Double-Robust Identification for Causal Panel Data Models*

Dmitry Arkhangelsky † Guido W. Imbens‡

September 2019

Abstract

We study identification and estimation of causal effects of a binary treatment in settings with panel data. We highlight that there are two paths to identification in the presence of unobserved confounders. First, the conventional path based on making assumptions on the relation between the potential outcomes and the unobserved confounders. Second, a design-based path where assumptions are made about the relation between the treatment assignment and the confounders. We introduce different sets of assumptions that follow the two paths, and develop double robust approaches to identification where we exploit both approaches, similar in spirit to the double robust approaches to estimation in the program evaluation literature.

Keywords: fixed effects, cross-section data, clustering, causal effects, treatment effects, unconfoundedness.

*This paper benefited greatly from our discussions with Manuel Arellano and David Hirshberg.
†Assistant Professor, CEMFI, darkhangel@cemfi.es.
‡Professor of Economics, Graduate School of Business and Department of Economics, Stanford University, SIEPR, and NBER, imbens@stanford.edu.

1 Introduction

Panel data are widely used to assess causal effects of policy interventions on economic outcomes. These data are particularly useful in settings where there is substantial heterogeneity both between units at the same point in time, as well as heterogeneity over time within units. Fundamentally the presence of panel data allows for two conceptually different comparisons to estimate causal effects. First, we can compare treated and control outcomes for the same unit at different points in time, that is, make across-time within-unit comparisons. Such comparisons are not possible in cross-section settings. Second, following approaches in cross-sectional settings, we can compare treated and control outcomes at the same point in time for different units, i.e., within-period across-unit comparisons. In that case, we use the panel data simply to allow for a richer set of controls than we would use in a cross-section setting. Different sets of assumptions justify the two approaches. In practice, researchers often make assumptions that simultaneously justify both types of comparisons. For example, many empirical papers use a linear two-way fixed effect specification that implicitly justifies both the within-unit and within-period comparisons:

$$Y_{it} = \alpha_i + \lambda_t + \tau W_{it} + \beta^\top X_{it} + \varepsilon_{it}. \tag{1.1}$$

Here W_{it} is an indicator for the treatment, with τ the causal effect of interest, and X_{it} are the time-unit specific control variables. In this specification, the α_i capture the permanent unit-specific effects, and the λ_t capture the common time effects. After removing the unit and time fixed effects, we can compare outcomes for treated units both to outcomes for the same unit in time periods where the unit was not treated, or to control units in the same time period.

In this paper, we take a different perspective, building on the program evaluation or causal inference literature. We start with the assumption that conditional on an unobserved unit-specific variable U_i (possibly vector-valued), the T-component vector of treatment assignments over time for unit i, $\underline{W}_i$, with t-th element equal to W_{it}, is independent of the vector of potential outcomes $\underline{Y}_i(\underline{w})$:

$$\underline{W}_i \perp\!\!\!\perp \left\{ \underline{Y}_i(\underline{w}) \right\}_{\underline{w}} \;\Big|\; U_i. \tag{1.2}$$

This assumption has no immediate content because we can make it hold by construction by setting U_i equal to the vector of assignments $\underline{W}_i$. Nevertheless, it clarifies what the issue is and why cross-section data alone are not sufficient: there is an unobserved variable U_i that invalidates comparisons of observed outcomes by treatment status because this unobserved variable is correlated both with the potential outcomes and with the treatment assignment. Although it is not always articulated in this form implicitly this conditional independence assumption is made in many of the approaches to identification in panel data settings used in the empirical literature.

For the case where (in contrast to the case we consider in the current paper) (1.2) holds with U_i observed, the program evaluation literature has developed a number of effective methods for estimating the average causal effect of W_{it} on Y_{it} (see Imbens [2004], Abadie and Cattaneo [2018] for reviews). One approach is to remove the association between U_i and the treatment W_{it} by using the propensity score either through weighting or through conditioning. Second, one can transform the outcome by removing the association between the outcome and U_i. This is typically done by subtracting from the outcome the conditional mean of the outcome Y_{it} given U_i. Third, and most effectively, one can use double robust methods and combine the propensity score adjustment and the outcome modeling/transformation. These methods inspire the proposals developed in the current paper for the case where U_i is not observed.

In the case where U_i is not observed one has to make additional assumptions to ensure point-identification. For the most part, applied researchers have been focusing on making assumptions regarding the relationship between the outcome and the unobserved characteristic. This approach is natural, often follows directly from an economic model, and is supported by the econometric theory (see, e.g., the surveys: Chamberlain [1984], Arellano and Honoré [2001], Arellano [2003], Arellano and Bonhomme [2011]). At the same time, such restrictions are very different from (1.2) because they are not motivated by a model of $\underline{W}_i$ (model of assignment). The point that we are making in this paper is that a model for $\underline{W}_i$ provides an alternative path to identification argument, and, moreover, it can be considered **separately** from the model for the outcome. We show that with panel data, one can base the **identification** argument on either the outcome model or the assignment model being correct. This is where our approach differs conceptually from the double robust estimation literature. Here both the design assumptions and the outcome modeling approaches are used in the identification stage.

First, analogous to the outcome modeling, we can use models and assumptions to motivate a transformation of the potential outcomes such that the unobserved component is independent of the transformed potential outcomes, and the transformed outcomes themselves are informative about the causal effect of interest. Formally,

$$U_i \perp\!\!\!\perp \; g\left(\left\{\underline{Y}_i(\underline{w})\right\}_{\underline{w}}\right), \tag{1.3}$$

for some function of the potential outcomes $g(\cdot)$, possibly after some conditioning. Many methods used in the empirical literature, including the two-way fixed effect estimator, can be thought of as fitting in this approach. For example, consider a two-period setting. The two-way fixed effect estimator transforms the outcomes by taking differences, *e.g.*, in the two period case $\Delta_i = g(\underline{Y}_i(\underline{w})) = Y_{i2}(\underline{w}) - Y_{i1}(\underline{w})$, so that Δ_i is free of dependence on the unobserved component U_i.

The second approach is design-based, where the goal is to find a set of conditioning variables S_i that removes the association between the treatment assignment and the unobserved component analogous to the propensity score approach.

$$U_i \;\perp\!\!\!\perp\; \underline{W}_i \;\Big|\; S_i. \tag{1.4}$$

A version of this assumption has been used in the panel literature before (e.g., the exchangeability assumption in Altonji and Matzkin [2005] or the exponential family assumption in Arkhangelsky and Imbens [2018]). In this paper, we argue that it holds for a variety of models that have been commonly used for binary data (e.g., Honoré and Kyriazidou [2000], Chamberlain [2010], Aguirregabiria et al. [2018]). In principle, the two-way fixed effect estimator can also be thought of as following this approach by comparing treated and control units at the same time within the set of units with the same fraction of treated periods, that is, conditioning on $S_i = \sum_{t=1}^{T} W_{it}$. However, as a general approach to identifying treatment effects in a panel data setting, this design-based approach that is common in the treatment effect literature has not been explored, and we do so in the current paper.

Third, we explore robust versions where we combine outcome modeling and assumptions on the assignment mechanism. Essentially there we develop models that justify (1.3) for some transformation, and models that justify (1.4) for some conditioning variables S_i, and then consider strategies that only require that the independence in (1.3) holds within subpopulations defined by S_i:

$$U_i \perp\!\!\!\perp \; g\left(\left\{\underline{Y}_i(\underline{w})\right\}_{\underline{w}}\right) \;\Big|\; S_i. \tag{1.5}$$

The paper fits in with the recent literature on causal inference in panel data settings, including the closely related synthetic control literature (Abadie et al. [2010], Arkhangelsky et al. [2019], Xu [2017], Ben-Michael et al. [2018]) difference in differences methods (de Chaisemartin and D'Haultfœuille [2018], Goodman-Bacon [2017], Athey and Imbens [2018], Athey et al. [2017]), and fixed effect methods (Imai and Kim [2019], Arkhangelsky and Imbens [2018]).

1.1 Notation

For $p \in [1, \infty]$ we use $L_p(\mathcal{P})$ to denote the space of all random variables X that satisfy $\mathbb{E}[\|X\|^p]^{\frac{1}{p}} < \infty$. For any two random variables $X_1, X_2 \in L_p(\mathcal{P})$ we use $\|X_1 - X_2\|_p$ to denote the $L_p(\mathcal{P})$ distance. For a random sample $\{X_i\}_{i=1}^n$ and any real-valued functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$ we define:

$$\begin{aligned} &\mathbb{P}_n f_1(X_i) := \frac{1}{N}\sum_{i=1}^{N} f_1(X_i) \\ &\|f_1 - f_2\|_{n,p} = \left(\mathbb{P}_n \left(f_1(X_i) - f_2(X_i)\right)^p\right)^{\frac{1}{p}} \end{aligned} \tag{1.6}$$

For a matrix A we use $\sigma_{\min}(A)$ to denote its smallest singular value.

2 Setup

We observe N units over T periods (i and t being a generic unit and period, respectively). We focus on settings with large N and fixed T. We are interested in the effect of a binary policy variable w on some economic outcome Y_{it}. To formalize this we consider a potential outcome framework (Imbens and Rubin [2015]). The policy can change over time, and so is indexed by unit i and time t, $W_{it} \in \{0, 1\}$. Let $\underline{w}^t \equiv (w_1, w_2, \ldots, w_t)$ denote the sequence of treatment exposures up to time t, with $\underline{w}$ as shorthand for the full vector of exposures $\underline{w}^T$. Define $\underline{W}_i \equiv (W_{i1}, \ldots, W_{iT})$ to be the full assignment vector for unit i. For the first part of the paper we assume that researchers do not observe additional unit-level covariates and explicitly introduce them in Section 4. In general, one can view all our identification results as conditional on covariates.

Let $Y_{it}(\underline{w}^t)$ denote the potential outcome for unit i at time t, given treatment history up to time t $\underline{w}^t$:

$$Y_{it}(\underline{w}^t) \equiv Y_{it}(w_1, w_2, \ldots, w_t). \tag{2.1}$$

In this paper we consider a static version of this general model.

Assumption 2.1. (No Dynamics) *For arbitrary $\underline{w}^t_{(1)}$ and $\underline{w}^t_{(2)}$ such that $w_{t1} = w_{t2}$ we have the following:*

$$Y_{it}(\underline{w}^t_{(1)}) = Y_{it}(\underline{w}^t_{(2)}) \tag{2.2}$$

This restriction implies that past treatment exposures do not affect contemporaneous outcomes. This assumption does not restrict time-series correlation in the realized outcomes and so on its own does not have any testable implications. However, given a particular assignment process, Assumption 2.1 can be tested. Since a substantial part of the empirical literature focuses on contemporaneous effects and assumes away dynamic effects, we view this as a natural starting point. The issues we raise are relevant for the dynamic treatment effect case as well but are discussed most easily in the static case.

