NBER WORKING PAPER SERIES

DOUBLE-ROBUST IDENTIFICATION FOR CAUSAL PANEL DATA MODELS

Dmitry Arkhangelsky
Guido W. Imbens

Working Paper 28364
http://www.nber.org/papers/w28364

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2021

This paper benefited greatly from our discussions with Manuel Arellano, Stéphane Bonhomme, and David Hirshberg. We are grateful for comments from seminar participants at CERGE-EI, University of Chicago, University of Georgia, Princeton University, and various conferences. This research was generously supported by ONR grant N00014-17-1-2131. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Double-Robust Identification for Causal Panel Data Models
Dmitry Arkhangelsky and Guido W. Imbens
NBER Working Paper No. 28364
January 2021
JEL No. C01,C1,C23

ABSTRACT

We study identification and estimation of causal effects in settings with panel data. Traditionally researchers follow model-based identification strategies relying on assumptions governing the relation between the potential outcomes and the unobserved confounders. We focus on a novel, complementary, approach to identification where assumptions are made about the relation between the treatment assignment and the unobserved confounders. We introduce different sets of assumptions that follow the two paths to identification, and develop a double robust approach. We propose estimation methods that build on these identification strategies.

Dmitry Arkhangelsky
CEMFI
5 Calle Casado del Alisal
Madrid 28014
Spain
darkhangel@cemfi.es

Guido W. Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
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 researchers are concerned with the presence of unobserved confounders that invalidate simple comparisons between treated and control outcomes. With Y_{it} the outcome of interest for unit i in period t, the general setup we consider is

$$Y_{it} = g_t(W_{it}, U_i, X_{it}, \varepsilon_{it}), \tag{1.1}$$

with W_{it} an indicator for the treatment, U_i the unobserved confounder, X_{it} the observed attributes, and ε_{it} an independent idiosyncratic error term. The possibility that U_i may be correlated with W_{it} even after controlling for observed characteristics prevents us from estimating the average effect of W_{it} on the outcome by comparing treated and control outcomes.

The conventional strategy to deal with the unobserved confounder is to impose restrictions on the outcome model $g_t(\cdot)$ that allow us to remove dependence on U_i. The most common approach in empirical work relies on a linear additive two-way fixed effect specification for the outcome model,

$$g_t(w, u, x, e) = \alpha(u) + \lambda_t + w\tau + x^\top \beta + e, \tag{1.2}$$

in combination with $\varepsilon_{it} \perp\!\!\!\perp \{(W_{il}, X_{il})\}_{l=1}^T$, so that the parameters can be estimated by least squares regression.

In this paper we focus on a different, what we call a design-based, strategy. Recall the basic linear regression omitted-variable-bias formula insight that the bias from an unobserved confounder comes from the combination of its correlation with the outcome and its correlation with W_{it}. As an alternative, or complement, to building a model for the outcomes that restricts the dependence of the outcomes on the unobserved confounders, we can therefore restrict the dependence of the assignment mechanism on the unobserved confounder to remove the bias. Let $\underline{W}_i$ be the T-vector of assignments with typical element W_{it}. The restrictions we consider are

of the form

$$\underline{W}_i \perp\!\!\!\perp U_i \,\Big|\, S_i, \tag{1.3}$$

where S_i is a known function of $\underline{W}_i$. For example, S_i may be equal to the average treatment assignment for unit i, $\overline{W}_i = \sum_t W_{it}/T$. This would amount to the assumption that units with the same fraction of treated periods are comparable. Later in the paper we provide several examples of statistical and structural economic models that justify (1.3) for a particular choice of S_i. If (1.3) holds, we can compare treated unit that are treated with control of unit as long as the units have the same value for S_i. Whether the restrictions in a model such as (1.2) are more plausible than a restriction as in (1.3) depends on the substantive application. For example when nonlinearities are important (*e.g.,* Løken et al. [2012]), the model-based approach may be challenging. The first contribution here is to point out that an alternative to the outcome-model based strategy is to build a model for the assignment mechanism.

The second contribution of the paper is the insight that one can combine the two strategies, model-based and design-based, into a single, doubly-robust identification strategy. Models such as (1.2) validate a particular set of comparisons between treated and control outcomes. The restriction in (1.3) validates a different set of comparisons between treated and control outcomes. In many cases we can focus on comparisons that are validated by both the model in (1.2) and the restriction in (1.3). Such comparisons remain valid as long as at least one of the models is correct.

To implement our strategy we restrict attention to linear estimators that are widely used in economics and statistics (*e.g.*, Donoho et al. [1994], Armstrong and Kolesár [2018b]). Our estimator has the following form:

$$\hat{\tau} = \frac{1}{NT}\sum_{i,t} \gamma_{it} Y_{it} \tag{1.4}$$

where researchers explicitly select the weights γ_{it} by solving a quadratic optimization problem. Our estimator remains consistent if either outcome or assignment model is correctly specified and is robust to arbitrary heterogeneity in treatment effects (thus addressing concerns expressed in de Chaisemartin and D'Haultfœuille [2018]). We also provide an extension to general non-binary treatments.

It is important to note that our strategy is not based on constructing consistent estimates for U_i and then controlling for it. In fact, in the fixed T case we consider, it is usually impossible to build an unbiased estimator for U_i. Instead, we leverage the fact that the distribution of U_i stays constant for units with different assignment paths $\underline{W}_i$ as long as we restrict our attention to subpopulations defined by S_i. This emphasizes the practical role that design assumptions can play in models with unobserved heterogeneity. We can interpret the conventional two-way fixed effect estimator also as following our approach. Specifically, it can be viewed as comparing treated and control units at the same point in time within the set of units with the same fraction of treated periods, that is, conditioning on $S_i = \sum_{t=1}^{T} W_{it}$ (*e.g.*, Mundlak [1978]). However, as we show by a simple example, the two-way fixed effect estimator is not doubly robust, primarily because it controls for S_i only linearly. Our proposed estimator explicitly deals with this problem, while also addressing potential issues that two-way fixed effect estimators have in the presence of heterogeneous treatment effects.

The additive structure in (1.2) has a long history in applied economics (going back at least to Mundlak [1961], Hoch [1962], Mundlak and Hoch [1965]) and econometrics (*e.g.*, Chamberlain [1984, 1992], Arellano and Bonhomme [2011], Graham and Powell [2012], Chernozhukov et al. [2013], Freyberger [2018]). The second approach justifies (1.3) by putting additional restrictions on the relationship between $\underline{W}_i$ and U_i, or, in other words, by formulating an assignment model. This type of restrictions on the assignment mechanism are at the heart of the "credibility revolution" in applied economics (Angrist and Pischke [2010]) that emphasizes the role of the research design. Many strategies currently used for causal inference with cross-sectional data are based on such design assumptions (see Angrist and Pischke [2008], Currie et al. [2020] for evidence on this). Although less common than outcome modeling in panel data settings, this design-based approach has been used to achieve identification in settings with grouped data, *e.g.*, the exchangeability assumption in Altonji and Matzkin [2005], or the exponential family assumption in Arkhangelsky and Imbens [2018] (see also Borusyak and Hull [2020]). We are making two contributions to this literature. First, building on the research on binary panel models (*e.g.*, Honoré and Kyriazidou [2000], Chamberlain [2010], Aguirregabiria et al. [2018]), we show how to use design-based assumptions to identify treatment effects in this setting. Second, we propose a doubly robust identification strategy that combines the models for outcomes and assignments and remains valid if either of them is correctly specified. In practice, this means that applied researchers can directly exploit information about economic mechanisms behind

different patterns in $\underline{W}_i$, without abandoning familiar outcome models such as (1.2).

This paper is also directly related to recent causal panel literature that focuses on two-way fixed effect estimators (de Chaisemartin and D'Haultfœuille [2018], Callaway and Sant'Anna [2019], Sant'Anna and Zhao [2020], Goodman-Bacon [2017], Athey and Imbens [2018]). Similarly to these papers, we use the two-way structure to model the baseline outcomes. However, our focus is quite different. First, we consider general designs, whereas the previous research has been focused on either block case or staggered adoption. Second, we develop a new estimator, that directly utilizes both assignment and outcome model. Finally, our identification strategy can be applied more broadly. In particular, although we focus on the two-way model for the baseline outcomes, the strategy can be extended to general factor models, thus connecting to the literature on synthetic control (*e.g.*, Abadie et al. [2010], Xu [2017], Athey et al. [2017], Abadie et al. [2010], Ben-Michael et al. [2018], Arkhangelsky et al. [2019], Chernozhukov et al. [2019]).

2 Setup

In a generic panel data set up we observe N units over T periods (i and t being a generic unit and period, respectively), *e.g.,* Chamberlain [1984], Arellano and Honoré [2001], Hsiao [2014], Baltagi [2008], Wooldridge [2010], Arellano et al. [2007]. 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}. Later we discuss settings with more general treatments or policies. To formalize this we consider a potential outcome framework (Rubin [1974], Imbens and Rubin [2015]). The policy or treatment can change over time, and so is indexed by unit i and time t, $W_{it} \in \{0, 1\}$. Let $\underline{w}^t := (w_1, w_2, \ldots, w_t)$ 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 $\underline{W}_i := (W_{i1}, \ldots, W_{iT})$ to be the full assignment vector for unit i. For the first part of the paper we abstract from the presence of additional unit-level covariates. We explicitly introduce them in Section 5. 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:

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

In this paper we consider a static version of this general model where some potential outcomes are identical.

Assumption 2.1. (No Dynamics) *For arbitrary t-compponent assignment vectors $\underline{w}^t$ and $\underline{w}^{t\prime}$ such that the period t assignment is the same, $w_t^t = w_t^{t\prime}$ the potential outcomes in period t are the same:*

$$Y_{it}(\underline{w}^t) = Y_{it}(\underline{w}^{t\prime}). \tag{2.2}$$

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. Because 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 conceptual 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 $w \in \{0, 1\}$. Define also $\underline{Y}_i(\underline{w}) \equiv (Y_{i1}(w_1), \ldots, Y_{it}(w_T))$ to be the vector of potential outcomes. In this setup we can be interested in various treatment effects. The main building block is the individual and time-specific treatment effect:

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

We focus primarily on identification and estimation of average treatment effects, typically a convex combination of individual effects τ_{it}.

Next we discuss to assumptions we maintain throughout the paper. 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 variable $U_i \in \mathbb{R}^d$ 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 implies 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: if we choose $U_i = \underline{W}_i$ this assumption is satisfied by construction.

We view U_i as a permanent (time-invariant) unit characteristic that we need to control for if we wish to compare outcomes across different 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. Define the infeasible generalized propensity score:

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

Assumption 2.3. (Latent Overlap) *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 $\underline{W}_i$. Such restrictions are common in the (cross-section) program evaluation literature: without an overlap assumption we would not be able to identify the average causal effect of the treatment without functional form restrictions even if we observed U_i. 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.

