The Role of the Propensity Score in Fixed Effect Models*

Dmitry Arkhangelsky † Guido W. Imbens‡

September 2018

Abstract

We develop a new approach for estimating average treatment effects in the observational studies with unobserved cluster-level heterogeneity. The previous approach relied heavily on linear fixed effect specifications that severely limit the heterogeneity between clusters. These methods imply that linearly adjusting for differences between clusters in average covariate values addresses all concerns with cross-cluster comparisons. Instead, we consider an exponential family structure on the within-cluster distribution of covariates and treatments that implies that a low-dimensional sufficient statistic can summarize the empirical distribution, where this sufficient statistic may include functions of the data beyond average covariate values. Then we use modern causal inference methods to construct flexible and robust estimators.

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

*We are grateful for comments by participants in the Harvard-MIT econometrics seminar, the SIEPR lunch at Stanford, the International Association of Applied Econometrics meeting in Montreal, Pat Kline, and Matias Cattaneo. We are also grateful to Greg Duncan for raising questions that this paper tries to answer. This research was generously supported by ONR grant N00014-17-1-2131.

†Assistant Professor, CEMFI, darkhangel@cemfi.es.

‡Professor of Economics, Graduate School of Business, Stanford University, SIEPR, and NBER, imbens@stanford.edu.

1 Introduction

Suppose a researcher is interested in the average causal effect of a binary treatment in a setting where the population of interest is partitioned into a number of subpopulations, clusters or strata. The potential outcome distributions, both marginal and conditional on observed pretreatment variables, as well as the marginal and conditional treatment assignment probabilities, may differ between the strata. The researcher has available a random sample of units from a randomly selected subset of clusters. A popular estimation strategy in such settings is fixed effect regression where the differences between the strata are assumed to be fully accounted for by additive stratum-specific components in the regression function. Under this approach, all clusters are assumed to be comparable once these additive stratum-specific components are removed. However, especially in settings with additional covariates (pretreatment) variables, the additivity and linearity assumptions imposed by such fixed effect methods impose strong, possibly undesirable, conditions on the relationship between the potential outcome distributions in the different strata.

In this paper, we develop methods for this setting that allow for heterogeneity between clusters beyond the additive component. This will imply that we cannot simply compare treated and control units in any pair of clusters once we remove the additive component. Instead only treated and control units in "similar" clusters are comparable. The key will be in constructing measures of similarity between clusters. If we wish to be very flexible in the amount of heterogeneity between clusters we allow for, we end up in the extreme case where only units within the same cluster are comparable. However, if there are few units per cluster in the sample we may in that case not be able to adjust flexibly for differences between treated and control units in terms of fixed pretreatment variables. We, therefore, may need to balance the desire to compare treated and control units in similar clusters and the desire to compare only treated and control units with similar covariates.

One important implication of the functional-form free nature of our approach is that it allows us to identify any moments of the marginal distributions of the potential outcomes in a unified way. In particular, our approach can be used to estimate quantile treatment effects as well as other distributional effects. These quantities can not be estimated using standard fixed effect regressions.

To relax the functional assumptions, we start by first noting that, by omitted variable bias

arguments, biases arise from differences in the conditional distributions of the potential outcomes in different strata only if the conditional assignment probabilities differ by stratum. In fact, it is obvious that, if the treatment is completely randomly assigned, one need not be concerned with systematic differences in potential outcome distributions between strata, additive or not. One can, therefore, remove biases from comparisons between treated and control units in the same or different clusters by ensuring that treatment/control comparisons are between units with identical assignment probabilities (identical values for the propensity score). We propose to exploit this idea, well known in the evaluation literature since Rosenbaum and Rubin [1983] by modeling and estimating the conditional assignment probabilities in the different strata. Using this strategy, we can relax the functional form assumptions on the potential outcome distributions substantially. The complication in following this approach is that in settings with few sampled units per cluster we cannot consistently estimate the population assignment probabilities for each unit. We address this problem by imposing an exponential family structure on the joint distribution of treatments and covariates within a cluster to achieve consistency of the estimator for the average treatment effect under asymptotic sequences with a fixed number of sampled units per cluster.