Given the no-dynamics assumption we can index the potential outcomes by a single binary argument w, so we write $Y_{it}(w)$, for $w \in \{0, 1\}$. In this setup we can be interested in various treatment effects. Define individual and time-specific treatment effects:

$$\tau_{it} \equiv Y_{it}(1) - Y_{it}(0) \tag{2.3}$$

We focus primarily on average treatment effects, typically a convex combination of individual effects τ_{it}. Define also $\underline{Y}_i(\underline{w}) \equiv (Y_{i1}(w_1), \ldots, Y_{it}(w_T))$ to be the vector of potential outcomes.

We make two additional assumptions. First, we restrict out attention to settings with strictly exogenous covariates (e.g., Arellano [2003]) and make the following assumption:

Assumption 2.2. (Latent Unconfoundedness) *There exist a random element $U_i \in \mathbb{U}$ such* *that the following conditional independence holds:*

$$\underline{W}_i \perp\!\!\!\perp \left\{\underline{Y}_i(\underline{w})\right\}_{\underline{w}} \Big| \, U_i \tag{2.4}$$

This assumption effectively says that once we control for U_i, then all the differences in the treatment paths $\underline{W}_i$ across units are unrelated to the potential outcomes. This type of assignment should be contrasted with the sequential assignment where W_{it} can depend on past outcomes and latent characteristics. See Arellano [2003] for a discussion in the linear case. On its own Assumption 2.2 is not restrictive because we allow U_i to be unobserved: we can mechanically choose $U_i = \underline{W}_i$ so that this assumption is satisfied by construction. There are multiple papers that essentially follow this road, going back at least to Chamberlain [1992] (also see Chernozhukov et al. [2013] for a very general version of this approach).

We view U_i as a unit characteristic that we need to control for if we wish to compare outcomes across units. We formalize this by making the following assumption on the (infeasible) generalized propensity score (Imbens [2000]) that ensures that in principle such comparisons are possible.

Assumption 2.3. (Latent Overlap) *Define the infeasible generalized propensity score:*

$$r^{\text{inf}}(\underline{w}, u) \equiv \text{pr}(\underline{W}_i = \underline{w} | U_i = u). \tag{2.5}$$

For any $u \in \mathbb{U}$:

$$\max_{\underline{w}}\{r^{\text{inf}}(\underline{w}, u)\} < 1 \tag{2.6}$$

This assumption essentially says that in the population there exist units with the same U_i but different values of $\underline{W}_i$. This type of assumption is common in the (cross-section) program evaluation literature: without such an overlap assumption even if we observed U_i we would not be able to identify the average causal effect of the treatment without functional form restrictions. However, this latent overlap assumption is not always maintained in the panel literature. For example, if only time-series variation is used to make causal statements, then one does not need to make Assumption 2.3. Of course, this comes at a cost – one has to restrict the way potential outcomes can change over time. At the same time, if one also wants to exploit the cross-sectional variation, then some version of Assumption 2.3 appears to be unavoidable, but the outcome model can be more flexible compared to the approaches that rely on over-time comparisons.

3 Double Robustness Identification

3.1 Preliminaries

Before we consider identification in various models we need to define additional objects. Let $\boldsymbol{W}$ be the support of the vector of assignments $\underline{W}_i$; we can think of $\boldsymbol{W}$ as a matrix with at most 2^T rows and T columns, where each row is an element of the support of $\underline{W}_i$. Let $\boldsymbol{W}_k$ be a k row of the matrix $\boldsymbol{W}$ – a T-dimensional vector of zeros and ones. Let $\pi_k \equiv \text{pr}(\underline{W}_i = \boldsymbol{W}_k) = \mathbb{E}\left[\mathbf{1}_{\underline{W}_i = \boldsymbol{W}_k}\right]$. All π_k are positive, otherwise the corresponding row of $\boldsymbol{W}$ can be dropped. Let K be the number of rows in $\boldsymbol{W}$.

For example, if $T = 3$ then $\boldsymbol{W}$ can have the following form:

$$\boldsymbol{W} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{3.1}$$

Each row of this matrix represents a possible assignment, and in this particular case only 4 out of the $2^3 = 8$ possible combinations have positive probability. For a particular unit i, let $k(i)$ be the row $\boldsymbol{W}_k$ of $\boldsymbol{W}$ such that $\boldsymbol{W}_k = \underline{W}_i$. For the identification argument we assume we know $\boldsymbol{W}$ and the probabilities π_k and consider estimation in Section 4.

We are interested in estimating weighted averages of the treatment effects τ_{it}. Our estimators will be linear in $\boldsymbol{Y}$, with weights that depend on $\underline{W}_i$:

$$\hat{\tau} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{it}Y_{it}.$$

Choosing an estimator therefore corresponds to choosing a set of weights ω_{it}. We maintain throughout this section the no-dynamics assumption (Assumption 2.1), latent unconfoundedness assumption (Assumption 2.2), and latent overlap (Assumption 2.3).

3.2 Three Estimation Strategies with Observed Confounders

As discussed briefly in the introduction, the latent unconfoundedness assumptiwon can be exploited in two directions. To build intuition, it is useful to briefly make an analogy to the conventional unconfoundedness case with observed confounders, in a cross-section setting.

Suppose we have unconfoundedness (Rosenbaum and Rubin [1983]) with an observed confounder X_i. Here we use its weak form (Imbens [2000]):

$$\mathbf{1}_{W_i=w} \perp\!\!\!\perp \; Y_i(w) \; \Big| \; X_i, \qquad \forall w. \tag{3.2}$$

In that case researchers have followed two approaches. One is to exploit the propensity score result that (irrespective of whether (3.2) holds),

$$\mathbf{1}_{W_i=w} \perp\!\!\!\perp \; X_i \; \Big| \; \mathrm{pr}(W_i = w|X_i), \tag{3.3}$$

where $\mathrm{pr}(W_i = w|X_i)$ is the generalized propensity score. (3.2) and (3.3) combined imply that conditional on the generalized propensity score we have

$$\mathbf{1}_{W_i=w} \perp\!\!\!\perp \; Y_i(w) \; \Big| \; \mathrm{pr}(W_i = w|X_i). \tag{3.4}$$

Thus, we can conditioning on a variable, here $\mathrm{pr}(W_i = w|X_i)$ such that the association of the treatment indicator, here $\mathbf{1}_{W_i=w}$ and the variable we originally need to condition on, here X_i, vanishes.

A second approach is to transform the potential outcomes. Define the conditional expectations $\mu(w, x) \equiv \mathbb{E}[Y_i(w)|X_i = x)$ and $e(X_i) \equiv \mathrm{pr}(W_i = w|X_i)$. We do not actually need the full independence assumption in (3.2), only the mean-independence since it implies, $\mathbb{E}[Y_i(w)|\mathbf{1}_{W_i=w}, X_i] = \mathbb{E}[Y_i(w)|X_i]$. Now define

$$\tilde{Y}_i(w) \equiv g(Y_i(w)) \equiv Y_i(w) - \mu(w, x) - \frac{\mathbb{E}[e(X_i)]^{W_i}(1 - \mathbb{E}[e(X_i)])^{1-W_i}}{e(X_i)^{W_i}(1 - e(X_i))^{1-W_i}} \Big(\mu(1, X_i) - \mu(0, X_i)\Big).$$

This transformation of the potential outcomes does not change mean-indepence of $\tilde{Y}_i(w)$ and

$\mathbf{1}_{W_i=w}$ conditional on X_i, and we have

$$\mathbb{E}[\tilde{Y}_i(w))|W_i, X_i] = \mathbb{E}[\tilde{Y}_i(w)|X_i].$$

Hwever, for this transformed outcome we have something much stronger. Here we do not need the conditioning on X_i to have the result that the expected value is free of dependence on W_i, and mean-independence holds without conditioning on X_i:

$$\mathbb{E}[\tilde{Y}_i(w))|W_i = 1] = \mathbb{E}[\tilde{Y}_i(w)|W_i = 0] = \mathbb{E}[\tilde{Y}_i(w)] = \mathbb{E}[Y_i(1) - Y_i(0)].$$

We can combine these two approaches and estimate

$$\mathbb{E}[\tilde{Y}_i(1)|W_i = 1, e(X_i)], \qquad \text{and} \;\; \mathbb{E}[\tilde{Y}_i(0)|W_i = 0, e(X_i)],$$

and average the difference over the marginal distribution of $e(X_i)$. This will have double robustness properties.

The first insight that we take to the panel data case is that we can either use the conditional distribution of the assignment given the confounder to remove biases associated with a direct comparison of treated and control units, or we can remove the dependence of the outcomes on the confounder. This general strategy works whether the confounder is observed or not, but implementing the two approaches is a bigger challenge if the confounder is not observed, and we need to make additional assumptions in order to do so. The second insight is that combining these two approaches may lead to more robust estimates of the treatment effects.

3.3 Double Robust Identification – An Example

In this section we consider a simple example that illustrates the main message of the paper. For simplicity we start assuming that $\tau_{it} = \tau$ – constant treatment effects – and no covariates X_i. At the end of the section we discuss heterogenity in treatment effects. We introduce covariates in Section 4.