Considered together Assumptions 2.2 and 2.3 have testable restrictions, because now we cannot simply choose $U_i = \underline{W}_i$. Crucially, Assumption 2.2 implies that the assignment process does not depend on past outcomes. This restriction is often unrealistic in settings where W_{it} can be viewed as a choice variable that agents use to optimize Y_{it}. In such cases, past outcomes might contain important information and thus are useful for decisions today. As a canonical example, consider a firm that selects inputs to optimize the output while facing uncertainty about future productivity (*e.g.*, Olley and Pakes [1992]). In this case, past outcomes can be informative about the unobserved productivity, affecting the decisions today. As another example, consider a well-

known empirical observation that earnings of labor programs' participants tend to decline right before the start of the program (*e.g.*, Ashenfelter [1978]). In this case, the decision to participate is likely directly affected by recent earnings.

Nothwithstanding these examples, Assumption 2.2 is quite natural when W_{it} is either driven by some (quasi)-experimental shocks or is a choice variable that is not used to optimize Y_{it} directly. Examples of the first type are common in the applied literature, especially when the shocks are aggregate, but some units are more exposed to them than others. As an example of the second type, consider a situation where W_{it} corresponds to national-level prices for product i in period t and Y_{it} is a measure of sales in a local-level market. In this case, unobserved quality U_i makes Y_{it} and W_{it} correlated (over i), despite the fact that W_{it} is not chosen to optimize Y_{it} directly. Overall, the relationship between W_{it} and U_i that satisfies Assumptions 2.2 and 2.3 can arise for a variety of reasons and in the subsequent sections we show why it is useful to model it explicitly.

3 Two paths to 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 $K \leq 2^T$ be the number of rows in $\boldsymbol{W}$.

For example, if $T = 3$ then a possible form for $\boldsymbol{W}$ is:

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

Each row of this matrix represents a possible assignment, and in this particular case only four

Table 1: Assignment process and weights

k	$\boldsymbol{W}_k$	π_k	$\gamma_{k1}^{(fe)}$	$\gamma_{k2}^{(fe)}$	$\gamma_{k3}^{(fe)}$
1	(0,0,0)	0.09	0.46	-0.64	0.18
2	(1,0,0)	0.04	5.70	-3.26	-2.44
3	(0,1,0)	0.11	-2.16	4.60	-2.44
4	(1,1,0)	0.14	3.08	1.98	-5.07
5	(0,0,1)	0.07	-2.16	-3.26	5.42
6	(1,0,1)	0.08	3.08	-5.88	2.80
7	(0,1,1)	0.15	-4.78	1.98	2.80
8	(1,1,1)	0.32	0.46	-0.64	0.18

out of the eight ($2^3 = 8$) possible combinations have positive probability. For a particular unit i, let $k(i)$ be the row $\boldsymbol{W}_k$ of the support matrix $\boldsymbol{W}$ such that $\boldsymbol{W}_k = \underline{W}_i$. For the identification argument we assume we know $\boldsymbol{W}$ and the assignment probabilities π_k. We consider the case with unknown π in Section 5.

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

$$\hat{\tau}(\gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_{it} Y_{it}.$$

For the estimators we consider, the weights γ are a function of the assignment matrix $\boldsymbol{w}$, $\gamma : \{0,1\}^{N \times T} \mapsto \mathbb{R}^{N \times T}$ (but do not depend on the outcomes data). Thus, choosing an estimator corresponds to choosing a weight function $\gamma_{it}(\boldsymbol{w})$. 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 Double Robust Identification – An Example

We start with an example that illustrates the main message of the paper. Suppose that $T = 3$ and $K = 8 = 2^T$, so $\underline{W}_i$ has full support. We assume that the distribution of $\underline{W}_i$ in population is given by the third column of Table 1. Suppose that potential outcomes $Y_{it}(w)$ have the following

structure:

$$\begin{aligned} &Y_{it}(w) = \alpha(U_i) + \lambda_t + \tau w + \varepsilon_{it} \\ &\mathbb{E}[\varepsilon_{it}|\underline{W}_i, U_i] = 0 \end{aligned} \tag{3.2}$$

and we use OLS with two-way fixed effects to "estimate" τ in population. This procedure leads to a particular set of weights $\gamma_{it}^{(\text{fe})}(\boldsymbol{W})$, and then to the following fixed effect estimand:

$$\tau^{\text{fe}} = \mathbb{E}\left[\frac{1}{T}\sum_t Y_{it}\gamma_{it}^{(\text{fe})}(\boldsymbol{W})\right], \tag{3.3}$$

where the expectation is taken over the U_i, ε_{it} and the assignment $\boldsymbol{W}$. If (3.2) is correctly specified then $\tau^{\text{fe}} = \tau$ – the average treatment effect – but in general this equality will not hold. For the distribution given above the weights are presented in the last three column of Table 1. By construction these weights sum up to 0 for every row and every column (once re-weighted by the marginal probability of $\underline{W}_i$).

In general, the model (3.2) is overidentified which means that we can use other weights to identify τ. The weights $\gamma_t^{(fe)}(\cdot)$ are selected for efficiency reasons, because under homoskedasiticity they lead to an estimator with the least possible variance. In practice, we can have concerns besides efficiency. In particular, we may be worried that the model (3.2) is misspecified.

To illustrate this, suppose that DGP for the assignment mechanism $\underline{W}_i$ has the following form:

$$\forall(t,t')\text{: } W_{it} \perp\!\!\!\perp W_{it'}|U_i, \qquad \mathbb{E}[W_{it}|U_i] = \frac{\exp(\alpha(U_i)+\lambda_t)}{1+\exp(\alpha(U_i)+\lambda_t)}. \tag{3.4}$$

As we show in the next section, this assumption implies the following conditional independence:

$$\underline{W}_i \perp\!\!\!\perp \left\{\underline{Y}_i(\underline{w})\right\}_{\underline{w}} \,\Big|\, \overline{W}_i \tag{3.5}$$

where $\overline{W}_i \equiv \sum_{t=1}^T W_{it}/T$ is the fraction of treated periods for unit i. Now we can articulate a key insight. If the two-way fixed effect outcome model (3.2) is misspecified, but the assignment mechanism (3.4) is correctly specified and the treatment effect is constant, $\tau_{it} = \tau$, then the fixed

Table 2: Aggregated weights

$\overline{W}_i$	$\mathbb{E}[\gamma_1^{(fe)}(\underline{W}_i)\|\overline{W}_i]$	$\mathbb{E}[\gamma_2^{(fe)}(\underline{W}_i)\|\overline{W}_i]$	$\mathbb{E}[\gamma_3^{(fe)}(\underline{W}_i)\|\overline{W}_i]$
0	0.46	-0.64	0.18
1/3	-0.73	0.60	0.13
2/3	-0.08	0.36	-0.28
1	0.46	-0.64	0.18

effect estimand is equal to the treatment effect, or $\tau^{fe} = \tau$, as long as the following condition is satisfied for all t and $\overline{W}_i$:

$$\mathbb{E}[\gamma_t^{(fe)}(\underline{W}_i)|\overline{W}_i] = 0 \tag{3.6}$$

This restriction requires the weights $\gamma_t^{(fe)}(\underline{W}_i)$ to balance out any time-specific function of $\overline{W}_i$. If it is not satisfied, then the differences in the outcomes for treated and control units can be attributed in the differences in the baseline outcomes $Y_{it}(0)$.

Table 2 shows that this condition does not hold not true for the fixed effect weights from Table 1. In other words, although fixed effects weights balance individual and time effects overall, they do not necessarily do so within subpopulations defined by $\overline{W}_i$. This is particularly striking, because the two-way estimator can be interpreted as controlling for $\overline{W}_i$. As shown in Mundlak [1978] the fixed effects estimand τ^{fe} is numerically equivalent to the estimand in the following linear regression:

$$Y_{it} = \alpha + \lambda_t + \tau W_{it} + \eta \overline{W}_i + \tilde{\epsilon}_{it} \tag{3.7}$$

As a result, when constructing τ^{fe} we control for $\overline{W}_i$ only linearly and this is not enough to enforce the necessary balancing property.

This example shows that the weights based on the outcome model (3.2) do not work if the assignment model (3.5) is correctly specified. As an alternative to the fixed effect weights one can use the assignment model (3.5) to construct weights that deliver the treatment effect τ if the design model (3.5) is correctly specified. Similarly to the current example, there is no guarantee that such weights will "work" for the outcome model (3.2). The question now arises whether we find the weights that deliver τ if either the fixed effect model (3.2) or the design process (3.5) is correctly specified. The answer is positive and a set of weights that satisfy this restriction

Table 3: Doubly robust weights

(W_1, W_2, W_3)	$\gamma_1^{(dr)}(\underline{W}_k)$	$\gamma_2^{(dr)}(\underline{W}_k)$	$\gamma_3^{(dr)}(\underline{W}_k)$
(0,0,0)	0.00	0.00	0.00
(1,0,0)	6.59	-3.95	-2.64
(0,1,0)	-1.46	4.10	-2.64
(1,1,0)	3.24	1.66	-4.90
(0,0,1)	-1.46	-3.95	5.42
(1,0,1)	3.24	-6.39	3.15
(0,1,1)	-4.81	1.66	3.15
(1,1,1)	0.00	0.00	0.00

are given in Table 3. It is evident that the weights some up to zero for each row and a simple calculation shows that $\mathbb{E}[\gamma_t^{(dr)}(\underline{W}_i)|\overline{W}_i] = 0$ 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 the two-way estimators have problems in cases with heterogenous treatment effects (*e.g.*, see de Chaisemartin and D'Haultfœuille [2018]). This is evident from looking at Table 1: in the last row we assign a negative weight to all treated units in the second period. To guarantee that this does not happen the following restriction can be enforced when the weights are constructed:

$$W_{it}\gamma_t(\underline{W}_i) \geq 0 \tag{3.8}$$

As we show in the next sections this does not make the problem much harder computationally and in fact the robust weights from Table 3 are constructed with this restriction in mind. As a result, in this case it easy to be robust to arbitrary heterogeneity in treatment effects. This should be contrasted with the well-known problem with potentially negative weights that arises in IV estimation (Imbens and Angrist [1994]).

It is also important to emphasize that with general heterogeneity in treatment effects the robust weights might not deliver the average treatment effect (the same is true for the fixed effect weights). Instead, we get a weighted average with observable, or at least estimable, weights. The fact that the weights are estimable is important, because in the empirical work these weights can be reported and analyzed.

3.3 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. Here we focus on an outcome model with the following structure:

Assumption 3.1. *The potential outcomes satisfy:*

$$\mathbb{E}[Y_{it}(w)|U_i] = \alpha(U_i) + \lambda_t + \tau_t(U_i)w. \tag{3.9}$$

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

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

which is unrestricted.

To motivate this outcome model, note that it has a long tradition in econometric literature. Chamberlain [1992] analyzed a more general factor model with an additional restriction $\tau_t(U_i) = \tau(U_i)$, and derived an efficiency bound for $\mathbb{E}[\tau(U_i)]$. For the two-way case his analysis was refined in Graham and Powell [2012] which looked at settings in which the semiparametric efficiency bound is equal to infinity, and thus τ cannot be estimated by a regular estimator. In Arellano and Bonhomme [2012] authors return to the general model of Chamberlain [1992], impose additional time-series restrictions on the errors, and show how they can be used to identify the whole distribution of $\tau(U_i)$.

In our analysis we depart from these papers in two important directions. First, we explicitly allow for $\tau_t(U_i)$ to depend on t in an unrestricted way. Second, we do not focus only on averages of $\mathbb{E}[\tau_t(U_i)]$. Instead we ask which convex combinations of $\tau_t(U_i)$ can be identified. Analysis in de Chaisemartin and D'Haultfœuille [2018] shows that both of these directions are important: there is evidence that treatment effects vary with time, and the standard estimators do not

deliver convex combinations of treatment effects in general.