In the fixed effect approach units in different clusters are directly comparable once we remove the additive fixed component. In our approach, the differences are more complex, and we rely on comparing units in similar clusters, requiring us to define smooth measures of the distance between clusters. Another alternative that also allows for heterogeneity between clusters beyond additive components is to assume that sets of clusters are similar and use the data to identify such sets, e.g., Bonhomme and Manresa [2015].

Although we focus in the current paper on a cross-section setting with clusters, as in Altonji and Mansfield [2014], the issues raised here are also relevant to proper panel or longitudinal data settings (Hsiao et al. [2012], Altonji and Matzkin [2005], Chamberlain [1984], Pesaran [2006], Arellano and Honoré [2001], Abadie [2005], Bertrand et al. [2004], as we discuss in Section 6.1. In that literature the paper fits into a recent set of studies Abadie et al. [2010], de Chaisemartin and D'Haultfœuille [2018], Bonhomme and Manresa [2015], Imai and Kim [2016], Athey and Imbens [2018] that connects more directly with the causal (treatment effect) literature than the earlier panel data literature by allowing for general heterogeneity beyond additive effects.

2 Fixed Effect versus Propensity Score Methods

In this section we set up the problem and introduce the notation. We then discuss fixed effect regression, and state assumptions on the potential outcome distributions that justify this estimation strategy. Next, we contrast these with estimation methods from the program evaluation literature under unconfoundedness.

2.1 The Set Up

Using the potential outcome set up (e.g., Imbens and Rubin [2015]), we consider a set up with a large, possibly infinite, population of units, characterized by a pair of potential outcomes $(Y_i(0), Y_i(1))$, and a K-component vector of pretreatment variables X_i. The population is partitioned into strata or clusters, with C_i indicating the stratum or cluster unit i is a member of. The number of strata in the population is large, and so is the number of units per cluster. We are interested in the average treatment effects. Ideally we might wish to estimate the population average effect,

$$\tau := \mathbb{E}[Y_i(1) - Y_i(0)],$$

but this may be challenging, and we may need to settle for some other average of $Y_i(1) - Y_i(0)$, *e.g.*, the average over some subpopulation defined in terms of clusters, covariates and assignments. Unit i receives treatment $W_i \in \{0, 1\}$. We first randomly sample C clusters, and then draw a random sample of size N from the subpopulation defined by the sampled clusters. For the sampled units we observe the quadruple (Y_i, W_i, X_i, C_i), $i = 1, \ldots, N$, where $Y_i := Y_i(W_i)$ is the realized outcome, that is, the potential outcome corresponding to the treatment received, and $C_i \in \{1, \ldots, C\}$ is the cluster label for unit i. Also define $C_{ic} = \mathbf{1}_{C_i=c}$ as the binary cluster indicators, and let $N_c := \sum_{i=1}^{N} C_{ic}$ be the number of sampled units in stratum c. For any variable Z_i, let $\overline{Z}_c := \sum_{i:C_i=c} Z_i/N_c$ be the corresponding cluster average in cluster c.

In the settings we are interested in the number of strata or clusters in the sample, C, may be substantial, on the order of hundreds or even thousands. The dimension of X_i may be modest. The number of units in the population in each cluster is large, but we observe only few units in each stratum, possibly as few as two or three. As a result methods that rely on accurate estimation of features of the population distribution of potential outcomes or treatments

conditional on covariates within clusters may have poor properties.

2.2 Fixed Effect Regression

A common approach to estimating the causal effect of W_i in this setting is to use a fixed effect regression (e.g., Hsiao [2014], Arellano [2003], Chamberlain [1984], Angrist and Pischke [2008], Wooldridge [2010]). Here the regression function is specified as

$$Y_i = \alpha_{C_i} + W_i\tau + X_i^\top\beta + \varepsilon_i = \sum_{c=1}^{C} C_{ic}\alpha_c + W_i\tau + X_i^\top\beta + \varepsilon_i, \tag{2.1}$$

with τ the object of interest, and β and the α_c nuisance parameters. The parameters (β, τ) and the fixed effects α_c, for $c = 1, \ldots, C$, are then estimated by least squares:

$$\left(\hat{\alpha}_c^{\text{fe}}, \hat{\beta}^{\text{fe}}, \hat{\tau}^{\text{fe}}\right) := \arg\min_{\alpha_c,\tau,\beta} \sum_{i=1}^{N} \left(Y_i - \sum_{c=1}^{C} C_{ic}\alpha_c - W_i\tau - X_i^\top\beta\right)^2. \tag{2.2}$$