Consider the case with three periods and suppose that the distribution of $\underline{W}_i$ is given by Table 1. A researher wants to use a standard fixed effects model and runs the following regression

Table 1: Assignment process and weights

$\mathbb{P}(\underline{W}_i)$	W_1	W_2	W_3	$\omega_1^{(fe)}(\underline{W}_i)$	$\omega_2^{(fe)}(\underline{W}_i)$	$\omega_3^{(fe)}(\underline{W}_i)$
0.09	0	0	0	0.46	-0.64	0.18
0.04	1	0	0	5.70	-3.26	-2.44
0.11	0	1	0	-2.16	4.60	-2.44
0.14	1	1	0	3.08	1.98	-5.07
0.07	0	0	1	-2.16	-3.26	5.42
0.08	1	0	1	3.08	-5.88	2.80
0.15	0	1	1	-4.78	1.98	2.80
0.32	1	1	1	0.46	-0.64	0.18

(in population):

$$\begin{aligned} &Y_{it} = \alpha_i + \lambda_t + \tau^{fe} W_{it} + \varepsilon_{it} \\ &\mathbb{E}[\varepsilon_{it} | \underline{W}_i, \alpha_i] = 0 \end{aligned} \tag{3.5}$$

Usual OLS logic implies that τ^{fe} has the following representation:

$$\tau^{fe} = \mathbb{E}[Y_{it} \omega_t^{(fe)}(\underline{W}_i)] \tag{3.6}$$

where $\omega_t^{(fe)}(\underline{W}_i)$ are **fixed effects weights** that depend **only** on the distribution of $\underline{W}_i$. For the distribution given above the weights are presented in Table 1. By construction these weights sum up to 0 for every row and every column (once reweighted by the probabilities). If the two-way model is correctly specified than the estimator based on a sample analog of these weights has excellent statistical properties (see e.g., Donoho et al. [1994], Armstrong and Kolesár [2018b], and references therein). At the same time, such estimator is not entirely satisfactory. In particular, assume that the assignment is random conditional on $\overline{W}_i \equiv \frac{1}{T} \sum_{t=1}^{T} W_{it}$:

$$\underline{W}_i \perp\!\!\!\perp \left\{ \underline{Y}_i(\underline{w}) \right\}_{\underline{w}} \Big| \, \overline{W}_i \tag{3.7}$$

In this case, the relevant outcome model has the following structure:

$$\begin{aligned} &Y_{it} = h_t(\overline{W}_i) + \tau W_{it} + \xi_{it} \\ &\mathbb{E}[\xi_{it} | \underline{W}_i] = 0 \end{aligned} \tag{3.8}$$

The estimator based on the fixed effect weights is consistent if the following condition is satisfied for every t and $\overline{W}_i$:

$$\mathbb{E}[\omega_t^{(fe)}(\underline{W}_i)|\overline{W}_i] = 0 \tag{3.9}$$

Table 2 shows that this is not true for the given distribution of $\underline{W}_i$. As a result, if the outcome

Table 2: Aggregated weights

$\overline{W}_i$	$\mathbb{E}[\omega_1^{(fe)}(\underline{W}_i)\|\overline{W}_i]$	$\mathbb{E}[\omega_2^{(fe)}(\underline{W}_i)\|\overline{W}_i]$	$\mathbb{E}[\omega_3^{(fe)}(\underline{W}_i)\|\overline{W}_i]$
0	0.46	-0.64	0.18
1	-0.73	0.60	0.13
2	-0.08	0.36	-0.28
3	0.46	-0.64	0.18

model is given by (3.8) then the fixed effect weights will give us an inconsistent estimator. This is not surprising because $\omega_t^{(fe)}(\underline{W}_i)$ are not constructed to deal with such outcome models.

At this point, it is natural to ask whether we can achieve both goals simultaneously, i.e., can we find the weights that "work" if either the fixed effect model (3.5) or the design process (3.7) is correctly specified? The answer is positive and the weights that satisfy this restriction are given in Table 3. It is evident that the weights some up to zero for each row and simple

Table 3: Doubly robust weights

$\omega_1^{(dr)}(\underline{W}_i)$	$\omega_2^{(dr)}(\underline{W}_i)$	$\omega_3^{(dr)}(\underline{W}_i)$
0.00	0.00	0.00
6.59	-3.95	-2.64
-1.46	4.10	-2.64
3.24	1.66	-4.90
-1.46	-3.95	5.42
3.24	-6.39	3.15
-4.81	1.66	3.15
0.00	0.00	0.00

calculation shows that $\mathbb{E}[\omega_t^{(dr)}(\underline{W}_i)|\overline{W}_i] = 0$ for every t and $\overline{W}_i$. As a result, there is no trade-off in terms of identification and we can construct the estimator that works for both models.

So far we have assumed that the treatment effects are constant. This assumption is very strong and it is well documented that two-way estimators have problems in case with heterogenous treatment effects (e.g., see de Chaisemartin and D'Haultfœuille [2018]). This is evident after looking at Table 1: in the last row we assign negative weight to treated units in the second period. In contrast to this, all treated units receive non-negative weight when we use doubly robust weights from Table 3. This is not a coincidence and below we discuss a procedure that guarantees that this property is satisfied.

3.4 Identification Through the Outcome Model

First we consider outcome models. Recall that by the no-dynamics assumption the potential outcomes $Y_{it}(w)$ are indexed by a binary treatment w. A common outcome model that goes back at least to Chamberlain [1992] is the following one:

Assumption 3.1. *The potential outcomes satisfy:*

$$\mathbb{E}[Y_{it}(w)|U_i] = \alpha(U_i) + \lambda_t + \tau(U_i)w. \tag{3.10}$$

Given Assumption 2.2 the content of this model is that it restricts the time-dependency of the conditional mean of the control outcome and the treatment effect. Rewriting the model we can see that more directly. The conditional control mean is

$$\mathbb{E}[Y_{it}(0)|U_i] = \alpha(U_i) + \lambda_t,$$

which is restricted to be additively separable in time, and the conditional treatment effect is

$$\mathbb{E}[\tau_{it}|U_i] = \tau(U_i),$$

which is restricted to be time-invariant.

We are interested in identifying a convex combination of the heterogenous treatment effects $\tau(U_i)$ (which itself is a convex combination of τ_{it}) in this model. We do this by using the weights ω_{kt} that satisfy the following restrictions:

$$\begin{aligned} &\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\pi_k\omega_{kt}\boldsymbol{W}_{kt} = 1, \quad &&\forall k, \frac{1}{T}\sum_{t}\omega_{kt}\boldsymbol{W}_{kt} \geq 0 \\ &\forall k, \frac{1}{T}\sum_{t=1}^{T}\omega_{kt} = 0, &&\forall t, \sum_{k=1}^{K}\pi_k\omega_{kt} = 0 \end{aligned} \tag{3.11}$$

Let $\mathbb{W}_{\text{outc}}$ be the set of weights $\{\omega_{tk}\}_{t,k}$ that satisfy these restrictions. We can evaluate these restrictions and thus we can construct this set. For any generic element $\omega \in \mathbb{W}_{\text{outc}}$ define the random variables $\omega_{k_{(i)}t}$:

$$\omega_{k_{(i)}t} \equiv \sum_{k=1}^{K} \omega_{kt}\{\underline{W}_i = \boldsymbol{W}_k\} \tag{3.12}$$

Using these stochastic weights we can compute the following expectation:

$$\tau(\omega) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\omega_{k_{(i)}t}\right] \tag{3.13}$$

Proposition 1. *Suppose Assumptions 2.1, 2.2, and 3.1 hold, and that* $\omega \in \mathbb{W}_{\text{outc}}$*. Then* $\tau(\omega)$ *is a convex combination of* $\tau(U_i)$*.*

As a result, a certain convex combination of $\tau(U_i)$ can be identified whenever $\mathbb{W}_{\text{outc}}$ is non-empty. A natural question when this is the case. The answer is quite simple: the matrix $\boldsymbol{W}$ should contain at least one of the following three submatrices (up to permutations):

$$\boldsymbol{W}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{W}_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{W}_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.14}$$

Consider each of these three cases separately. In the first case there are adoptors of the treatment with $(W_{it} = 0, W_{it+1} = 1)$ and in the same periods t and $t+1$ non-adoptors with $(W_{it} = 0, W_{it+1} = 0)$. In the second case there are adoptors of the treatment with $(W_{it} = 0, W_{it+1} = 1)$ and in the same periods t and $t+1$ units who have already adopted and keep the treatment, with $(W_{it} = 1, W_{it+1} = 1)$. In the last case there are adopters with $(W_{it} = 0, W_{it+1} = 1)$ and units who switch out with $(W_{it} = 1, W_{it+1} = 0)$. To put this discussion in perspective, it is not sufficient to have assignment matrices of the type

$$\boldsymbol{W}_4 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{W}_5 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

where with the first design some units are always in the control group and all others are always in the treatment group, and where with the second design all units adopt the treatment at exactly the same time.

3.5 Identification Through Design

In this section we consider assignment processes that satisfy a certain sufficiency property. We state it as a high-level assumption and then show examples of economic models that satisfy this assumption:

Assumption 3.2. (SUFFICIENCY) *There exist a known $\underline{W}_i$-measurable sufficient statistic $S_i \in \mathbb{S}$ and a subset $\mathbb{A} \subset \mathbb{S}$ such that: (i)*

$$\underline{W}_i \perp\!\!\!\perp U_i \;\Big|\; S_i, \tag{3.15}$$

and (ii), for all $s \in \mathbb{A}$:

$$\max_{\underline{w}}\{r(\underline{w}, s)\} < 1. \tag{3.16}$$

where $r(\underline{w}, s)$ is the feasible generalized propensity score:

$$r(\underline{w}, s) \equiv \mathrm{pr}(\underline{W}_i = \underline{w} | S_i = s). \tag{3.17}$$

This assumption might look restrictive, but an S_i such that conditional on S_i the treatment $\underline{W}_i$ and the unboserved variable U_i are independent always exists, namely $S_i^{\mathrm{gen}} \equiv f_{U|\underline{W}}(\cdot|\underline{W}_i)$, where $f_{U|\underline{W}}(x|y)$ is the conditional distribution of U_i given $\underline{W}_i$. In general, S_i^{gen} is an infinite-dimensional object (a function) and is unknown, because $f_{U|\underline{W}}(x|y)$ is unknown. As a result, the first restriction that we make in Assumption 3.2 is that S_i is known. Part (ii) does not allow for $S_i = S_i^{\mathrm{gen}}$ because we require $\underline{W}_i$ to have a non-generate distribution given S_i. Below we consider various assignment models that are common in the empirical panel data literature and demonstrate that in all of them there exist S_i that one can easily compute.

The main implication of the Assumption 3.2 coupled with Assumption 2.2 is summarized in the following proposition:

Proposition 2. *Suppose Assumptions 2.1, 2.2, and 3.2 hold. Then for any $\underline{w}$:*

$$\mathbf{1}_{\underline{W}_i=\underline{w}} \perp\!\!\!\perp \underline{Y}_i(\underline{w}) \,\Big|\, S_i. \tag{3.18}$$

This proposition demonstrates that unconfoundedness conditional on U_i can be transformed into undonfoundedness conditional on S_i under the additional assumption that restricts the assignment process.

The assignment models that we consider in this section are restrictive, in a sense that they must satisfy Assumption 3.2. At the same time, most of the models for the binary time-series process W_{it} that are used in the applied and theoretical literature actually satisfy these restrictions (see, e.g., Honoré and Kyriazidou [2000], Chamberlain [2010], Aguirregabiria et al. [2018]). In fact, in certain cases existence of a sufficient statistic is a necessary requirement for estimation of common parameters (e.g., Magnac [2004]). This is especially relevant, because many of such models have an underlying economic intuition and can be interpreted as models of optimal choice.