To identify a convex combination of $\tau(U_i)$ we consider the weights γ_{kt} that satisfy the following four restrictions:

$$\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\pi_k\gamma_{kt}\boldsymbol{W}_{kt}=1, \tag{3.10}$$

$$\forall k, \sum_{t}\gamma_{kt}=0, \tag{3.11}$$

$$\forall t, \sum_{k=1}^{K}\pi_k\gamma_{kt}=0, \tag{3.12}$$

$$\forall (t,k),\ \gamma_{kt}\boldsymbol{W}_{kt}\geq 0 \tag{3.13}$$

These constraints are natural given the outcome model described above. The first and the last restriction insure that we focus on a convex combination of treatment effects. The second and the third restriction guarantee that weights balance out the systematic variation in the baseline outcomes $Y_{it}(0)$. By construction, any weights that satisfy these restrictions lead to within-unit comparisons. We do not include the analogue of non-negativity constraint for control units, thus allowing for extrapolation. Depending on application, one might want to impose such constraint as well.

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

$$\gamma_{k_{(i)}t}(\gamma) := \sum_{k=1}^{K}\gamma_{kt}\mathbf{1}_{\underline{W}_i=\boldsymbol{W}_k} \tag{3.14}$$

Using these stochastic weights we can compute the following expectation:

$$\tau(\gamma) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\gamma_{k_{(i)}t}(\gamma)\right] \tag{3.15}$$

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

As a result, a certain convex combination of $\tau(U_i)$ can be identified whenever $\mathbb{W}_{\text{outc}}$ is non-empty. A necessary and sufficient condition for this is simple: the matrix $\boldsymbol{W}$ should contain at least one of the following two submatrices (up to permutation of columns):

$$\boldsymbol{W}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{W}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.16}$$

Consider each of these cases separately. In the first case there are adoptors i of the treatment, and periods t and t' with $(W_{it} = 0, W_{it'} = 1)$ and in the same periods t and t' non-adoptors i' with $(W_{i't} = 0, W_{i't'} = 0)$. In the second 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

$$\boldsymbol{W}_3 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{W}_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{W}_5 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

where with the first design some units are always in the control group and all others are always in the treatment group, in the second design all units adopt the treatment at exactly the same time. The third design is more complicated, because at a first sight $\boldsymbol{W}_5$ looks very similar to $\boldsymbol{W}_1$; the key difference is that with $\boldsymbol{W}_1$ we have a control period and this allows us to deal with unobserved unit-specific differences. With $\boldsymbol{W}_5$ this is no longer feasible and we need to use negative weights that are not allowed. Standard two-way fixed effect estimator treats $\boldsymbol{W}_1$ and $\boldsymbol{W}_5$ symmetrically and this is the reason why the resulting estimand might be outside of the convex hull of treatment effects.

3.4 Identification Through Design

In this section we consider assignment processes that satisfy a certain sufficiency property. Here we state it as a high-level assumption, and provide examples of economic models that satisfy this assumption in the next section.

Define $r(\underline{w}, s)$ to be the feasible generalized propensity score:

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

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)*

$$\underline{W}_i \perp\!\!\!\perp U_i \;\Big|\; S_i, \tag{3.18}$$

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

$$\max_{\underline{w}} \{r(\underline{w}, s)\} < 1. \tag{3.19}$$

This assumption might look restrictive, but S_i that satisfies the conditional independence assumption (3.18) always exists. The obvious choice is $S_i = \underline{W}_i$ that satisfies this restriction by construction. Of course, in this case the overlap condition is not satisfied. Alternatively, one can consider $S_i^{\text{gen}} \equiv f_{U|\underline{W}}(\cdot|\underline{W}_i)$, where $f_{U|\underline{W}}(x|y)$ is the conditional distribution of U_i given $\underline{W}_i$. While less restrictive than $\underline{W}_i$, S_i^{gen} is not feasible, because $f_{U|\underline{W}}(x|y)$ is unknown. More generally, we need to find different values for $\underline{W}_i$ that generate the same distribution of U_i, but still exhibit variation in some W_{it}. For example, in addition to conditioning on the fraction of treated periods, $\overline{W}_i$, we may want to condition on the number of transitions, $\sum_{t=1}^{T-1} W_{it}(1 - W_{it+1})$. In the next section we describe examples that show that S_i arises naturally in different empirical settings.

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 $\underline{w}$:*

$$\mathbf{1}_{\underline{W}_i = \underline{w}} \perp\!\!\!\perp \underline{Y}_i(\underline{w}) \;\Big|\; S_i. \tag{3.20}$$

This proposition demonstrates that unconfoundedness conditional on U_i (1.3) can be transformed into undonfoundedness conditional on S_i (3.20) under the additional assumption that restricts the assignment process.

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 $\boldsymbol{W}^s_k$ 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:

$$\boldsymbol{W}^0 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}, \boldsymbol{W}^{2/3} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \boldsymbol{W}^1 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \tag{3.21}$$

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):

$$\frac{1}{T}\sum_{tk} \pi_k \gamma_{kt} \boldsymbol{W}_{kt} = 1, \tag{3.22}$$

$$\sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \gamma_{kt} = 0. \tag{3.23}$$

$$\sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \gamma_{kt} \boldsymbol{W}_{kt} \geq 0, \tag{3.24}$$

Let $\mathbb{W}_{\text{design}}$ be the set of weights $\{\gamma_{tk}\}_{t,k}$ 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 $\boldsymbol{W}^s$ contains at least two rows. This is guaranteed by the second part of Assumption 3.2. For any $\gamma \in \mathbb{W}_{\text{design}}$ define the random variables $\gamma_{k_{(i)}t}$ in the same way as before and consider the following expectation:

$$\tau(\gamma) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\gamma_{k_{(i)}t}\right] \tag{3.25}$$

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

3.5 Examples

In this section we consider various examples of assignment models. We show that Assumption 3.2 holds in a wide range of examples and discuss their applicability to various applied problems.

3.5.1 Aggregate shocks

As a first case, we consider a situation, where W_{it} is determined by an observed p-dimensional aggregate shock ψ_t, and idiosyncratic noise ν_{it}. Importantly, different units are affected by ψ_t in a way that is determined by their unobserved U_i. Formally, we have the following model that includes a latent index:

$$\begin{aligned} &W_{it}^\star = \theta_t + \alpha_1(U_i) + \alpha_2^T(U_i)\psi_t + \nu_{it}, \\ &\{\nu_{it}\}_t \perp\!\!\!\perp \{\alpha(U_i), \underline{Y}_i(w)\} \\ &W_{it} = \mathbf{1}_{W_{it}^\star > 0} \end{aligned} \tag{3.26}$$

Here $\alpha(U_i) = (\alpha_1(U_i), \alpha_2(U_i)) \in \mathbb{R}^{p+1}$ captures the exposure of different units to aggregate shocks ψ_t and ν_{it} represents an independent mean-zero idiosyncratic component. Because W_{it} is binary, the model is nonlinear: instead of the latent index $W_{it}^\star$ we observe only $W_{it} = \mathbf{1}_{W_{it}^\star > 0}$.

In this model endogeneity arises because the exposure $\alpha(U_i)$ can be correlated with the potential outcomes. This should be compared with a situation where $\alpha(U_i) = \alpha$, but ν_{it} can be correlated with $\underline{Y}_i(w)$. In the latter case, ψ_t can be used as an aggregate instrument. In our model the situation is different and we can use ψ_t to control for $\alpha(U_i)$. To gain some intuition for why this can be done assume that $W_{it}^\star$ is observed. In this case, we can construct OLS estimates for $\alpha(U_i)$. More precisely, let $\tilde{\psi}_t = (1, \psi_t)$ and define the following estimates:

$$\hat{\alpha}(U_i) = \left(\sum_{t=1}^T \tilde{\psi}_t \tilde{\psi}_t^\top \right)^{-1} \sum_{t=1}^T \tilde{\psi}_t (W_{it}^\star - \overline{W}_t^\star) \tag{3.27}$$

We can use $\hat{\alpha}(U_i)$ instead of $\alpha(U_i)$ to control for U_i. Since neither $\overline{W}_t^\star$ nor $\left(\sum_t \psi_t \psi_t^\top\right)^{-1}$ vary over

i, this is equivalent to controlling for $(\sum_t W_{it}^\star/T, \sum_{t=1}^T \psi_t W_{it}^\star/T)$. This is an intuitive strategy: if endogeneity arises because of the differential exposure to ψ_t, then we can simply estimate this exposure and control for it.

There are two major problems with this strategy: first, we do not observe $W_{it}^\star$, second, $\hat{\alpha}(U_i)$ is not equal to $\alpha(U_i)$. To solve both problems we need to make two additional assumptions. First, we assume that ν_{it} is independent over t, second, we assume that ν_{it} has a logistic distribution. Together these two assumptions lead to the following model:

$$\begin{aligned} &\mathbb{E}[W_{it}|U_i] = \frac{\exp(\alpha_1(U_i) + \theta_t + \alpha_2^T(U_i)\psi_t)}{1 + \exp(\alpha_1(U_i) + \theta_t + \alpha_2^T(U_i)\psi_t)}, \\ &W_{it} \perp\!\!\!\perp \{W_{il}\}_{l \neq t} \,\Big|\, U_i, \end{aligned} \tag{3.28}$$

which is a natural generalization of the familiar two-way logit model

$$\mathbb{E}[W_{it}|U_i] = \frac{\exp(\alpha_1(U_i) + \theta_t)}{1 + \exp(\alpha_1(U_i) + \theta_t)}. \tag{3.29}$$

The key feature of (3.28) is that now the version of the previously outlined strategy works without any additional caveats. In particular, define the following statistic:

$$S_i = \left(\sum_{t=1}^T W_{it}/T, \sum_{t=1}^T \psi_t W_{it}/T \right).$$

S_i has the same interpretation as $\hat{\alpha}(U_i)$: it measures the exposure of unit i to aggregate shocks. S_i depends on observable quantities only and can be computed, thus solving the first problem outlined above. It also solves the second problem, because it is easy to show that the following independence condition holds:

$$U_i \perp\!\!\!\perp \underline{W}_i \,\Big|\, S_i \tag{3.30}$$

and thus S_i satisfies Assumption 3.2.

Undoubtedly the assumptions that make S_i a sufficient statistic are strong. In particular, the logistic distribution of ν_{it} is a functional form assumption on the unobserved idiosyncratic errors. The discussion above shows that we make this assumption to deal with the infeasibility

of $\hat{\alpha}(U_i)$. At the same time, one can interpret S_i as a scale-free estimator for $\alpha(U_i)$ in the spirit of maximum score estimation (Manski [1975], Horowitz [1992]). As a result, we expect it to capture essential aspects of the unobserved heterogeneity even if the logistic assumption is violated.

