\gamma^*_t + (1-X_{i't'})C^*_{i't'}(0)\,\gamma^*_t\right\} \;=\; 0. \qquad (30)
\end{aligned}$$

The third algebraic equality we prove is the following,

$$
\begin{aligned}
& \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
= & \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'} + (1-X_{i't'})w_{it}^{i't'}\right) \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left[X_{i't'}\left\{C_{i't'}^{*} + C_{i't'}^{*} + \sum_{(i,t)\in\mathcal{N}_{i't'}^{*}} v_{i't'}^{it'}C_{i't'}^{*} - \sum_{(i,t)\in\mathcal{A}_{i't'}^{*}} v_{i't'}^{it'}C_{i't'}^{*}\right\}\right. \\
& \qquad \left. + (1-X_{i't'})\left\{-C_{i't'}^{*} + C_{i't'}^{*} + \sum_{(i,t)\in\mathcal{N}_{i't'}^{*}} v_{i't'}^{it'}C_{i't'}^{*} + \sum_{(i,t)\in\mathcal{A}_{i't'}^{*}} v_{i't'}^{it'}C_{i't'}^{*}\right\}\right] \\
= & \sum_{i'=1}^{N}\sum_{t'=1}^{T}(2X_{i't'}C_{i't'}^{*} + 2(1-X_{i't'})C_{i't'}^{*}) = 2\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}^{*}. \qquad (31)
\end{aligned}
$$

Next, define $\delta_i = \alpha_i + \beta/2$ and $\xi = \beta/2$, and consider the following weighted least squares estimates,

$$
(\bar{\delta}, \bar{\gamma}, \bar{\xi}) \;=\; \underset{(\delta,\gamma,\xi)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(Y_{it} - \delta_i - \gamma_t - (2X_{it}-1)\xi)^2
$$

where $\delta = (\delta_1, \ldots, \delta_N)$. Clearly, $\bar{\alpha}_i^{M^*} = \bar{\delta}_i - \ddot{\beta}^{M^*}/2$ and $\ddot{\beta}^{M^*} = 2\bar{\xi}$ because (δ, ξ) is a linear one-to-one transformation of (α, β). Thus, using the above algebraic equalities, we can derive the desired result,

$$
\begin{aligned}
\ddot{\beta}^{M^*} &= \frac{2\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{W_{it}(2X_{it}-1)Y_{it} - W_{it}(2X_{it}-1)\bar{\delta}_i - W_{it}(2X_{it}-1)\bar{\gamma}_t\right\}}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}^{*}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} w_{it}^{i't'}\right)(2X_{it}-1)Y_{it} \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}^{*}} \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}^{i't'}(2X_{it}-1)Y_{it}\right) \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}^{*}} \sum_{i'=1}^{N}\sum_{t'=1}^{T}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{i't'}w_{it}^{i't'}(2X_{it}-1)Y_{it} + (1-X_{i't'})w_{it}^{i't'}(2X_{it}-1)Y_{it}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it}^{*}} \sum_{i'=1}^{N}\sum_{t'=1}^{T} \Bigg[X_{i't'} C_{i't'}^{*} \left\{ Y_{i't'} - Y_{it'-1} - \sum_{(i,t)\in\mathcal{N}_{i't'}^{*}} v_{i't'}^{it'} Y_{it} + \sum_{(i,t)\in\mathcal{A}_{i't'}^{*}} v_{i't'}^{it'} Y_{it} \right\} \\
&\qquad\qquad + (1 - X_{i't'}) C_{i't'}^{*} \left\{ Y_{it'-1} + \sum_{(i,t)\in\mathcal{N}_{i't'}^{*}} v_{i't'}^{it'} Y_{it} - \sum_{(i,t)\in\mathcal{A}_{i't'}^{*}} v_{i't'}^{it'} Y_{it} - Y_{i't'} \right\} \Bigg] \\
&= \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} C_{it} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} C_{it} D_{it} \left(\widehat{Y_{it}(1)} - \widehat{Y_{it}(0)} \right)
\end{aligned}
$$

where the second equality follows from equations (29), (30), and (31). □