We are not interested in estimating common parameters of the model for $\underline{W}_i$, which is the standard object in non-linear panel analysis. Instead, we only require that the conditional distribution of $\underline{W}_i$ admits a certain representation. Parameters of this representation are not identified with fixed T, but they do not play any role in Proposition 2, which is the only result that we need.

Static model. As a first example that we consider a static logit model with heterogeneity over time. Formally, we consider the following model:

$$\begin{aligned} &\mathbb{E}[W_{it}|U_i] = \frac{\exp(\alpha^T(U_i)\psi(t)+\lambda_t)}{1+\exp(\alpha^T(U_i)\psi(t)+\lambda_t)} \\ &W_{it} \perp\!\!\!\perp \{W_{il}\}_{l\neq t} \,\Big|\, U_i \end{aligned} \tag{3.19}$$

where $\psi(t)$ is a known function of t. It is easy to demonstrate that in this model

$$S_i = \sum_{t=1}^{T} \psi(t) W_{it}/T.$$

This model is a generalization of the standard fixed-effects logit model analyzed in Chamberlain [2010]. □

Dynamic model. Next we consider a time homogenous Markov model:

$$\mathbb{E}[W_{it}|U_i, W_i^{t-1}] = \frac{\exp(\alpha(U_i) + \gamma(U_i)W_{it-1})}{1 + \exp(\alpha(U_i) + \gamma(U_i)W_{it-1})}$$
$$W_{it} \perp\!\!\!\perp \{W_{il}\}_{l>t} \Big| U_i, W_i^{t-1} \tag{3.20}$$

In this model

$$S_i = \left(\sum_{t=2}^{T-1} W_{it}, \sum_{t=2}^{T} W_{it}W_{it-1}, W_{i1}, W_{iT}\right).$$

□

General case For sufficiency we need the following representation for the conditional distribution of $\underline{W}_i$:

$$\log\left(\mathbb{P}(\underline{W}_i|U_i)\right) = S(\underline{W}_i)^\top \alpha(U_i) + \beta(U_i) + \gamma(\underline{W}_i) \tag{3.21}$$

where $S(\cdot)$ is a known function of $\underline{W}_i$. All previous examples have this representation. More generally, Aguirregabiria et al. [2018] show that this structure arises in flexible models of dynamic choice. □

Let S_i be a potential sufficient statistics. Let $\boldsymbol{W}^s$ be a matrix representation of the support of $\underline{W}_i$ conditional on $S_i = s$ and $\boldsymbol{W}_k^s$ be a generic row (element of the support). For example, if $S_i = \sum_t W_{it}$ and $\boldsymbol{W}$ is given by (3.1) then S_i takes 3 possible values and we have the following:

$$
\begin{aligned}
\boldsymbol{W}^0 &= \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \\
\boldsymbol{W}^2 &= \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \\
\boldsymbol{W}^3 &= \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}
\end{aligned} \tag{3.22}
$$

When considering identification strategy based on design assumptions we do not restrict potential outcomes, but instead require that assumptions behind Proposition 2 are satisfied. In this case, one can identify a convex combination of individual treatment effects using the weights that satisfy the following restrictions (for all k, s and t):

$$
\begin{aligned}
&\frac{1}{T}\sum_{tk} \pi_k \omega_{kt} \boldsymbol{W}_{kt} = 1 \\
&\sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \omega_{kt} \boldsymbol{W}_{kt} \geq 0 \\
&\sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \omega_{kt} = 0
\end{aligned} \tag{3.23}
$$

Let $\mathbb{W}_{\text{design}}$ be the set of weights $\{\omega_{tk}\}_{t,k}$ that satisfy these restrictions. It is easy to see that $\mathbb{W}_{\text{design}}$ is nonempty whenever there exists at least one s such that $\mathbb{W}_s$ contains at least two rows. This is guaranteed by the second part of Assumption 3.2. For any $\omega \in \mathbb{W}_{\text{design}}$ define the random variables $\omega_{k_{(i)}t}$ in the same way as before and consider the following expectation:

$$
\tau(\omega) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\omega_{k_{(i)}t}\right] \tag{3.24}
$$

Proposition 3. *Suppose Assumptions 2.1, 2.2, 2.3, and 3.2 hold, and that* $\omega \in \mathbb{W}_{\text{design}}$. *Then* $\tau(\omega)$ *is a convex combination of treatment effects.*

3.6 Double robustness

The sets of $\mathbb{W}_{\text{outc}}$ and $\mathbb{W}_{\text{design}}$ are motivated by different models and in general do not need to be similar. In some sense, one can say that the weights in $\mathbb{W}_{\text{outc}}$ target within-unit comparisons, while those in $\mathbb{W}_2$ target within-period comparisons. This interpretation is convenient, but is not entirely correct because in general $\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}}$ is not empty. Consequently, one does not need to take a stand on what comparisons to use: those based on looking at the same units across time or at different units for a fixed time period. As a result, we suggest using the weights in $\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}}$. In fact, we restrict this set even further and define the following one:

$$
\begin{aligned}
&\mathbb{W}_{\text{dr}} \equiv \{\omega\} \\
\text{subject to: } & \frac{1}{T|W|}\sum_{tk} \pi_k \omega_{kt} \boldsymbol{W}_{kt} = 1, \quad \frac{1}{T}\sum_{t=1}^{T} \omega_{kt} = 0 \\
& \sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \omega_{kt} = 0, \qquad \omega_{kt}\boldsymbol{W}_{kt} \geq 0
\end{aligned}
\tag{3.25}
$$

Denote this set by $\mathbb{W}_{\text{dr}}$, and note that $\mathbb{W}_{\text{dr}} \subset (\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}})$. The difference between $\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}}$ and $\mathbb{W}_{\text{dr}}$ is quite small – we simply impose the additional restriction that every treated unit receives a non-negative weight. Note that neither weights in $\mathbb{W}_{\text{outc}}$ nor in $\mathbb{W}_{\text{design}}$ in general satisfy this restriction. This is important in practice, because we want to be robust to arbitrary heterogeneity in treatment effects.

When is the set $\mathbb{W}_{\text{dr}}$ non-empty? Combining earlier discussion of $\mathbb{W}_{\text{outc}}$ and $\mathbb{W}_{\text{design}}$ it is easy to see that a necessary and sufficient condition for $\mathbb{W}_{\text{dr}}$ to be non-empty is that there exists an s such that the corresponding $\boldsymbol{W}^s$ contains at least one of the following two sub-matrices (up to permutations):

$$
\boldsymbol{W}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \boldsymbol{W}_3 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}
\tag{3.26}
$$

In particular, note that the matrix $\boldsymbol{W}_2$ from (3.14) is not sufficient. The reason for this is that we require the weights for treated units to be non-negative and sum up to zero for each row. This implies that the first row should receive a zero weight and thus we cannot make cross-sectional comparisons. The requirement for $\boldsymbol{W}^s$ to contain these sub-matrices is in general more demanding than the second part of Assumption 3.2. At the same time, if S_i includes $\overline{W}_i$ then for any s, $\boldsymbol{W}^s$ can contain $\boldsymbol{W}_3$ only if it contains $\boldsymbol{W}_1$ and this is equivalent to the overlap condition.

Finally we can state the main identification result. The following theorem is a direct consequence of Propositions 1 and 3:

Theorem 1. *Suppose Assumptions 2.1, 2.2, and 2.3 hold, and either 3.1, or Assumption 3.2, or both hold. Then for any $\omega \in \mathbb{W}_{\mathrm{dr}}$, the estimand $\tau(\omega)$ is a convex combination of treatment effects.*

4 Estimation and inference

4.1 Statistical framework

We assume that we observe a random sample $\{\underline{Y}_i, \underline{W}_i, X_i\}_{i=1}^N$ from some distribution $\mathcal{P}$ with T (number of periods) being fixed. We assume that a researcher has constructed sufficient statistics $S_i \equiv S(\underline{W}_i, X_i)$ based on a design model. We maintain Assumption 2.1 and additionally restrict the outcome model:

Assumption 4.1. *For each t one of the following outcome models is correct. Either there exist a sufficient statistic S_i such that the following is true:*

$$\begin{aligned} &Y_{it}(0) = \beta_t + \psi_0(X_i, t)^\top \delta + \psi_1(X_i, S_i, t)^\top \gamma + \xi_{it} \\ &\mathbb{E}[\xi_{it} | X_i, S_i] = 0 \\ &(\xi_{i1}, \dots \xi_{iT}) \perp\!\!\!\perp \underline{W}_i | X_i, S_i \end{aligned} \tag{4.1}$$

or $U_i = (\underline{W}_i, X_i)$ and we have the following:

$$\begin{aligned} &Y_{it}(0) = \alpha(\underline{W}_i, X_i) + \beta_t + \psi_0(X_i, t)^\top \delta + \varepsilon_{it} \\ &\mathbb{E}[\varepsilon_{it} | X_i, \underline{W}_i] = 0 \end{aligned} \tag{4.2}$$

where $\psi_0(X_i, t)$ and $\psi_0(X_i, S_i, t)$ are known p-dimensional functions.

This assumption allows for our design model to be correct, so that we only need to control for (S_i, X_i), or the more traditional fixed effects model to be correct. We do not impose any restrictions on $Y_{it}(1)$ and thus on heterogeneity in treatment effects. For simplicity we assume that in both cases the conditional expectations are linear in parameters with respect to a known finite-dimensional dictionary. Since all our identification results hold conditional on X_i this assumption is not necessary and the estimation procedure below can be adopted to allow for unknown ψ_0 and ψ_1. At the same time, we believe that our estimator is a natural alternative for the current status quo which is a two-way fixed effect model estimated by OLS which is based on (4.2). We leave further nonparametric generalizations to future work.