3.5.2 Stationary dynamics

In the previous example W_{it} was mainly determined by aggregate exogenous shocks. In this section, we consider an opposite situation and assume that W_{it} is mainly determined by its past. In particularly, we consider the following structure:

$$\begin{aligned} &W_{it} \perp\!\!\!\perp \{W_{il}\}_{l>t} \;\Big|\; U_i, W_i^{t-1} \\ &\mathbb{E}[W_{it}|U_i, W_{it-1}] = \frac{\exp(\alpha(U_i) + \eta(U_i)W_{it-1})}{1+\exp(\alpha(U_i) + \eta(U_i)W_{it-1})} \end{aligned} \tag{3.31}$$

This assumption describes a stationary first-order Markov dynamic model with rich unobserved heterogeneity. As we show below it arises naturally when W_{it} is a solution of a dynamic optimization problem, with past choices determining the current state.

As a stylized example, consider a sales manager for product i who decides on the price w_t (high or low) solving a dynamic optimization problem with a discount factor β, and the following instantaneous payoff:

$$\Pi_{it}(w_t, w_{t-1}) = R(w_t, U_i) - c(w_t, w_{t-1}, U_i) + \nu_{it} \tag{3.32}$$

where U_i is the unobserved quality of the product, ν_{it} is an independent (of U_i and over time) idiosyncratic shock. Here the function $R(w_t, U_i)$ reflects the instantaneous profit, and function $c(w_t, w_{t-1}, U_i)$ captures costs of price adjustments. Bellman's principle implies that the optimal W_{it} is a solution of the following problem:

$$V(W_{it-1}, U_i, \nu_{it}) = \max_w \{\Pi_{it}(w, W_{it-1}) + \beta \mathbb{E}[V(\nu_{it+1}, w, U_i)|U_i, \nu_{it}]\} \tag{3.33}$$

where V is a value function. The optimal policy is a function of the state:

$$W_{it} = f_t(W_{it-1}, U_i, \nu_{it}) \tag{3.34}$$

and as long as ν_{it} are i.i.d. (over time) we get that W_{it} satisfies the following restrictions:

$$\begin{aligned} &W_{it} \perp\!\!\!\perp \{W_{il}\}_{l>t} \;\Big|\; U_i, W_i^{t-1} \\ &\mathbb{E}[W_{it}|U_i, W_i^{t-1}] = \mathbb{E}[W_{it}|U_i, W_{it-1}] \end{aligned} \tag{3.35}$$

Let $\pi_i(W_{it-1}) := \mathbb{E}[W_{it}|U_i, W_{it-1}]$ and observe that $\log\left(\frac{\pi_i(W_{it-1})}{1-\pi_i(W_{it-1})}\right)$ is a linear function of W_{it-1}:

$$\log\left(\frac{\pi_i(W_{it-1})}{1-\pi_i(W_{it-1})}\right) = \alpha(U_i) + \eta(U_i)W_{it-1} \tag{3.36}$$

This can be expressed in a familiar logit form:

$$\mathbb{E}[W_{it}|U_i, W_{it-1}] = \frac{\exp(\alpha(U_i) + \eta(U_i)W_{it-1})}{1 + \exp(\alpha(U_i) + \eta(U_i)W_{it-1})} \tag{3.37}$$

which together with (3.35) implies (3.31).

Define the following statistic:

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

It is not hard to show that, as long as conditions (3.31) are satisfied, S_i is a sufficient statistic that satisfies Assumption 3.2. In fact, the discussion above shows that (3.31) is equivalent to (3.35). This means that contrary to our previous example, in this model, sufficiency does not follow from a functional form assumption. The only real restriction that we impose is that W_{it} is a stationary first-order Markov process.

3.5.3 Discussion

Examples from Sections 3.5.1 and 3.5.2 illustrate two different empirical settings in which one can use our approach. The first example emphasizes the role of exogenous aggregate shocks that are frequently used in applied literature to identify policy effects (*e.g.*, Duflo and Pande [2007], Dube and Vargas [2013], Nunn and Qian [2014], Nakamura and Steinsson [2014]). Our approach is applicable as long as the primary reason for endogeneity is differential exposure of different

units to these shocks. The second example emphasizes the role of the structural assumptions, such as Markov restrictions, thus providing a way of combining structural choice models with the estimation of treatment effects.

From the formal point of view all the models considered above share the same structure. In all of them the conditional distribution of $\underline{W}_i$ has the following representation:

$$\log\left(\mathbb{P}(\underline{W}_i|U_i)\right) = S(\underline{W}_i)^\top \alpha(U_i) + \beta(U_i) + \eta(\underline{W}_i) \tag{3.38}$$

In other words, the distribution of W_i belongs to an exponential family, with $S(\underline{W}_i)$ begin the sufficient statistic. Sufficiency arguments have a long tradition in econometrics of binary panel models (*e.g.*, Andersen [1970], Chamberlain [1984], Honoré and Kyriazidou [2000], Chamberlain [2010]) where they have been used to obtain consistent estimators for common parameters. More recently, sufficiency arguments were used by Aguirregabiria et al. [2018] to identify common parameters in dynamic structural models. We are using sufficiency differently: instead of using S_i to identify common parameters, we use it to condition on unobserved heterogeneity, similarly to Arkhangelsky and Imbens [2018].

4 Double robustness

In this section we build on our previous results and present a doubly-robust identification argument. We then propose a natural algorithm that implements our strategy.

4.1 Identification

The sets of $\mathbb{W}_{\text{outc}}$ and $\mathbb{W}_{\text{design}}$ described in Section 3 are motivated by different models and in general do not need to be similar. The first set is built with within-unit comparisons in mind, while the second one is based on within-period comparisons. The non-negativity constraints are also different: in $\mathbb{W}_{\text{outc}}$ we require every treated unit to have a non-negative weight, while in $\mathbb{W}_{\text{design}}$ this property only holds for the weights averages within a subpopulation described by S_i. Nevertheless, these sets are not entirely different, because $\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}}$ can be non-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.

Let $\mathbb{W}_{\text{dr}} = (\mathbb{W}_{\text{outc}} \cap \mathbb{W}_{\text{design}})$ and observe that combining the restrictions in (3.10)-(3.13) and (3.22)-(3.23) we get that any $\gamma \in \mathbb{W}_{\text{dr}}$ satisfies the following restrictions:

$$\text{Target} : \frac{1}{T} \sum_{tk} \pi_k \gamma_{kt} \boldsymbol{W}_{kt} = 1, \tag{4.1}$$

$$\text{Within} - \text{unit balance} : \frac{1}{T} \sum_{t=1}^{T} \gamma_{kt} = 0, \tag{4.2}$$

$$\text{Within} - \text{period balance} : \sum_{k:\boldsymbol{W}_k \in \boldsymbol{W}^s} \pi_k \gamma_{kt} = 0, \tag{4.3}$$

$$\text{Non} - \text{negativity} : \gamma_{kt} \boldsymbol{W}_{kt} \geq 0. \tag{4.4}$$

The set $\mathbb{W}_{\text{dr}}$ treats units and periods asymmetrically. The weights need to balance arbitrary functions of S_i within each period, but only need to balance unit fixed effects for every unit. Of course, this is a direct consequence of the two-way model that we consider for the outcomes. If the underlying model is more complicated – *e.g.*, there are interactive fixed effects – then it will introduce additional restrictions. While we do not pursue such extensions in this paper, they can be addressed within our framework using the ideas from Arellano and Bonhomme [2011] and Freyberger [2018].

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 a value s for the sufficient statistic S_i such that the corresponding $\boldsymbol{W}^s$ contains at least one of the following two sub-matrices (up to permutations):

$$\boldsymbol{W}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \boldsymbol{W}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{4.5}$$

The requirement that for some s the set $\boldsymbol{W}^s$ contains at least one of these sub-matrices is in

general more demanding than the assumption that $\boldsymbol{W}$ contains one of these submatrices. It is also more demanding than the overlap restriction in Assumption 3.2. At the same time, if S_i includes $\overline{W}_i$ then for any s, $\boldsymbol{W}^s$ can contain $\boldsymbol{W}_1$ only if it contains $\boldsymbol{W}_2$ 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 3.2, or both hold. Then for any $\gamma \in \mathbb{W}_{\text{dr}}$, the estimand $\tau(\gamma)$ is a convex combination of treatment effects.*

It is important to compare this theorem with other doubly-robust results in the panel literature (*e.g.*, Sant'Anna and Zhao [2020]). The conventional interpretation of double robustness (*e.g.*, Robins and Rotnitzky [1995], Chernozhukov et al. [2016, 2018b,a]) is based on two different ways of using covariates in estimation. One can either demean the outcomes, thus making units directly comparable, or re-weight the units to guarantee that the differences between them average out. To ensure good statistical properties (*e.g.*, semiparametric efficiency), we need to combine both of these ideas. In this case, the bias of the resulting estimator depends on the product of two errors. As a result, one can trade a less accurate outcome model for more precise weighting and vice versa.

Our interpretation of double robustness is different. We have not explicitly introduced the covariates, and thus any discussion on how to use them is irrelevant for our identification result. More precisely, our approach is based on combining two different identification arguments, whereas the traditional double robustness uses a single identification assumption (a version of conditional independence). In principle, one can combine traditional double robustness with ours, but this lies outside of the scope of this paper.

4.2 Algorithm

Our estimator uses the panel data $\{Y_{it}, W_{it}, X_i\}_{i,t}$, where we now explicitly introduce time invariant covariates X_i. We assume that a researcher has constructed a sufficient statistic S_i. To incorporate covariates we consider two p-dimensional functions of (X_i, S_i, t) and (X_i, S_i): $\psi^{(1)}(X_i, S_i, t) = (\psi_1^{(1)}(X_i, S_i, t), \ldots, \psi_p^{(1)}(X_i, S_i, t))$, $\psi^{(2)}(X_i, t) = (\psi_1^{(2)}(X_i, t), \ldots, \psi_p^{(2)}(X_i, t))$, and define $\psi_{it} \equiv (\psi^{(1)}(X_i, S_i, t), \psi^{(2)}(X_i, t))$. In cases where X_i is discrete these functions can be simply set to time-specific indicators for each value of X_i and S_i.

Given these inputs, our estimator is defined in the following way:

$$\hat{\tau} := \frac{1}{NT} \sum_{it} \hat{\gamma}_{it} Y_{it} \tag{4.6}$$

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

$$\begin{aligned} \{\hat{\gamma}_{it}\}_{it} = \arg\min_{\{\gamma_{it}\}_{it}} & \frac{1}{(NT)^2} \sum_{it} \gamma_{it}^2 \\ \text{subject to: } & \frac{1}{nT} \sum_{it} \gamma_{it} W_{it} = 1, \quad \frac{1}{T} \sum_{i} \gamma_{it} = 0, \\ & \frac{1}{N} \sum_{t} \gamma_{it} = 0, \qquad \frac{1}{NT} \sum_{it} \gamma_{it} \psi_{it} = 0, \\ & \gamma_{it} W_{it} \geq 0, \end{aligned} \tag{4.7}$$

The weights $\hat{\gamma}_{it}$ are related to weights produced by OLS. The key difference is that 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]).

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 is that we need to balance unit-specific functions and explicitly impose non-negativity constraints. At the same time, we only balance a parametric class of functions of (X_i, S_i), rather than a general nonparametric class (as in Hirshberg and Wager [2017], Armstrong and Kolesár [2018a])

The weights that we get from (4.7) have the least possible norm, subject to balancing conditions motivated by Theorem 1. Our choice of the objective function is justified by statistical reasons – the variance of any linear estimator is directly related to the norm of the weights. While we do not pursue this in the current work, one can consider alternative objective functions that are motivated by economic reasons, *e.g.*, a version of empirical welfare.

5 Inference

5.1 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 Assumptions 2.1, 2.2,2.3 and additionally restrict the outcome model:

Assumption 5.1. *Either there exist a sufficient statistic S_i such that the following is true:*

$$\begin{aligned} &Y_{it}(0) = \beta_t + \psi^{(2)}(X_i, t)^\top \delta + \psi^{(1)}(X_i, S_i, t)^\top \eta + u_{it} \\ &\mathbb{E}[u_{it}|X_i, S_i] = 0 \\ &(u_{i1}, \dots u_{iT}) \perp\!\!\!\perp \underline{W}_i | X_i, S_i \end{aligned} \tag{5.1}$$

or $U_i = (\alpha_i, X_i)$ and we have the following:

$$\begin{aligned} &Y_{it}(0) = \alpha_i + \beta_t + \psi^{(2)}(X_i, t)^\top \delta + u_{it} \\ &\mathbb{E}[u_{it}|X_i, \alpha_i] = 0 \\ &(u_{i1}, \dots u_{iT}) \perp\!\!\!\perp \underline{W}_i | X_i, \alpha_i \end{aligned} \tag{5.2}$$

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. This is without loss of generality if both X_i and S_i are discrete.

5.2 Formal resuls

To state the inference results we make several statistical assumptions:

Assumption 5.2. *(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 condi-*

tions:

$$\begin{aligned} &0 < \underline{\sigma}_u^2 \leq \mathbb{E}[u_{it}^2 | \underline{W}_i, X_i] \leq \overline{\sigma}_u^2 < \infty \\ &\mathbb{E}[u_{it}^4] < \infty \end{aligned} \tag{5.3}$$

For each i, t define the following $2 \times p$-dimensional random vector:

$$\Gamma_{it} \equiv (1 - W_{it})\psi_{it} - \frac{\sum_{l=1}^{T}(1 - W_{il})\psi_{il}}{\sum_{l=1}^{T}(1 - W_{il})} \tag{5.4}$$

Assumption 5.3. *(a) S_i includes $\overline{W}_i$; (b) for all t and $\eta > 0$ we have $\mathbb{E}[W_{it}|S_i, X_i] \leq 1 - \eta$; (c) the following holds:*

$$\sigma_{\min}\left(\sum_{t=1}^{T} \mathbb{E}\left[\Gamma_{it}\Gamma_{it}^{\top}\right]\right) \geq \kappa > 0 \tag{5.5}$$

Define the target parameter:

$$\tau_{cond} = \frac{1}{NT}\sum_{it} \hat{\omega}_{it} W_{it} \mathbb{E}[\tau_{it} | \underline{W}_i, X_i] \tag{5.6}$$

This is a conditional weighted average treatment effect, where the weights are directly observed (and equal to $\hat{\omega}_{it}W_{it}$). By construction these weights are nonnegative. Next theorem shows $\hat{\tau}$ is close to τ_{cond}:

Theorem 2. *Suppose Assumptions 5.1, 5.2, 5.3 are satisfied. Then there exist a collection of random variables $\{\omega_t^\star(X_i, \underline{W}_i, t)\}_{t=1}^{T}$ such that the following holds:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{\omega}_t - \omega_t^\star\|_2 = o_p(1) \tag{5.7}$$

and the following convergence in distribution holds:

$$\sqrt{n}\left(\hat{\tau} - \tau_{cond}\right) \to \mathcal{N}(0, \sigma^2) \tag{5.8}$$

The variance has the following form:

$$\sigma^2 = \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\omega_{it}^{\star}\left(\left(u_{it} + W_{it}\left(\tau_{it} - \mathbb{E}[\tau_{it}|\underline{W}_i, X_i]\right)\right)\right)^2\right] \tag{5.9}$$

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.

To conduct inference we need to construct an estimator for σ^2. Our next results shows that conventional unit-level bootstrap can be used for this purpose.

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

$$\hat{\sigma}^2 := \frac{N}{B}\sum_{b=1}^{B}\left(\hat{\tau}_{(b)} - \hat{\tau}\right)^2 \tag{5.10}$$

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.

Theorems 2 and 3 imply that one can construct asymptotically conservative confidence intervals by standard methods. In particular, let z_α be an α-level quantile of the standard normal distribution. Then the following interval has an asymptotic coverage of at least $1 - \alpha$:

$$\tau_{cond} \in \hat{\tau} \pm \sqrt{\frac{\hat{\sigma}^2}{N}} z_{\alpha/2} \tag{5.11}$$

6 Extensions and Experiments

6.1 Non-binary treatment

In applications, the treatment W_{it} is often non-binary, and the results discussed so far are not directly applicable. One possibility is to binarize the treatment, but this process will change both the outcome and the assignment models. This section discusses an alternative strategy for dealing with a general treatment.

To proceed we need to specify the outcome model and the assignment process for general non-binary treatment W_{it}. For the outcome model we resort to the two-way linear structure:

$$\begin{aligned} &Y_{it}(w) = \alpha(U_i) + \lambda_t + \tau_t(U_i)w + \epsilon_{it} \\ &\mathbb{E}[\epsilon_{it}|U_i] = 0 \end{aligned} \tag{6.1}$$

thus abstracting away from potential non-linear effects of w. This is the standard assumption made in applications, and it does not take us far from the current empirical practice.

For the assignment model we consider a baseline distribution $f_0(w)$ that has the same support as W_{it}. If W_{it} is non-negative, then this can be an exponential distribution, if W_{it} represents counts of certain events, then $f_0(w)$ can be Poisson. We then assume that the distribution of W_i conditional on U_i belongs to the following exponential family:

$$f(W_i|U_i) = \exp\left\{\sum_t \beta^\top(U_i)\psi_t(W_{it}) - \psi(U_i)\right\}\prod_t f_0(W_{it}) \tag{6.2}$$

where $\psi_t(\cdot)$ is a known function. This structure directly generalizes the example presented in Section 3.5.1. In particular, if we observe aggregate shocks Z_t then it is natural to consider $\psi_t(W_{it}) = Z_t W_{it}$.

Exponential structure of the assignment model implies the general unconfoundedness condition:

$$W_i \perp\!\!\!\perp \{Y_i(w)\}_w | S_i \tag{6.3}$$

where $S_i = \sum_t \psi_t(W_{it})$. Given S_i we can proceed by identifying the effect by running the standard two-way regression:

$$Y_{it} = \alpha_i + \lambda_t + \tau_{it} W_{it} + \epsilon_{it} \tag{6.4}$$

withing the subpopulations defined by S_i. This approach delivers meaningful causal effects if τ_{it} does not vary in the subpopulations defined by S_i. In practice, we can split the data into clusters with similar values of S_i, run OLS separately for each cluster, and then aggregate the effects. This approach connects us with the recent work on fixed-effect models (Bonhomme and

Manresa [2015], Bonhomme et al. [2017]), where the authors argue for using K-means algorithm to classify units into clusters, as a way to estimate computationally challenging fixed-effect models. Our results shows how to use the assignment model to derive the characteristics that can be used for such classification.

The model (6.2) can be considerably generalized. Instead of the aggregate shocks, one can consider stationary dynamic models from Section 3.5.2. In general, one can use a rich class of generalized linear models (Nelder and Wedderburn [1972], Efron and Hastie [2016]) to adapt the assignment process to the particular structure of W_{it}. These models are commonly used in applied data analysis to understand complex data structures, and our results show how to exploit them for identification purposes.

If we use the normal distribution as the baseline for W_{it} then the assignment model reduces to linear regression:

$$W_{it} = \beta(U_i)^\top \psi_t + \nu_{it}$$

implying that W_{it} can be decomposed into a low-rank component $\beta(U_i)^\top \psi_t$ and idiosyncratic noise ν_{it}. This suggests that one can use interactive fixed effects regressions to estimate the treatment effects. This choice is attractive for some applications, but the panels that we have in mind have small T and large n, rendering both conventional interactive fixed effects regressions (*e.g.*, Bai [2009]) and its regularized analogs (*e.g.*, Chernozhukov et al. [2019]) inconsistent. Also, we do not restrict the persistence in the errors of the potential outcomes; thus, the GMM estimators in the spirit of Holtz-Eakin et al. [1988], Freyberger [2018] are inapplicable as well.

6.2 Empirical illustration

To illustrate our approach at work in a real application we consider data from Charles and Stephens Jr [2013]. In the paper authors analyze the relationship between the local voting preferences (expressed by turnout) and local economic outcomes (such as earnings or employment). In particular, the stylized version of the main regressions that is proposed in the paper has the following form:

$$Y_{it} = \alpha_i + \lambda_t + \tau W_{it} + \epsilon_{it} \tag{6.5}$$

where the unit of observation i corresponds to the U.S. counties, Y_{it} measures the local turnout (we will focus on the presidential elections), W_{it} measure the log-income per capita in the corresponding county. Authors estimate τ by IV, using aggregate shocks to construct instruments. In particular, their first stage model has the following form:

$$\Delta W_{it} = \theta_t + \gamma_1^\top D_{1i} \Delta Z_{1t} + \gamma_2^\top D_{2i} \Delta Z_{2t} + \nu_{it} \tag{6.6}$$

where $\{Z_{1t}, Z_{2t}\}_t$ correspond to nation-level oil and coal prices, and $D_{ki} = (D_{1ki}, D_{2ki})$ are indicators for the importance of oil and coal for the county i (medium or large). As a result, authors use the variation in ΔW_{it} that is "cleaned" from ν_{it} and θ_t to identify τ. This variation is coming from two sources: the variation in D_{ki} over i and ΔZ_t over t. The underlying identification assumption behind this approach is that the endogeneity problem arises from ν_{it} being correlated with $\Delta\epsilon_{it}$, while D_{ki} is not.

Our approach to identification is different: instead of (6.6) we consider the following first stage model:

$$W_{it} = \beta_i + \theta_t + \gamma_{1i} Z_{1t} + \gamma_{2i} Z_{2t} + \upsilon_{it} \tag{6.7}$$

and assume that $(\beta_i, \gamma_{1i}, \gamma_{2i})$ are correlated with the potential outcomes, while $\{\upsilon_{it}\}_t$ are not. Using our previous notation, we can express this in the following way:

$$U_i = (\beta_i, \gamma_{1i}, \gamma_{2i}) \tag{6.8}$$

As a result, our approach is complimentary to that of Charles and Stephens Jr [2013]. While they exploit the variation in D_{ki} – which can be viewed as a proxy for γ_{ki} – we instead control for it and exploit the variation in υ_{it}.[1]

We use (6.7) to construct the sufficient statistic for U_i:

$$S_i := \left(\sum_{t \le T} W_{it}, \sum_{t \le T} Z_{1t} W_{it}, \sum_{t \le T} Z_{1t} W_{it} \right) \tag{6.9}$$

[1] In principle one can utilize variation in Z_{kt} only, but more time periods are needed for this approach to be practically useful. See Arkhangelsky and Korovkin [2020] for more details.

	estimate	s.e.
$\hat{\tau}_{DR}$	0.003	0.007

Table 4: The results are based on the data from $n = 2994$ counties over $T = 8$ presidential elections (1972-2000). The outcome is the turnout in the presidential elections at the county level, and the treatment is the log-earnings. Sufficient statistic S_i is constructed using log(national employment) in coal and gas industries. K-means algorithm is used to split counties into $K = 1000$ groups based on S_i. Standard errors are computed using county-level bootstrap.