4.2 Estimator

Our estimator is defined in the following way:

$$\hat{\tau} := \frac{1}{NT} \sum_{it} \hat{\omega}_{it} Y_{it} \tag{4.3}$$

where the weights $\{\hat{\omega}_{it}\}_{it}$ solve the optimization problem:

$$\begin{aligned}
\{\hat{\omega}_{it}\}_{it} = \arg\min_{\{\omega_{it}\}_{it}} & \frac{1}{(NT)^2} \sum_{it} \omega_{it}^2 \\
\text{subject to: } & \frac{1}{nT} \sum_{it} \omega_{it} W_{it} \geq 1 \\
& \frac{1}{T} \sum_{i} \omega_{it} = 0 \\
& \frac{1}{N} \sum_{t} \omega_{it} = 0 \\
& \frac{1}{NT} \sum_{it} \omega_{it} \psi(X_i, S_i, t) = 0 \\
& \omega_{it} W_{it} \geq 0
\end{aligned} \tag{4.4}$$

where we define $\psi(X_i, S_i, t) := (\psi_0(X_i, t), \psi_1(X_i, S_i, t))$. At the optimum the first inequality is binding and we write it down in this form to simplify the dual representation below. The weights $\hat{\omega}_{it}$ are related to standard OLS fixed effects weights, but here we are explicitly looking for weights that balance out functions of S_i, not only fixed attributes X_i, and satisfy certain inequality constraints. The last restriction is crucial, because it is well documented that the standard OLS estimators with fixed effects in general do not correspond to reasonable estimands if the effects are heterogeneous (see e.g., de Chaisemartin and D'Haultfœuille [2018]).

It is natural to ask if the weights that solve the problem above exist. In Lemma A.1 we show that a necessary and sufficient condition for the existence is that the control and treated units satisfy a certain overlap condition. In particular, there is **no** $\{\lambda_i, \mu_t, \gamma\}_{i,t}$ such that the following is true:

$$\begin{aligned} &\lambda_i + \mu_t + \psi_{it}^\top \gamma \geq 0 \\ &W_{it} = \{\lambda_i + \mu_t + \psi_{it}^\top \gamma > 0\} \end{aligned} \tag{4.5}$$

This is a very mild overlap condition that is likely to be satisfied for any reasonable assignment process.

Our estimator fits naturally into recent theoretical literature on balancing weights (e.g., Imai and Ratkovic [2014], Zubizarreta [2015], Athey et al. [2016], Hirshberg and Wager [2017], Chernozhukov et al. [2018a,b], Armstrong and Kolesár [2018a]). The main technical difference between our approach and the ones proposed in the literature is that we need to balance unit-specific functions and explicitly impose non-negativity constraints. At the same time, we only balance a small parametric class of functions of (X_i, S_i), while others consider much more general functional classes. We leave this generalization to future research.

4.3 Dual representation

The Lagrangian saddle-point problem for the program (4.4) has the following form:

$$\begin{aligned} &\inf_{\omega_{it}} \sup_{\lambda_{(t)}, \lambda_{(i)}, \gamma, \mu_{it} \geq 0, \pi \geq 0} \frac{1}{(NT)^2} \sum_{it} \omega_{it}^2 + \frac{1}{N} \sum_i \lambda_{(i)} \left(\frac{1}{T} \sum_i \omega_{it} \right) + \\ &\frac{1}{T} \sum_t \lambda_{(t)} \left(\frac{1}{N} \sum_t \omega_{it} \right) + \pi \left(1 - \frac{1}{NT} \sum_{it} \omega_{it} W_{it} \right) - \\ &\qquad \gamma^\top \left(\frac{1}{NT} \sum_{it} \omega_{it} \psi_{it} \right) - \frac{1}{NT} \sum_{it} \mu_{it} \omega_{it} W_{it} \end{aligned} \tag{4.6}$$

where we use ψ_{it} as a shorthand for $\psi(X_i, S_i, t)$. In Lemma A.1 we show that strong duality holds and we can rearrange the minimization and maximization:

$$\sup_{\lambda_{(t)},\lambda_{(i)},\gamma,\mu_{it}\geq 0,\pi\geq 0} \inf_{\omega_{it}} \frac{1}{(NT)^2}\sum_{it}\omega_{it}^2 + \frac{1}{N}\sum_{i}\lambda_{(i)}\left(\frac{1}{T}\sum_{i}\omega_{it}\right) + \frac{1}{T}\sum_{t}\lambda_{(t)}\left(\frac{1}{N}\sum_{t}\omega_{it}\right) - \pi\left(\frac{1}{NT}\sum_{it}\omega_{it}W_{it} - 1\right) - \gamma^\top\left(\frac{1}{NT}\sum_{it}\omega_{it}\psi_{it}\right) - \frac{1}{NT}\sum_{it}(\mu_{it}\omega_{it}W_{it}) \quad (4.7)$$

Solving this in terms of ω_{it} (an unconstrained quadratic problem) we get the following representation:

$$\inf_{\lambda_{(t)},\lambda_{(i)},\gamma,\mu_{it}\geq 0,\pi\geq 0} \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\left(\pi W_{it} - \lambda_{(t)} - \lambda_{(i)} - \gamma^\top\psi_{it} - \mu_{it}W_{it}\right)^2\right] - \frac{4\pi}{N} \quad (4.8)$$

We can further simplify this expression by concentrating out μ_{it} and π. To this end, define the following loss function:

$$\rho_z(x) := x^2(1-z) + x_+^2 z \quad (4.9)$$

After some algebra we get the following:

$$\inf_{\lambda_{(t)},\lambda_{(i)},\gamma} \mathbb{P}_n\left(\frac{1}{T}\sum_{t=1}^{T}\rho_{W_{it}}\left(W_{it} - \lambda_{(t)} - \lambda_{(i)} - \gamma^\top\psi_{it}\right)\right) \quad (4.10)$$

Let $\{\hat{\lambda}_{(t)}, \hat{\lambda}_{(i)}, \hat{\gamma}\}_{i,t}$ be the solutions to this problem. The optimal unnormalized weights are equal to the following:

$$\hat{\omega}_{it}^{(un)} = \left(W_{it} - \hat{\lambda}_{(t)} - \hat{\lambda}_{(i)} - \hat{\gamma}^\top\psi_{it}\right)(1 - W_{it}) + \left(W_{it} - \hat{\lambda}_{(t)} - \hat{\lambda}_{(i)} - \hat{\gamma}^\top\psi_{it}\right)_+ W_{it} \quad (4.11)$$

and the optimal weights are given by the normalization:

$$\hat{\omega}_{it} := \frac{\hat{\omega}_{it}^{(un)}}{\frac{1}{NT}\sum_{it}\hat{\omega}_{it}^{(un)}W_{it}} \tag{4.12}$$

By construction the weights are non-negative for the treated units and sum up to one once multiplied by W_{it}. The denominator is strictly positive under the conditions of Lemma A.1.

4.4 Inference

In order to state the inference results we need to make several statistical assumptions:

Assumption 4.2. *(a) $\mathcal{P}$-a.s. $(X_i, S_i) \in \Omega$ – compact subset of some metric space; (b) $\psi(X_i, S_i, t)$ is a continuous function of its arguments (on Ω); errors u_{it} satisfy the following moment conditions:*

$$\begin{aligned} &\mathbb{E}[u_{it}^2|\underline{W}_i, X_i] \le \overline{\sigma}_u^2 < \infty \\ &\mathbb{E}[u_{it}^4] < \infty \end{aligned} \tag{4.13}$$

Part of the assumption about u_{it} is standard in the literature on projection estimators. We assume compactness to streamline the proofs and we think that it covers most problems that researchers face in applications. There is no doubt that it can be considerably relaxed.

Assumption 4.3. *(a) S_i includes $\overline{W}_i$; (b) for all t and $\eta > 0$ we have $\mathbb{E}[W_{it}|S_i, X_i] \le 1-\eta$; (c) the following holds:*

$$\begin{aligned} &\Gamma_{it} := (1-W_{it})\psi_{it} - \frac{\sum_{l=1}^T (1-W_{il})\psi_{il}}{\sum_{l=1}^T (1-W_{il})} \\ &\sigma_{\min}\left(\sum_{t=1}^T \mathbb{E}\left[\Gamma_{it}\Gamma_{it}^\top\right]\right) \ge \kappa > 0 \end{aligned} \tag{4.14}$$

Next theorem states properties of $\hat{\tau}$ and $\hat{\omega}_{it}$:

Theorem 2. *Suppose Assumptions 4.1, 4.2, 4.3 are satisfied. Then there exist a collection of* *random variables* $\{\omega^\star(X_i, \underline{W}_i, t)\}_{t=1}^T$ *such that the following holds:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{\omega}_t - \omega_t^\star\|_2 = o_p(1) \tag{4.15}$$

Define the following conditional estimand:

$$\tau_{emp} = \frac{1}{NT}\sum_{it} \hat{\omega}_{it} W_{it} \mathbb{E}[\tau_{it}|\underline{W}_i, X_i] \tag{4.16}$$

the scaled difference between the estimator and τ_{emp} *converges in distribution to a normal random variable:*

$$\sqrt{n}\left(\hat{\tau} - \tau_{emp}\right) \to \mathcal{N}(0, \sigma^2) \tag{4.17}$$

where the variance has the following form:

$$\sigma^2 := \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} \omega_{it}^\star \left((u_{it} + W_{it}\left(\tau_{it} - \mathbb{E}[\tau_{it}|\underline{W}_i, X_i]\right)\right)\right)^2\right] \tag{4.18}$$

where $\omega_{it}^\star := \omega^\star(X_i, \underline{W}_i, t)$*, and* u_{it} *is equal to either* ξ_{it} *or* ε_{it}*.*

This theorem describes the performance of our estimator in larger samples. The population weights $\omega^\star$ depend on $(X_i, \underline{W}_i)$, not only on S_i which is an implication of the fact that we need to deal with individual fixed effects.

Our next result shows that standard nonparametrtic bootstrap provides a conservative estimator for σ^2.

Theorem 3. *Let* $\{\hat{\tau}_{(b)}\}_{b=1}^B$ *be a set of non-parametric (unit-level) bootstrap analogs of* $\hat{\tau}$*. Define:*

$$\hat{\sigma}^2 := \frac{N}{B}\sum_{b=1}^{B}\left(\hat{\tau}_{(b)} - \hat{\tau}\right)^2 \tag{4.19}$$

and suppose that assumption of Theorem 2 hold. Then if $\mathbb{E}[\tau_{it}|\underline{W}_i, X_i] = \tau$ $\hat{\sigma}^2$ *is consistent for* σ^2*; otherwise* $\hat{\sigma}^2$ *is conservative.*

5 Conclusion

In this paper, we propose a novel identification argument that can be used to evaluate a causal effect using panel data. We show that one can naturally combine familiar restrictions on the relationship between the outcome and the unobserved unit-level characteristics with reasonable economic models of the assignment. Our approach allows us to construct a doubly robust identification argument: out estimand has causal interpretation if either the outcome model is correct, or the assignment model is correct (or both). Using these results, we construct a natural generalization of the standard two-way fixed effects estimator that is robust to arbitrary heterogeneity in treatment effects and show that it has reasonable theoretical properties.

References

A Abadie and MD Cattaneo. Econometric methods for program evaluation. Annual Review of Economics, 18, 2018.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490):493–505, 2010.

Victor Aguirregabiria, Jiaying Gu, and Yao Luo. Sufficient statistics for unobserved heterogeneity in structural dynamic logit models. arXiv preprint arXiv:1805.04048, 2018.

Joseph G Altonji and Rosa L Matzkin. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica, 73(4):1053–1102, 2005.

Manuel Arellano. Panel data econometrics. Oxford university press, 2003.

Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data analysis. 2011.

Manuel Arellano and Bo Honoré. Panel data models: some recent developments. Handbook of econometrics, 5:3229–3296, 2001.

Dmitry Arkhangelsky and Guido Imbens. The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research, 2018.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Timothy Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. 2018a.

Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. Econometrica, 86(2):655–683, 2018b.

Susan Athey and Guido Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. 2018.

Susan Athey, Guido Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. arXiv preprint arXiv:1604.07125, 2016.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. arXiv preprint arXiv:1710.10251, 2017.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. arXiv preprint arXiv:1811.04170, 2018.

Stephen Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

Gary Chamberlain. Panel data. Handbook of econometrics, 2:1247–1318, 1984.

Gary Chamberlain. Efficiency bounds for semiparametric regression. Econometrica: Journal of the Econometric Society, pages 567–596, 1992.

Gary Chamberlain. Binary response models for panel data: Identification and information. Econometrica, 78(1):159–168, 2010.

Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel models. Econometrica, 81(2):535–580, 2013.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. arXiv preprint arXiv:1802.08667, 2018a.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning l2 continuous regression functionals via regularized riesz representers. arXiv preprint arXiv:1809.05224, 2018b.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. 2018.

David L Donoho et al. Statistical estimation and optimal recovery. The Annals of Statistics, 22 (1):238–270, 1994.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. Technical report, Working Paper, 2017.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. arXiv preprint arXiv:1712.00038, 2017.

Bo E Honoré and Ekaterini Kyriazidou. Panel data discrete choice models with lagged dependent variables. Econometrica, 68(4):839–874, 2000.

Kosuke Imai and In Song Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? American Journal of Political Science, 63(2):467–490, 2019.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263, 2014.

Guido Imbens. The role of the propensity score in estimating dose–response functions. Biometrika, 87(0):706–710, 2000.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, pages 1–29, 2004.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Thierry Magnac. Panel binary variables and sufficiency: generalizing conditional logit. Econometrica, 72(6):1859–1876, 2004.

Shahar Mendelson. Learning without concentration. In Conference on Learning Theory, pages 25–39, 2014.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

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

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511):910–922, 2015.

6 Appendix

6.1 Propositions

Proof of Proposition 1: For any $\omega \in \mathbb{W}_{\text{outc}}$ we defined the random variables

$$\omega_{k_{(i)}t} \equiv \sum_{k=1}^{K} \omega_{kt}\{\underline{W}_i = \boldsymbol{W}_k\} \tag{A.1}$$

and considered the following estimator:

$$\tau(\omega) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\omega_{k_{(i)}t}\right] \tag{A.2}$$

By assumption we have the representation:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\omega_{k_{(i)}t}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i) + \lambda_t + \tau(U_i)W_{it} + \varepsilon_{it})\omega_{k_{(i)}t}\right] = \\
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i) + \lambda_t + \tau(U_i)W_{it} + \varepsilon_{it})\sum_{k=1}^{K}\omega_{kt}\{\underline{W}_i = \boldsymbol{W}_k\}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i)\omega_{kt}\{\underline{W}_i = \boldsymbol{W}_k\})\right] + \\
&\frac{1}{T}\sum_{t=1}^{T}\lambda_t\sum_{k=1}^{K}\mathbb{E}\left[\omega_{kt}\{\underline{W}_i = \boldsymbol{W}_k\}\right] + \mathbb{E}\left[\tau(U_i)\{\underline{W}_i = \boldsymbol{W}_k\}\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\boldsymbol{W}_{kt}\omega_{kt}\right] = \\
&\qquad \frac{1}{T}\sum_{t=1}^{T}\lambda_t\sum_{k=1}^{K}\pi_k\omega_{kt} + \mathbb{E}\left[\tau(U_i)\xi(\underline{W}_i)\right] = \mathbb{E}\left[\tau(U_i)\xi(\underline{W}_i)\right]
\end{aligned} \tag{A.3}$$

where $\xi(\underline{W}_i) := \{\underline{W}_i = \boldsymbol{W}_k\}\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\boldsymbol{W}_{kt}\omega_{kt} \geq$. The first equality follows from the restrictions on the outcome model, the second – by definition of the weights, the third – because $\mathbb{E}[\varepsilon_i|U_i] = 0$ and strict exogeneity assumption; finally the last two equalities follow by construction of weights. By construction we also have that $\xi(\underline{W}_i) \geq 0$ and $\mathbb{E}[\xi(\underline{W}_i)] = 1$. This proves the claim.

Proof of Proposition 3: The proof is very similar to the one above and is omitted.

Proof of Proposition 2: We need to prove the following for arbitrary $\underline{w}$ and measurable A_0, A_1:

$$\mathbb{E}[\{\underline{W}_i = \underline{w}\}\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|S_i] = \mathbb{E}\{\underline{W}_i = \underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|S_i] \tag{A.4}$$

We have the following chain of equalities that proves the claim.

$$\begin{aligned}
&\mathbb{E}[\{\underline{W}_i = \underline{w}\}\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|S_i] = \\
&\mathbb{E}[\{\underline{W}_i = \underline{w}\}\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|S_i, U_i, \underline{W}_i]|S_i] = \\
&\mathbb{E}[\{\underline{W}_i = \underline{w}\}\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|U_i, S_i]|S_i] = \\
&\mathbb{E}\mathbb{E}[\{\underline{W}_i = \underline{w}\}|S_i, U_i]\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|U_i, S_i]|S_i] = \\
&\mathbb{E}[\mathbb{E}[\{\underline{W}_i = \underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|U_i, S_i]|S_i] = \\
&\qquad \mathbb{E}\{\underline{W}_i = \underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0) \in A_0, \underline{Y}_i(1) \in A_1\}|S_i]
\end{aligned} \tag{A.5}$$

where the second inequality follows by strict exogeneity, the fourth one – by sufficiency.

6.2 Lemmas

Lemma A.1. *Suppose that $\{W_{it}\}_{i,t}$ are such that there is no $\{\alpha_i, \beta_t, \gamma\}_{i,t}$ such that the following is true:*

$$\begin{aligned}
&\alpha_i + \beta_t + \psi_{it}^\top \gamma \geq 0 \\
&W_{it} = \{\alpha_i + \beta_i + \psi_{it}^\top \gamma > 0\}
\end{aligned} \tag{A.6}$$

Then (a) the primal problem always has a unique solution and (b) the strong duality holds, i.e., for a function

$$\begin{aligned}
&h(\lambda, \mu, \pi, \gamma, \omega) := \frac{1}{(nT)^2} \sum_{it} \omega_{it}^2 + \frac{1}{n} \sum_i \lambda_{(i)} \left(\frac{1}{T} \sum_i \omega_{it} \right) + \\
&\frac{1}{T} \sum_t \lambda_{(t)} \left(\frac{1}{n} \sum_t \omega_{it} \right) + \pi \left(1 - \frac{1}{nT} \sum_{it} \omega_{it} W_{it} \right) - \\
&\qquad \gamma^\top \left(\frac{1}{nT} \sum_{it} \omega_{it} \psi_{it} \right) - \frac{1}{nT} \sum_{it} \mu_{it} \omega_{it} W_{it}
\end{aligned} \tag{A.7}$$

we have

$$\inf_{\omega_{it}} \sup_{\lambda_{(t)}, \lambda_{(i)}, \gamma, \mu_{it} \geq 0, \pi \geq 0} h(\lambda, \mu, \pi, \gamma, \omega) = \sup_{\lambda_{(t)}, \lambda_{(i)}, \gamma, \mu_{it} \geq 0, \pi \geq 0} \inf_{\omega_{it}} h(\lambda, \mu, \pi, \gamma, \omega) \tag{A.8}$$

Proof. Direct application of Generalized Farkas' lemma implies that the constraint set is empty iff there exist $(\alpha_i^\star, \beta_t^\star, \gamma^\star)$ such that the following is true:

$$\begin{aligned}
&\alpha_i^\star + \beta_t^\star + \psi_{it}^\top \gamma^\star \geq 0 \\
&W_{it} = \{\alpha_i^\star + \beta_t^\star + \psi_{it}^\top \gamma^\star > 0\}
\end{aligned} \tag{A.9}$$

By assumption such $(\alpha_i^\star, \beta_t^\star, \gamma^\star)$ does not exist and thus the constraint set is not empty and convex. Since the objective function is strictly convex we have that the primal problem has the unique solution. Since all the inequality constrains are affine strong duality holds (see 5.2.3 in Boyd and Vandenberghe [2004]) and we have the result. □

Lemma A.2. *For arbitrary γ define $g(X, \underline{W}, \gamma)$ in the following way:*

$$g(X, \underline{W}, \gamma) \in \arg\min_{\alpha} \left\{ \frac{1}{T} \sum_{t=1}^{T} \rho_{W_t}(W_t - \alpha - \psi_t^\top \gamma) \right\} \tag{A.10}$$

Then for any $\underline{W}$ such that $\overline{W} < 1$ this function is uniquely defined. Also if $\|\psi_t\|_\infty < K$ then $g(X, \underline{W}, \gamma)$ is $\mathcal{P}$ a.s. uniformly (in $(X, \underline{W})$) Lipschitz in γ.