Note that if $\{\nu_{it}\}_t$ has normal distribution then S_i is sufficient for U_i, otherwise one can justify using S_i with the logic from Section 3.5.1. S_i is a 3-dimensional object and to control for it we use K-means algorithm to classify n units into K groups with similar values of S_i. Once these groups are defined, we proceed by estimating 6.5 by OLS with two-way fixed effects within each group. The results of the estimation are presented in Table 4. These results are qualitatively similar to those obtained by Charles and Stephens Jr [2013] who also do not find significant effects for the presidential elections.

6.3 Simulations

We use the data from Charles and Stephens Jr [2013] as a basis for a simulation. Let $\boldsymbol{Y}$ and $\boldsymbol{W}$ be $n \times T$ matrices with enties equal to Y_{it} and W_{it}, respectively. We standardize each of these matrices by subtracting the overall mean, and dividing by the overall standard deviation. We then decompose them into three components:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{F}_Y + \boldsymbol{L}_Y + \boldsymbol{E}_Y \\ \boldsymbol{W} &= \boldsymbol{F}_W + \boldsymbol{L}_W + \boldsymbol{E}_w \end{aligned} \tag{6.10}$$

where $F_k = \alpha_i^{(k)} + \lambda_t^{(k)}$ is the two-way matrix, L_k is a matrix of rank 5, and $\boldsymbol{E}_k$ captures the residual variation. We compute the size and the correlation between the residuals and elements

of matrices L_k:

$$\begin{aligned}
\sigma_k^2(E) &= \frac{\sum_{it} E_{k,it}^2}{nT} \\
\rho(E) &= \frac{\sum_{it} E_{y,it} E_{w,it}}{nT\sigma_w(E)\sigma_y(E)} \\
\sigma_k^2(L) &= \frac{\sum_{it} L_{k,it}^2}{nT} \\
\rho(L) &= \frac{\sum_{it} L_{y,it} L_{w,it}}{nT\sigma_w(L)\sigma_y(L)}
\end{aligned} \tag{6.11}$$

and then simulate the data using the following model:

$$\begin{aligned}
Y_{it} &= F_{Y,it}^{(b)} + \frac{1}{c(\zeta)} \left(\sqrt{(1-\zeta^2)} L_{Y,it}^{(b)} + \zeta \frac{\sigma_Y(L)}{\sigma_W(L)} L_{W,it}^{(b)} \right) + \epsilon_{it} \\
W_{it} &= F_{W,it}^{(b)} + L_{W,it}^{(b)} + \upsilon_{it}
\end{aligned}$$

where $(F_{Y,it}^{(b)}, L_{Y,it}^{(b)}, F_{W,it}^{(b)}, L_{W,it}^{(b)})$ are sampled uniformly from the rows of F_k, L_k, while $(\epsilon_{it}, \nu_{it})$ have a joint normal distribution with the covariance matrix implied by σ_y, σ_w, ρ. Parameter ζ controls the excess selection bias that is not present in the real data. Parameter $c(\zeta)$ normalizes this component to have the expected sum of squares equal to $\sigma_Y^2(L)$ to keep the relative sizes of the fixed effects and the low-rank component constant. We consider two designs: in the first one $\zeta = 0$, in the second it is equal to 0.05. Note that in this simulation the effect of the treatment is equal to zero, which is natural given the results presented in the previous section.

The summary of the results over 1000 simulations is presented in Table 5. We use two benchmarks: the standard two-way OLS regression and the TSLS regression implemented in the original paper. In the baseline case, our estimator and the standard TW perform equally well. Both exhibit certain bias, which is not surprising given the presence of the low-rank components and the correlation between the errors. Once we introduce the additional selection bias ($\zeta = 0.05$), the performance of TW estimator deteriorates considerably (1300% increase in RMSE), while our estimator continues to perform well (50% increase in RMSE).

We emphasize that we do not generate the treatment using the first stage described in the previous section:

$$W_{it} = \theta_t + \gamma_{1i} Z_{1t} + \gamma_{2i} Z_{2t} + \upsilon_{it}$$

	$\rho(L)$	$\rho(E)$	RMSE			Bias		
			DR	TW	TSLS	DR	TW	TSLS
Design 1 ($\zeta = 1$)	0.039	-0.038	0.016	0.016	0.521	-0.01	0.013	-0.005
Design 2 ($\zeta = 0.95$)	0.351	-0.038	0.023	0.205	0.505	0.02	0.205	0.222

Table 5: Results are based on 1000 simulations, with $n = 2994$ and $T = 8$; size of the outcome and assignment models components: $(\sigma^2_Y(L), \sigma^2_W(L), \sigma^2_Y(E), \sigma^2_W(E)) = (0.12, 0.09, 0.02, 0.06)$.

Instead, we use the actual data to extract the systematic components of W_{it}. Moreover, we do not set the correlation between υ_{it} and ϵ_{it} to zero. There are two reasons for the success of our estimator. First, the contribution of the errors $\upsilon_{it}, \epsilon_{it}$ to the corresponding outcomes is small compared to the contribution from $L_{k,it}$ and $F_{k,it}$. In particular, in both cases, the two-way fixed effects play a key role, explaining 85% of the variation in the data. This drives the good behavior of both TW and DR when $\zeta = 0$. Once we scale the correlation between $L_{W,it}$ and $L_{Y,it}$ the low-rank component starts to play a role, and TW estimator does not do anything about it. In contrast, the aggregate shocks (Z_{1t}, Z_{2t}) allow us to extract important components of the low-rank matrix and control for them.

7 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, prove that it is asymptotically normal, and show how to conduct inference of it.

References

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.

Erling Bernhard Andersen. Asymptotic properties of conditional maximum-likelihood estimators. Journal of the Royal Statistical Society: Series B (Methodological), 32(2):283–301, 1970.

Joshua Angrist and Steve Pischke. Mostly Harmless Econometrics: An Empiricists' Companion. Princeton University Press, 2008.

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of economic perspectives, 24(2):3–30, 2010.

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

Manuel Arellano and Stéphane Bonhomme. Identifying distributional characteristics in random coefficients panel data models. The Review of Economic Studies, 79(3):987–1020, 2011.

Manuel Arellano and Stéphane Bonhomme. Identifying distributional characteristics in random coefficients panel data models. The Review of Economic Studies, 79(3):987–1020, 2012.

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

Manuel Arellano, Jinyong Hahn, et al. Understanding bias in nonlinear panel models: Some recent developments. Econometric Society Monographs, 43:381, 2007.

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 and Vasily Korovkin. On policy evaluation with aggregate time-series shocks. CERGE-EI Working Paper Series, (657), 2020.

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.

Orley Ashenfelter. Estimating the effect of training programs on earnings. The Review of Economics and Statistics, pages 47–57, 1978.

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.

Jushan Bai. Panel data models with interactive fixed effects. Econometrica, 77(4):1229–1279, 2009.

Badi Baltagi. Econometric analysis of panel data. John Wiley & Sons, 2008.

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

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. Econometrica, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. University of Chicago, Becker Friedman Institute for Economics Working Paper, (2019-16), 2017.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. NBER Working Paper, (w27845), 2020.

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

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. Available at SSRN 3148250, 2019.

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.

Kerwin Kofi Charles and Melvin Stephens Jr. Employment, wages, and voter turnout. American Economic Journal: Applied Economics, 5(4):111–43, 2013.

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, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and Robins James. Double machine learning for treatment and causal parameters. arXiv preprint arXiv:1608.00060, 2016.

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.

Victor Chernozhukov, Christian Bailey Hansen, Yuan Liao, and Yinchu Zhu. Inference for heterogeneous effects using low-rank estimations. Technical report, cemmap working paper, 2019.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: mining text to track methods. Technical report, National Bureau of Economic Research, 2020.

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.

Oeindrila Dube and Juan F Vargas. Commodity price shocks and civil conflict: Evidence from colombia. The review of economic studies, 80(4):1384–1421, 2013.

Esther Duflo and Rohini Pande. Dams. The Quarterly Journal of Economics, 122(2):601–646, 2007.

Bradley Efron and Trevor Hastie. Computer Age Statistical Inference, volume 5. Cambridge University Press, 2016.

Joachim Freyberger. Non-parametric panel data models with interactive fixed effects. The Review of Economic Studies, 85(3):1824–1851, 2018.

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

Bryan S Graham and James L Powell. Identification and estimation of average partial effects in irregular correlated random coefficient panel data models. Econometrica, 80(5):2105–2152, 2012.

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

Irving Hoch. Estimation of production function parameters combining time-series and cross-section data. Econometrica: journal of the Econometric Society, pages 34–53, 1962.

Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen. Estimating vector autoregressions with panel data. Econometrica: Journal of the econometric society, pages 1371–1395, 1988.

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

Joel L Horowitz. A smoothed maximum score estimator for the binary response model. Econometrica: journal of the Econometric Society, pages 505–531, 1992.

Cheng Hsiao. Analysis of panel data. Number 54. Cambridge university press, 2014.

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 W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. Econometrica, 62(2):467–475, 1994.

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

Katrine V Løken, Magne Mogstad, and Matthew Wiswall. What linear estimators miss: The effects of family income on child outcomes. American Economic Journal: Applied Economics, 4(2):1–35, 2012.

Charles F Manski. Maximum score estimation of the stochastic utility model of choice. Journal of econometrics, 3(3):205–228, 1975.

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

Yair Mundlak. Empirical production function free of management bias. Journal of Farm Economics, 43(1):44–56, 1961.

Yair Mundlak. On the pooling of time series and cross section data. Econometrica: journal of the Econometric Society, pages 69–85, 1978.

Yair Mundlak and Irving Hoch. Consequences of alternative specifications in estimation of cobb-douglas production functions. Econometrica: Journal of the Econometric Society, pages 814–828, 1965.

Emi Nakamura and Jon Steinsson. Fiscal stimulus in a monetary union: Evidence from us regions. American Economic Review, 104(3):753–92, 2014.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3):370–384, 1972.

Nathan Nunn and Nancy Qian. Us food aid and civil conflict. American Economic Review, 104 (6):1630–66, 2014.

G Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. Technical report, National Bureau of Economic Research, 1992.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(1):122–129, 1995.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. Journal of Econometrics, 219(1):101–122, 2020.

Jeffrey M Wooldridge. Econometric analysis of cross section and panel data. MIT press, 2010.

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.