Proof. If $\overline{W} < 1$ then the minimized function is strictly convex with a unique minimum. Define $h_t := W_t - \psi_t^\top \gamma$; and let $\tilde{h}_{(1)}, \dots, \tilde{h}_{(\sum_{t=1}^T W_t)}$ be the decreasing ordering of h_t for units with $W_t = 1$; let $\tilde{h}_{(0)} = 0$. For $k = 0, \dots, \sum_{t=1}^T W_t$ define the following functions:

$$g_k(X, \underline{W}, \gamma) := \frac{\sum_{t=1}^T (1 - W_{it}) h_t + \sum_{l=0}^k \tilde{h}_{(l)}}{\sum_{t=1}^T (1 - W_{it}) + k} \tag{A.11}$$

It is easy to see that we have the following:

$$g(X, \underline{W}, \gamma) = g_0(X, \underline{W}, \gamma) + \sum_{l=1}^{k} \{\tilde{h}_{(l)} \geq g_{(l-1)}\}(g_l(X, \underline{W}, \gamma) - (g_{l-1}(X, \underline{W}, \gamma)) \tag{A.12}$$

From this representation if follows that $g(X, \underline{W}, \gamma)$ is differentiable and $\mathcal{P}$-a.s. uniformly (in $(X, \underline{W})$) Lipschitz in γ. □

Lemma A.3. *Let $\{\underline{W}_i, X_i\}$ be distributed according to $\mathcal{P}$; assume that S_i includes $\overline{W}_i$ and $\mathbb{E}[W_{it}|S_i, X_i] < 1 - \eta$ $\mathcal{P}$ a.s. for $\eta > 0$. Then there exist a $\sigma(\underline{W}_i, X_i)$-measurable random variable $\alpha_i^\star$ and a vector $\gamma^\star$*

such that the following conditions are satisfied:

$$
\begin{aligned}
&\xi_{it} := W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \\
&\mathbb{E}\left[\sum_{t=1}^{T} \xi_{it}\psi_{it}(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\})\right] = 0 \\
&\sum_{t=1}^{T} \xi_{it}(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}) = 0
\end{aligned}
\tag{A.13}
$$

Proof. Define $\mathcal{F} := \{f \in L_2(\mathcal{P})^T : f_t = g(\underline{W}_i, X_i) + h_t(S_i, X_i), g, h_t \in L_\infty(\mathcal{P})\}$, similarly define $\mathcal{G} := \{g = (g_1, \dots g_T) : g_t = f + \psi_t^\top\gamma, f \in L_2(\mathcal{P}), \gamma \in \mathbb{R}^p\}$.
Consider the following optimization program:

$$
\inf_{g \in \mathcal{G}} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \rho_{W_{it}}(W_{it} - g_{it})\right]
\tag{A.14}
$$

and let $r^\star$ be the value of infimum. We prove that there exists a function $g^\star \in \mathcal{G}$ that solves this problem. This is not entirely trivial because $\mathcal{G}$ is not compact and the loss function is not quadratic so we cannot directly use neither Weierstrass nor the standard projection theorem.
Consider the set $\mathcal{F}(r^\star) := \{f \in \mathcal{F} : \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \rho_{W_{it}}(W_{it} - f_{it})\right] \leq r^\star\}$. It is straightforward to see that this set is convex and because $R(f)$ is continuous on $\mathcal{L}_2^T(\mathcal{P})$ it follows that $f \in \overline{\mathcal{F}(r^\star)} \Rightarrow R(f) \leq r^\star$. The set $\overline{\mathcal{F}(r^\star)}$ is closed and convex. Now assume that $g^\star$ does not exist and thus $\overline{\mathcal{F}(r^\star)} \cap \mathcal{G} = \emptyset$. By construction $\mathcal{G}$ is closed (in $L_2(\mathcal{P})$) and convex; as a result we have two closed convex sets with empty intersection.
Assume that $\overline{\mathcal{F}(r^\star)}$ is weakly compact then by strict separating hyperplane theorem it follows that there exist $h^\star \in L_2^T(\mathcal{P})$ and $a \in \mathbb{R}$ such that $\sup_{f \in \overline{\mathcal{F}(r^\star)}}(f, h^\star) < a_1 < a_2 < \inf_{g \in \mathcal{G}}(g, h^\star)$. Assume that there exist a function $f^\star \in \mathcal{F}(r^\star) \cup \mathcal{G}^0$ such that $R(f^\star) \leq R(f)$ for any function $f \in \mathcal{F}(r^\star) \cup \mathcal{G}^0$. Fix an $\epsilon > 0$ and consider a function $g_\varepsilon \in \mathcal{G}$ such that $R(g_\varepsilon) < r^\star + \varepsilon$. Using this function construct $g_\varepsilon^0 \in \mathcal{G}^0$ such that $R(g_\varepsilon^0) < r^\star + \varepsilon$. For $t \in [0, 1]$ consider a function $r(t) = R(f^\star + t(f^\star - g_\varepsilon^0))$. By convexity of t it follows that $r(t)$ is convex and by definition of $f^\star$ it follows that $r(t)$ has a minimum at zero.
For $t \in [0, 1]$ consider a function:

$$
(h^\star, f^\star + t(g_\varepsilon^0 - f^\star)) =: a + bt
\tag{A.15}
$$

and define $t_1 := \frac{a_1 - a}{b}$ and $t_2 := \frac{a_2 - a}{b}$. It follows that $\frac{t_2 - t_1}{t_1} = \frac{a_2 - a_1}{a_1 - a} > 0$ – does not depend on g^0_ε. By construction it follows that $r(t_1) \geq r^\star$ and $r(t_2) < r^\star + \varepsilon$ and by convexity we have $r(t_2) \geq r(t_1) + \frac{r(t_1)) - r(0)}{t_1} \times (t_2 - t_1) \geq r^\star + \frac{r^\star - R(f^\star)}{t_1} \times (t_2 - t_1)$. The RHS of this inequality does not depend on ε which leads to contradiction.

To finish the proof we need to show that (a) $f^\star$ exists and is unique and (b) that $\overline{\mathcal{F}(r^\star)}$ is weakly compact. The latter statement will follow if we prove that $\mathcal{F}(r^\star)$ is bounded in $L_2(\mathcal{P})$. This follows because $R(f)$ is convex and has a unique minimum at $f^\star$ in $\mathcal{F}(r^\star)$.

Finally we prove the $R(f)$ has a unique minimum at $f^\star$. Consider $f^\star$ such that $f^\star_t := \mathbb{E}[W_{it}|S_i, X_i]$. Because S_i includes $\overline{W}_i$ it follows that $\frac{1}{T}\sum_{t=1}^T f^\star_t = \overline{W}_i$. Take any function $f \in \mathcal{F}$ and consider a convex combination $f(\lambda) := f^\star + \lambda(f - f^\star)$. Because $f \in L_\infty(\mathcal{P})$ and $f^\star_t \leq 1 - \eta$ it follows that for all $\lambda < \lambda_0$ we have $f_t(\lambda) < 1$ almost surely. For any $\lambda < \lambda_0$ we have that $R(f(\lambda)) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (W_t - f^\star_t)^2\right] + \mathbb{E}\left[\sum_{t=1}^T (f^\star_t - f_t(\lambda))^2\right] > R(f^\star)$. By convexity of $R(f)$ it follows that $R(f) > R(f^\star)$ which proves that $g^\star$ exists. The final result follows because $R(f)$ is Gato-differentiable on $\mathcal{F}$ and the results follows by taking first order conditions. $\square$

6.3 Theorems

Proof of Theorem 2: We split the proof into two parts. First, we assume that $\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2 = o_p(1)$, $(\omega^\star_{it})^{un}$ is uniformly bounded, and $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (\omega^\star_{it})^{un} W_{it}\right] > 0$, and prove the normality result. Then we prove the first statement.

Part 1: Assume that $\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2 = o_p(1)$.

For the estimator $\hat{\tau}$ we have the following:

$$\hat{\tau} = \frac{1}{nT}\sum_{it}\hat{\omega}_{it}Y_{it} = \frac{1}{nT}\sum_{it}\hat{\omega}_{it}\tau_{it}W_{it} + \frac{1}{nT}\sum_{it}\hat{\omega}_{it}u_{it} = \tau_{emp} + \frac{1}{nT}\sum_{it}\hat{\omega}_{it}u_{it} = \\ \tau_{emp} + \frac{1}{\mathbb{P}_n \frac{1}{T}\sum_{t=1}^T \hat{\omega}^{un}_{it}W_{it}}\left(\frac{1}{nT}\sum_{it}(\omega^\star_{it})^{un}u_{it} + \frac{1}{nT}\sum_{it}(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}\right) \quad \text{(A.16)}$$

By construction and assumption we have the following:

$$\mathbb{E}[(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}|\{\underline{W}_j, X_j\}_{j=1}^n] = (\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})\mathbb{E}[u_{it}|\{\underline{W}_j, X_j\}_{j=1}^n] = \\ (\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})\mathbb{E}[u_{it}|\underline{W}_i, X_i] = 0 \quad \text{(A.17)}$$

This implies that by conditional Chebyshev inequality we have the following:

$$\zeta_n(\epsilon) := \mathbb{E}\left[\left\{\sqrt{n}\left|\mathbb{P}_n\frac{1}{T}\sum_{t=1}^T(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}\right| \geq \epsilon\right\}|\{\underline{W}_j, X_j\}_{j=1}^n\right] \leq \\ \frac{\mathbb{P}_n\mathbb{E}\left[\left(\sum_{t=1}^T(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un}\right)^2|\{\underline{W}_j, X_j\}_{j=1}^n\right]}{T^2\epsilon^2} \leq \frac{\overline{\sigma}_u^2}{T\epsilon^2}\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2^2 = o_p(1) \quad \text{(A.18)}$$

Since indicator is a bounded function it follows that for any $\epsilon > o$

$$\mathbb{E}[\zeta_n(\epsilon)] = o(1) \quad \text{(A.19)}$$

and thus we have $\frac{1}{nT}\sum_{it}\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2 u_{it} = o_p\left(\frac{1}{\sqrt{n}}\right)$. Finally we need to check that CLT applies to $\frac{1}{nT}\sum_{it}(\omega^\star_{it})^{un}u_{it}$. The mean of each summand is zero and the variance is bounded:

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^T(\omega^\star_{it})^{un}u_{it}\right)^2\right] \leq \frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[((\omega^\star_{it})^{un}u_{it})^2\right] \leq \sum_{t=1}^T\sqrt{\mathbb{E}[u_{it}^4]\mathbb{E}[((\omega^\star_{it})^{un})^4]} < \infty \quad \text{(A.20)}$$

Finally, define:

$$\omega^\star_{it} := \frac{(\omega^\star_{it})^{un}}{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (\omega^\star_{it})^{un} W_{it}\right]} \tag{A.21}$$

It is easy to see that we have:

$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^T \hat{\omega}^{un}_{it} W_{it} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (\omega^\star_{it})^{un} W_{it}\right] + o_p(1) \tag{A.22}$$

and thus we have the following:

$$\begin{aligned} &\|\omega^\star - \hat{\omega}\|_2 = o_p(1) \\ &\sqrt{n}(\hat{\tau} - \tau_{emp}) \to \mathcal{N}(0, \sigma_\tau^2) \end{aligned} \tag{A.23}$$

which concludes the first part.

Part 2: In this part we prove that $\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2 = o_p(1)$, $(\omega^\star_{it})^{un}$ is uniformly bounded, and $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (\omega^\star_{it})^{un} W_{it}\right] > 0$. We use the dual representation derived in Section 4.3 and show that the solution converges to a population one.

The proof below shows that empirical weights converge to oracle weights that solve a certain problem in population. We use a natural adaptation of the "small-ball" argument from Mendelson [2014]. This is not necessary and most likely one can construct a simpler proof using classical results for GMM estimators. We present a different argument because it can be naturally generalized to handle more sophisticated estimation procedures – something that we want to address in future work.

We start by defining relevant oracle weights. Consider $(\{\alpha^\star_i\}_{i=1}^n, \gamma^\star)$ that satisfy the following restrictions:

$$\begin{aligned} &\xi_{it} := W_{it} - \alpha^\star_i - \psi_{it}^\top \gamma^\star \\ &\mathbb{E}\left[\sum_{t=1}^T \xi_{it}\psi_{it}(1 - W_{it}\{W_{it} - \alpha^\star_i - \psi_{it}^\top\gamma^\star \le 0\})\right] = 0 \\ &\sum_{t=1}^T \xi_{it}(1 - W_{it}\{W_{it} - \alpha^\star_i - \psi_{it}^\top\gamma^\star \le 0\}) = 0 \end{aligned} \tag{A.24}$$

Where we include time fixed effects λ_t into the definition of ψ_{it}, since T is fixed this does not create any problems. We prove that oracle weights that satisfy these restrictions exists in Lemma A.3. Using these parameters we consider a lower bound on individual components of the loss function:

$$
\begin{aligned}
&\rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^\top \gamma) = (W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i - \psi_{it}^\top \gamma \le 0\}\right) = \\
&(W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right) + \\
&(W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 W_{it} \left(\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\} - \{W_{it} - \alpha_i - \psi_{it}^\top \gamma \le 0\}\right) \ge \\
&(W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right) - \\
&\qquad (W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 W_{it}\{\alpha_i^\star + \psi_{it}^\top \gamma^\star < 1 \le \alpha_i + \psi_{it}^\top \gamma\} \quad \text{(A.25)}
\end{aligned}
$$

Using this and the properties of the oracle weights we get the following inequality for the excess loss for unit i:

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(\rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^\top \gamma) - \rho_{W_{it}}(W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star)\right) \ge \\
&\sum_{t=1}^{T} \left(((\alpha_i^\star - \alpha_i) + \psi_{it}^\top(\gamma^\star - \gamma))^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right)\right) + \\
&\sum_{t=1}^{T} \left(\xi_{it}(\alpha_i^\star - \alpha_i^\star) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right)\right) + \\
&\sum_{t=1}^{T} \left(\xi_{it}\psi_{it}^\top(\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right)\right) - \\
&\sum_{t=1}^{T} \left((W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 W_{it}\{\alpha_i^\star + X_i^\top \gamma^\star < 1 \le \alpha_i + \psi_{it}^\top \gamma\}\right) = \\
&\sum_{t=1}^{T} \left(((\alpha_i^\star - \alpha_i) + \psi_{it}^\top(\gamma^\star - \gamma))^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right)\right) + \\
&\sum_{t=1}^{T} \left(\xi_{it}\psi_{it}^\top(\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right)\right) - \\
&\qquad \sum_{t=1}^{T} \left((W_{it} - \alpha_i - \psi_{it}^\top \gamma)^2 W_{it}\{\alpha_i^\star + \psi_{it}^\top \gamma^\star < 1 \le \alpha_i + \psi_{it}^\top \gamma\}\right) \quad \text{(A.26)}
\end{aligned}
$$

Note that the last equality follows by definition of ξ_{it} and $(\{\alpha_i^\star\}_{i=1}^n, \gamma^\star)$.

In Lemma A.2 we show that $\alpha_i^\star$ is a function of $\gamma^\star$ and data for unit i:

$$\alpha_i^\star = g(X_i, W_i, \gamma^\star) \tag{A.27}$$

and prove that g is uniformly Lipschitz. By construction for every γ we only need to consider α_i that satisfies the following equality:

$$\alpha_i = g(X_i, W_i, \gamma) \tag{A.28}$$

Define:

$$\begin{aligned} f_{it} &= \alpha_i + \psi_{it}^\top \gamma \\ f_{it}^\star &= \alpha_i^\star + \psi_{it}^\top \gamma^\star \end{aligned} \tag{A.29}$$

and observe that we have the following:

$$\begin{aligned} \mathbb{P}_n \sum_{t=1}^{T} (1 - W_{it}\{W_{it} < f_{it}^\star\})(f_{it} - f_{it}^\star)^2 \ge \mathbb{P}_n \sum_{t=1}^{T} (1 - W_{it})(f_{it} - f_{it}^\star)^2 \ge \\ (\gamma - \gamma^\star)^\top \left(\sum_{t=1}^{T} \mathbb{P}_n \Gamma_{it} \Gamma_{it}^\top \right) (\gamma - \gamma^\star) = \kappa \|\gamma - \gamma^\star\|_2^2 + o_p(\|\gamma - \gamma^\star\|_2^2) \end{aligned} \tag{A.30}$$

where

$$\Gamma_{it} := (1 - W_{it})\psi_{it} - \frac{\sum_{l=1}^{T} (1 - W_{il})\psi_{il}}{\sum_{l=1}^{T} (1 - W_{il})} \tag{A.31}$$

Assume that $\|\gamma - \gamma^\star\|_2^2 = r^2$, which implies that $|\alpha_i - \alpha_i^\star| \le C_1 r$. Assumptions guarantee that ψ_{it} is bounded and thus $\sum_{t=1}^{T} \|f_t - f_t^\star\|_\infty \le C_2 r$. Using CS we get the following inequality:

$$\begin{aligned} \mathbb{P}_n \xi_{it} \psi_{it}^\top (\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right) \le \\ \|\gamma^\star - \gamma\|_2 \times \left\| \mathbb{P}_n \xi_{it} \psi_{it} \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \le 0\}\right) \right\|_2 \end{aligned} \tag{A.32}$$

We also have the following inequality:

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\alpha_i-\psi_{it}^\top\gamma)^2 W_{it}\{\alpha_i^\star+\psi_{it}^\top\gamma^\star<1\leq\alpha_i+\psi_{it}^\top\gamma\}\right]\leq$$
$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(f_{it}^\star-f_{it})^2\{f_{it}^\star<1\leq f_{it}\}\right]\leq\|f^\star-f\|_\infty^2\times\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}\}\right] \quad \text{(A.33)}$$

where the first implication follows because of the indicator, and the the second one follows by Holder inequality. Since $\|f^\star-f\|_\infty\leq C_2 r$ we have the following:

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}\}\right]\leq\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2 r\}\right] \quad \text{(A.34)}$$

DKW inequality implies that we have the following with high probability:

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2 r\}\right]\leq\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2 r\}\right]+\frac{C_3}{\sqrt{n}} \quad \text{(A.35)}$$

It is now easy to see that if r is greater than $O\left(\frac{1}{\sqrt{n}}\right)$ then the excess loss is positive with high probability. Since the loss function is convex this implies that optimum should belong to a ball of radius $\frac{1}{\sqrt{n}}$ around $(\{\alpha_i^\star\}_{i=1}^n,\gamma^\star)$ with high probability which proves that for all t $\|\hat{\omega}_t^{(un)}-(\omega_t^\star)^{un}\|_2=o_p(1)$.

Proof of Theorem 3:

Part 1 For each observation i define M_i – the number of times this observation is sampled in a bootstrap sample. Using this notation we can define bootstap analogs of α_i and γ from the proof of Theorem 2:

$$\{\alpha_i^{(b)}, \gamma^{(b)}\}_{i=1}^n = \arg\min \mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^T \rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^T\gamma) \tag{A.36}$$

in case if $M_i = 0$ we define $\alpha_i^{(b)}$ using the function $g(X_i, W_i, \gamma^\star)$ from 2. It is straightforward to extend the proof of Theorem 2 and show that bootstrap weights converge to population ones. Most part follow because of two key properties of $\{M_i\}_{i=1}^n$:

$$\begin{aligned} &\mathbb{P}_n M_i X_i = \mathbb{E}[X_i] + o_p(1) \\ &\mathbb{P}_n M_i \varepsilon_i = O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.37}$$

for any square integrable X_i and any square integrable mean-zero ε_i (all independent of M_i). The second inequality follows by applying Chebyshev inequality, the first one follows from the second one. The only additional result that we need is the following one:

$$\begin{aligned} &\mathbb{P}_n M_i \left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] = \mathbb{P}_n (M_i - 1)\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] + \\ &\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\} - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right]\right] + \\ &\qquad \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.38}$$

where the last line follows by DKW inequality, the fact that the set of intervals is Donsker, and the multiplier process converges to same limit process as the standard empirical one. It follows that we have convergence results:

$$\begin{aligned} &\|\omega^{(b)} - \omega^\star\|_\infty = o_p(1) \\ &\|\omega^{(b)} - \omega^\star\|_2 = O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.39}$$

Part 2: By construction of bootstrap estimator we have the following representation:

$$\hat{\tau}^{(b)} - \hat{\tau} = \mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \tau_{it} W_{it} +$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} u_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} u_{it} =$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} (\tau_{it} - \mathbb{E}[\tau_{it}]) W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} (\tau_{it} - \mathbb{E}[\tau_{it}]) W_{it} +$$
$$\mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} u_{it} + o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.40)}$$

From this representation it follows that if $\tau_{it} = \mathit{const}$ then the bootstrap estimator is consistent for the asymptotic variance of $\hat{\tau}$. In case if τ_{it} is heterogenous we further expand the first term. Define $\tau_t(\underline{W}_i, X_i) := \mathbb{E}[\tau_{it} | \underline{W}_i, X_i]$ and $\eta_{it} := \tau_{it} - \tau_t(\underline{W}_i, X_i)$. We have the following:

$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \tau_{it} W_{it} =$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_t(\underline{W}_i, X_i) W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \tau_t(\underline{W}_i, X_i) W_{it} +$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \eta_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \eta_{it} W_{it} =$$
$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} (M_i \omega_{it}^{(b)} - \hat{\omega}_{it}) \tau_t(\underline{W}_i, X_i) W_{it} + \mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} \eta_{it} W_{it} + o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.41)}$$

It follows that we have the following:

$$\hat{\tau}^{(b)} - \hat{\tau} = \mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} (\eta_{it} W_{it} + u_{it}) +$$
$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} (M_i \omega_{it}^{(b)} - \hat{\omega}_{it}) \tau_t(\underline{W}_i, X_i) W_{it} + \text{small order terms} \quad \text{(A.42)}$$

Since the second summand is uncorrelated with the first one we have that the bootstrap variance is a conservative estimator of the correct variance.