8 Appendix

8.1 Dual representation

The Lagrangian saddle-point problem for the program (4.7) has the following form:

$$\inf_{\gamma_{it}} \sup_{\lambda_{(t)},\lambda_{(i)},\eta,\mu_{it}\geq 0,\pi\geq 0} \frac{1}{(NT)^2}\sum_{it}\gamma_{it}^2 + \frac{1}{N}\sum_i \lambda_{(i)}\left(\frac{1}{T}\sum_i \gamma_{it}\right) + \frac{1}{T}\sum_t \lambda_{(t)}\left(\frac{1}{N}\sum_t \gamma_{it}\right) + \pi\left(1 - \frac{1}{NT}\sum_{it}\gamma_{it}W_{it}\right) - \eta^\top\left(\frac{1}{NT}\sum_{it}\gamma_{it}\psi_{it}\right) - \frac{1}{NT}\sum_{it}\mu_{it}\gamma_{it}W_{it} \quad \text{(A.1)}$$

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:

$$\sup_{\lambda_{(t)},\lambda_{(i)},\eta,\mu_{it}\geq 0,\pi\geq 0} \inf_{\gamma_{it}} \frac{1}{(NT)^2}\sum_{it}\gamma_{it}^2 + \frac{1}{N}\sum_i \lambda_{(i)}\left(\frac{1}{T}\sum_i \gamma_{it}\right) + \frac{1}{T}\sum_t \lambda_{(t)}\left(\frac{1}{N}\sum_t \gamma_{it}\right) - \pi\left(\frac{1}{NT}\sum_{it}\gamma_{it}W_{it} - 1\right) - \eta^\top\left(\frac{1}{NT}\sum_{it}\gamma_{it}\psi_{it}\right) - \frac{1}{NT}\sum_{it}(\mu_{it}\gamma_{it}W_{it}) \quad \text{(A.2)}$$

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

$$\inf_{\lambda_{(t)},\lambda_{(i)},\eta,\mu_{it}\geq 0,\pi\geq 0} \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\left(\pi W_{it} - \lambda_{(t)} - \lambda_{(i)} - \eta^\top\psi_{it} - \mu_{it}W_{it}\right)^2\right] - \frac{4\pi}{N} \quad \text{(A.3)}$$

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

$$\rho_z(x) := x^2(1-z) + x_+^2 z \quad \text{(A.4)}$$

After some algebra we get the following:

$$\inf_{\lambda_{(t)},\lambda_{(i)},\eta} \mathbb{P}_n \left(\frac{1}{T} \sum_{t=1}^{T} \rho_{W_{it}} \left(W_{it} - \lambda_{(t)} - \lambda_{(i)} - \eta^\top \psi_{it} \right) \right) \tag{A.5}$$

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

$$\hat{\gamma}_{it}^{(un)} = \left(W_{it} - \hat{\lambda}_{(t)} - \hat{\lambda}_{(i)} - \hat{\eta}^\top \psi_{it} \right) (1 - W_{it}) + \left(W_{it} - \hat{\lambda}_{(t)} - \hat{\lambda}_{(i)} - \hat{\eta}^\top \psi_{it} \right)_+ W_{it} \tag{A.6}$$

and the optimal weights are given by the normalization:

$$\hat{\gamma}_{it} := \frac{\hat{\gamma}_{it}^{(un)}}{\frac{1}{NT} \sum_{it} \hat{\gamma}_{it}^{(un)} W_{it}} \tag{A.7}$$

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.

8.2 Propositions

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

$$\omega_{k_{(i)}t} := \sum_{k=1}^{K} \omega_{kt} \{ \underline{W}_i = \boldsymbol{W}_k \} \tag{A.8}$$

and considered the following estimator:

$$\tau(\omega) = \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^{T} Y_{it} \omega_{k_{(i)}t} \right] \tag{A.9}$$

By assumption we have the representation:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Y_{it}\omega_{k_{(i)}t}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i)+\lambda_t+\tau(U_i)W_{it}+\varepsilon_{it})\omega_{k_{(i)}t}\right] =$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i)+\lambda_t+\tau(U_i)W_{it}+\varepsilon_{it})\sum_{k=1}^{K}\omega_{kt}\{\underline{W}_i=\boldsymbol{W}_k\}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\alpha(U_i)\omega_{kt}\{\underline{W}_i=\boldsymbol{W}_k\})\right] +$$
$$\frac{1}{T}\sum_{t=1}^{T}\lambda_t\sum_{k=1}^{K}\mathbb{E}\left[\omega_{kt}\{\underline{W}_i=\boldsymbol{W}_k\}\right] + \mathbb{E}\left[\tau(U_i)\{\underline{W}_i=\boldsymbol{W}_k\}\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\boldsymbol{W}_{kt}\omega_{kt}\right] =$$
$$\frac{1}{T}\sum_{t=1}^{T}\lambda_t\sum_{k=1}^{K}\pi_k\omega_{kt} + \mathbb{E}\left[\tau(U_i)\xi(\underline{W}_i)\right] = \mathbb{E}\left[\tau(U_i)\xi(\underline{W}_i)\right] \quad \text{(A.10)}$$

where $\xi(\underline{W}_i) := \{\underline{W}_i=\boldsymbol{W}_k\}\frac{1}{T}\sum_{k=1}^{K}\sum_{t=1}^{T}\boldsymbol{W}_{kt}\omega_{kt} \geq 0$. 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:

$$\mathbb{E}[\{\underline{W}_i=\underline{w}\}\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|S_i] = \mathbb{E}\{\underline{W}_i=\underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|S_i] \quad \text{(A.11)}$$

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

$$\mathbb{E}[\{\underline{W}_i=\underline{w}\}\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|S_i] =$$
$$\mathbb{E}[\{\underline{W}_i=\underline{w}\}\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|S_i,U_i,\underline{W}_i]|S_i] =$$
$$\mathbb{E}[\{\underline{W}_i=\underline{w}\}\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|U_i,S_i]|S_i] =$$
$$\mathbb{E}\mathbb{E}[\{\underline{W}_i=\underline{w}\}|S_i,U_i]\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|U_i,S_i]|S_i] =$$
$$\mathbb{E}[\mathbb{E}[\{\underline{W}_i=\underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|U_i,S_i]|S_i] =$$
$$\mathbb{E}\{\underline{W}_i=\underline{w}\}|S_i]\mathbb{E}[\{\underline{Y}_i(0)\in A_0,\underline{Y}_i(1)\in A_1\}|S_i] \quad \text{(A.12)}$$

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

8.3 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:*

$$\begin{aligned} &\alpha_i + \beta_t + \psi_{it}^\top \gamma \geq 0 \\ &W_{it} = \{\alpha_i + \beta_i + \psi_{it}^\top \gamma > 0\} \end{aligned} \tag{A.13}$$

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

$$\begin{aligned} h(\lambda, \mu, \pi, \gamma, \omega) := \frac{1}{(nT)^2} \sum_{it} \omega_{it}^2 + \frac{1}{n} \sum_i \lambda_{(i)} \left(\frac{1}{T} \sum_i \omega_{it} \right) + \\ \frac{1}{T} \sum_t \lambda_{(t)} \left(\frac{1}{n} \sum_t \omega_{it} \right) + \pi \left(1 - \frac{1}{nT} \sum_{it} \omega_{it} W_{it} \right) - \\ \gamma^\top \left(\frac{1}{nT} \sum_{it} \omega_{it} \psi_{it} \right) - \frac{1}{nT} \sum_{it} \mu_{it} \omega_{it} W_{it} \end{aligned} \tag{A.14}$$

we have

$$\inf_{\omega_{it}} \sup_{\lambda_{(t)}, \lambda_{(i)}, \gamma, \mu_{it} \geq 0, \pi \geq 0} h(\lambda, \mu, \pi, \gamma, \omega) = \sup_{\lambda_{(t)}, \lambda_{(i)}, \gamma, \mu_{it} \geq 0, \pi \geq 0} \inf_{\omega_{it}} h(\lambda, \mu, \pi, \gamma, \omega) \tag{A.15}$$

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:

$$\begin{aligned} &\alpha_i^\star + \beta_t^\star + \psi_{it}^\top \gamma^\star \geq 0 \\ &W_{it} = \{\alpha_i^\star + \beta_t^\star + \psi_{it}^\top \gamma^\star > 0\} \end{aligned} \tag{A.16}$$

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:*

$$g(X, \underline{W}, \gamma) \in \arg\min_{\alpha} \left\{ \frac{1}{T} \sum_{t=1}^{T} \rho_{W_t}(W_t - \alpha - \psi_t^\top \gamma) \right\} \tag{A.17}$$

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 $\tilde{h}_{(1)}, \ldots, \tilde{h}_{(\sum_{t=1}^T W_t)}$ be the decreasing ordering of h_t for units with $W_t = 1$; let $\tilde{h}_{(0)} = 0$. For $k = 0, \ldots, \sum_{t=1}^T W_t$ define the following functions:

$$g_k(X, \underline{W}, \gamma) := \frac{\sum_{t=1}^T (1 - W_{it}) h_t + \sum_{l=0}^k \tilde{h}_{(l)}}{\sum_{t=1}^T (1 - W_{it}) + k} \tag{A.18}$$

It is easy to see that we have the following:

$$g(X, \underline{W}, \gamma) = g_0(X, \underline{W}, \gamma) + \sum_{l=1}^{k} \{\tilde{h}_{(l)} \geq g_{(l-1)}\} (g_l(X, \underline{W}, \gamma) - (g_{l-1}(X, \underline{W}, \gamma)) \tag{A.19}$$

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:*

$$\begin{aligned}
&\xi_{it} := W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \\
&\mathbb{E}\left[\sum_{t=1}^{T} \xi_{it} \psi_{it} (1 - W_{it} \{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \leq 0\}) \right] = 0 \\
&\sum_{t=1}^{T} \xi_{it} (1 - W_{it} \{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \leq 0\}) = 0
\end{aligned} \tag{A.20}$$

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 $\mathcal{G} := \{g = (g_1, \ldots g_T) : g_t = f + \psi_t^\top \gamma, f \in L_2(\mathcal{P}), \gamma \in \mathbb{R}^p\}$.

Consider the following optimization program:

$$\inf_{g\in\mathcal{G}} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\rho_{W_{it}}(W_{it}-g_{it})\right] \tag{A.21}$$

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:

$$(h^\star, f^\star + t(g_\varepsilon^0 - f^\star)) =: a + bt \tag{A.22}$$

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_ε^0. 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 that $R(f)$ has a unique minimum at $f^\star$. Consider $f^\star$ such that $f_t^\star := \mathbb{E}[W_{it}|S_i, X_i]$. Because S_i includes $\overline{W}_i$ it follows that $\frac{1}{T}\sum_{t=1}^{T} f_t^\star = \overline{W}_i$. Take any function $f \in \mathcal{F}$ and consider a convex

combination $f(\lambda) := f^\star + \lambda(f - f^\star)$. Because $f_t \in L_\infty(\mathcal{P})$ and $f_t^\star \leq 1 - \eta$ it follows that for all $\lambda < \lambda_0$ we have $f_t(\lambda) < 1$ almost surely. For any $\lambda < \lambda_0$ we have that $R(f(\lambda)) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (W_t - f_t^\star)^2\right] + \mathbb{E}\left[\sum_{t=1}^T (f_t^\star - f_t(\lambda))^2\right] > R(f^\star)$. 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. $\square$

8.4 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:

$$\hat{\tau} = \frac{1}{nT}\sum_{it}\hat{\omega}_{it}Y_{it} = \frac{1}{nT}\sum_{it}\hat{\omega}_{it}\tau_{it}W_{it} + \frac{1}{nT}\sum_{it}\hat{\omega}_{it}u_{it} = \tau_{emp} + \frac{1}{nT}\sum_{it}\hat{\omega}_{it}u_{it} = \tau_{emp} + \frac{1}{\mathbb{P}_n\frac{1}{T}\sum_{t=1}^{T}\hat{\omega}^{un}_{it}W_{it}}\left(\frac{1}{nT}\sum_{it}(\omega^\star_{it})^{un}u_{it} + \frac{1}{nT}\sum_{it}(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}\right) \quad (A.23)$$

By construction and assumption we have the following:

$$\mathbb{E}[(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}|\{\underline{W}_j, X_j\}_{j=1}^n] = (\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})\mathbb{E}[u_{it}|\{\underline{W}_j, X_j\}_{j=1}^n] = (\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})\mathbb{E}[u_{it}|\underline{W}_i, X_i] = 0 \quad (A.24)$$

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

$$\zeta_n(\epsilon) := \mathbb{E}\left[\left\{\sqrt{n}\left|\mathbb{P}_n\frac{1}{T}\sum_{t=1}^{T}(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un})u_{it}\right| \geq \epsilon\right\}|\{\underline{W}_j, X_j\}_{j=1}^n\right] \leq \frac{\mathbb{P}_n\mathbb{E}\left[\left(\sum_{t=1}^{T}(\hat{\omega}^{un}_{it} - (\omega^\star_{it})^{un}\right)^2|\{\underline{W}_j, X_j\}_{j=1}^n\right]}{T^2\epsilon^2} \leq \frac{\overline{\sigma}^2_u}{T\epsilon^2}\|(\omega^\star)^{un} - \hat{\omega}^{un}\|_2^2 = o_p(1) \quad (A.25)$$

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

$$\mathbb{E}[\zeta_n(\epsilon)] = o(1) \quad (A.26)$$

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:

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}(\omega^\star_{it})^{un}u_{it}\right)^2\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[((\omega^\star_{it})^{un}u_{it})^2\right] \leq \sum_{t=1}^{T}\sqrt{\mathbb{E}[u^4_{it}]\mathbb{E}[((\omega^\star_{it})^{un})^4]} < \infty \quad (A.27)$$

Finally, define:

$$\omega_{it}^{\star} := \frac{(\omega_{it}^{\star})^{un}}{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\omega_{it}^{\star})^{un}W_{it}\right]} \tag{A.28}$$

It is easy to see that we have:

$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T}\hat{\omega}_{it}^{un}W_{it} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\omega_{it}^{\star})^{un}W_{it}\right] + o_p(1) \tag{A.29}$$

and thus we have the following:

$$\begin{aligned} &\|\omega^{\star} - \hat{\omega}\|_2 = o_p(1) \\ &\sqrt{n}(\hat{\tau} - \tau_{emp}) \to \mathcal{N}(0, \sigma_{\tau}^2) \end{aligned} \tag{A.30}$$

which concludes the first part.

Part 2: In this part we prove that $\|(\omega^{\star})^{un} - \hat{\omega}^{un}\|_2 = o_p(1)$, $(\omega_{it}^{\star})^{un}$ is uniformly bounded, and $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\omega_{it}^{\star})^{un}W_{it}\right] > 0$. We use the dual representation derived in Section 8.1 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 $(\{\alpha_i^{\star}\}_{i=1}^{n}, \gamma^{\star})$ that satisfy the following restrictions:

$$\begin{aligned} &\xi_{it} := W_{it} - \alpha_i^{\star} - \psi_{it}^{\top}\gamma^{\star} \\ &\mathbb{E}\left[\sum_{t=1}^{T}\xi_{it}\psi_{it}(1 - W_{it}\{W_{it} - \alpha_i^{\star} - \psi_{it}^{\top}\gamma^{\star} \le 0\})\right] = 0 \\ &\sum_{t=1}^{T}\xi_{it}(1 - W_{it}\{W_{it} - \alpha_i^{\star} - \psi_{it}^{\top}\gamma^{\star} \le 0\}) = 0 \end{aligned} \tag{A.31}$$

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:

$$\begin{aligned}
&\rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^\top\gamma) = (W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i - \psi_{it}^\top\gamma \leq 0\}\right) = \\
&(W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right) + \\
&(W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 W_{it} \left(\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\} - \{W_{it} - \alpha_i - \psi_{it}^\top\gamma \leq 0\}\right) \geq \\
&(W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right) - \\
&\qquad\qquad (W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 W_{it}\{\alpha_i^\star + \psi_{it}^\top\gamma^\star < 1 \leq \alpha_i + \psi_{it}^\top\gamma\} \quad \text{(A.32)}
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{t=1}^T \left(\rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^\top\gamma) - \rho_{W_{it}}(W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star)\right) \geq \\
&\sum_{t=1}^T \left((\alpha_i^\star - \alpha_i) + \psi_{it}^\top(\gamma^\star - \gamma))^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right)\right) + \\
&\sum_{t=1}^T \left(\xi_{it}(\alpha_i^\star - \alpha_i^\star) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right)\right) + \\
&\sum_{t=1}^T \left(\xi_{it}\psi_{it}^\top(\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right)\right) - \\
&\sum_{t=1}^T \left((W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 W_{it}\{\alpha_i^\star + X_i^\top\gamma^\star < 1 \leq \alpha_i + \psi_{it}^\top\gamma\}\right) = \\
&\sum_{t=1}^T \left((\alpha_i^\star - \alpha_i) + \psi_{it}^\top(\gamma^\star - \gamma))^2 \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right)\right) + \\
&\sum_{t=1}^T \left(\xi_{it}\psi_{it}^\top(\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top\gamma^\star \leq 0\}\right)\right) - \\
&\qquad\qquad \sum_{t=1}^T \left((W_{it} - \alpha_i - \psi_{it}^\top\gamma)^2 W_{it}\{\alpha_i^\star + \psi_{it}^\top\gamma^\star < 1 \leq \alpha_i + \psi_{it}^\top\gamma\}\right) \quad \text{(A.33)}
\end{aligned}$$

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:

$$\alpha_i^\star = g(X_i, W_i, \gamma^\star) \tag{A.34}$$

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

$$\alpha_i = g(X_i, W_i, \gamma) \tag{A.35}$$

Define:

$$\begin{aligned} f_{it} &= \alpha_i + \psi_{it}^\top \gamma \\ f_{it}^\star &= \alpha_i^\star + \psi_{it}^\top \gamma^\star \end{aligned} \tag{A.36}$$

and observe that we have the following:

$$\begin{aligned} \mathbb{P}_n \sum_{t=1}^T (1 - W_{it}\{W_{it} < f_{it}^\star\})(f_{it} - f_{it}^\star)^2 \geq \mathbb{P}_n \sum_{t=1}^T (1 - W_{it})(f_{it} - f_{it}^\star)^2 \geq \\ (\gamma - \gamma^\star)^\top \left(\sum_{t=1}^T \mathbb{P}_n \Gamma_{it} \Gamma_{it}^\top \right) (\gamma - \gamma^\star) = \kappa \|\gamma - \gamma^\star\|_2^2 + o_p(\|\gamma - \gamma^\star\|_2^2) \end{aligned} \tag{A.37}$$

where

$$\Gamma_{it} := (1 - W_{it})\psi_{it} - \frac{\sum_{l=1}^T (1 - W_{il})\psi_{il}}{\sum_{l=1}^T (1 - W_{il})} \tag{A.38}$$

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

$$\begin{aligned} \mathbb{P}_n \xi_{it} \psi_{it}^\top (\gamma^\star - \gamma) \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \leq 0\}\right) \leq \\ \|\gamma^\star - \gamma\|_2 \times \left\| \mathbb{P}_n \xi_{it} \psi_{it} \left(1 - W_{it}\{W_{it} - \alpha_i^\star - \psi_{it}^\top \gamma^\star \leq 0\}\right) \right\|_2 \end{aligned} \tag{A.39}$$

We also have the following inequality:

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\alpha_i-\psi_{it}^\top\gamma)^2W_{it}\{\alpha_i^\star+\psi_{it}^\top\gamma^\star<1\leq\alpha_i+\psi_{it}^\top\gamma\}\right]\leq$$
$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(f_{it}^\star-f_{it})^2\{f_{it}^\star<1\leq f_{it}\}\right]\leq\|f^\star-f\|_\infty^2\times\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}\}\right] \quad \text{(A.40)}$$

where the first implication follows because of the indicator, and the the second one follows by Holder inequality. Since $\|f^\star-f\|_\infty\leq C_2r$ we have the following:

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}\}\right]\leq\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2r\}\right] \quad \text{(A.41)}$$

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

$$\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2r\}\right]\leq\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{f_{it}^\star<1\leq f_{it}^\star+C_2r\}\right]+\frac{C_3}{\sqrt{n}} \quad \text{(A.42)}$$

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:

$$\{\alpha_i^{(b)}, \gamma^{(b)}\}_{i=1}^n = \arg\min \mathbb{P}_n M_i \frac{1}{T} \sum_{t=1}^T \rho_{W_{it}}(W_{it} - \alpha_i - \psi_{it}^T \gamma) \tag{A.43}$$

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 $\{M_i\}_{i=1}^n$:

$$\begin{aligned} &\mathbb{P}_n M_i X_i = \mathbb{E}[X_i] + o_p(1) \\ &\mathbb{P}_n M_i \varepsilon_i = O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.44}$$

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:

$$\begin{aligned} &\mathbb{P}_n M_i \left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] = \mathbb{P}_n (M_i - 1) \left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] + \\ &\mathbb{P}_n \left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\} - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right]\right] + \\ &\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{f_{it}^\star < 1 \le f_{it}^\star + C_2 r\}\right] + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.45}$$

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:

$$\begin{aligned} &\|\omega^{(b)} - \omega^\star\|_\infty = o_p(1) \\ &\|\omega^{(b)} - \omega^\star\|_2 = O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned} \tag{A.46}$$

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

$$\hat{\tau}^{(b)} - \hat{\tau} = \mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it}\tau_{it} W_{it} +$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} u_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} u_{it} =$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} (\tau_{it} - \mathbb{E}[\tau_{it}]) W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} (\tau_{it} - \mathbb{E}[\tau_{it}]) W_{it} +$$
$$\mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} u_{it} + o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.47)}$$

From this representation it follows that if $\tau_{it} = const$ 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:

$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it}\tau_{it} W_{it} =$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \tau_t(\underline{W}_i, X_i) W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \tau_t(\underline{W}_i, X_i) W_{it} +$$
$$\mathbb{P}_n M_i \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{(b)} \eta_{it} W_{it} - \mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} \hat{\omega}_{it} \eta_{it} W_{it} =$$
$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} (M_i \omega_{it}^{(b)} - \hat{\omega}_{it}) \tau_t(\underline{W}_i, X_i) W_{it} + \mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} \eta_{it} W_{it} + o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{(A.48)}$$

It follows that we have the following:

$$\hat{\tau}^{(b)} - \hat{\tau} = \mathbb{P}_n (M_i - 1) \frac{1}{T}\sum_{t=1}^{T} \omega_{it}^{\star} (\eta_{it} W_{it} + u_{it}) +$$
$$\mathbb{P}_n \frac{1}{T}\sum_{t=1}^{T} (M_i \omega_{it}^{(b)} - \hat{\omega}_{it}) \tau_t(\underline{W}_i, X_i) W_{it} + \text{small order terms} \quad \text{(A.49)}$$

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