This set up is the starting point of the discussion in this paper. The fixed effect specification and corresponding estimator are widely used in the empirical literature. In the absence of the fixed effects the concern is that comparisons of treated and control units, say based on least squares regression using the same specification of the regression function other than the omission of the fixed effects,

$$Y_i = \alpha + W_i\tau + X_i^\top\beta + \varepsilon_i,$$

would not have a credible causal interpretation.

In typical applications the number of strata is substantial, the dimension of the covariates is modest, and the number of sampled units per stratum is modest. Asymptotic approximations are often based on the number of strata increasing proportional to the number of sampled units, so that the average number of sampled units per stratum converges to a finite limit. This leads to the incidental parameters problem (Neyman and Scott [1948], Bonhomme [2012], Bonhomme and Manresa [2015], Arellano and Hahn [2006], Hsiao [2014], Hahn and Newey [2004]). However, the incidental parameter problem does not create complications for estimators of all parameters, and in this case it does not compromize our ability to get a consistent estimator for τ.

To motivate the paper, we first present a set of assumptions that justify the fixed effects estimator $\hat{\tau}^{\text{fe}}$ as an estimator for the causal effect τ, and then discuss some concerns with these assumptions. Note that we are not concerned with the properties of $\hat{\alpha}_c^{\text{fe}}$ or $\hat{\beta}^{\text{fe}}$, solely with the estimator $\hat{\tau}^{\text{fe}}$ for the treatment effect. In fact there are no consistent estimators for α_c under asymptotic sequences of the type we consider where the number of units per stratum remains finite. To formally justify the fixed effect estimator for τ we can make the following assumptions. First, unconfoundedness, which implies that the comparison of treated and control units within the same stratum, and with the same value for the pretreatment variables, has a causal interpretation:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, C_i. \tag{2.3}$$

The second assumption adds functional form restrictions. Define the within-stratum conditional expectation of the potential outcomes given X_i:

$$\mu_w(x, c) = \mathbb{E}[Y_i(w)|X_i = x, C_i = c]. \tag{2.4}$$

Then assume

$$\mu_w(x, c) := \alpha_c + x^\top \beta + w\tau. \tag{2.5}$$

Both the unconfoundedness assumption and the functional form assumption are strong and often controversial. In the current discussion however, we focus on the functional form assumption and maintain the unconfoundedness assumption. For discussions of the unconfoundedness assumption the reader is referred to the general treatment effect literature, e.g., Imbens and Rubin [2015], Morgan and Winship [2014]. There are multiple concerns with the functional form assumption. First, the strata may differ not only in the level of the outcome, but also in the response to the treatment, so that the constant treatment effect assumption is violated. The strata may also differ in the association between other covariates and the outcome, so that the additivity and linearity assumptions are violated. Under the fixed effect assumptions treated units in a cluster with 90% of the units treated can be compared to control units in a cluster with 10% of the units treated, as long as we remove the corresponding α_c. Such strong im-

plications may be unrealistic, and it may be more reasonable to compare treated and control units in clusters that are similar. The question is how to define and operationalize the notion of similarity of clusters in a setting with few sampled units per cluster.

Note that in the case without pretreatment variables these functional form assumption matter substantially less. In the absence of covariates, the fixed effects estimator is no longer unbiased for the average effect of the treatment if the effects of the treatment differ between strata. Nevertheless, the fixed effects estimator does estimate a weighted average of the within-stratum average effects, with the weights equal to the inverse of the within-stratum variances, and so that it has a meaningful causal interpretation (e.g., Angrist and Pischke [2008]). However, this result does not extend to the case with covariates that enter into a nonlinear or interactive way if the specification of the regression function is linear. In that case the fixed effect estimators may have substantial bias if the functional form assumption is violated, and the sign of the probability limit of the fixed effect estimator can be of the opposite sign, even if all the within-cluster average treatment effects are the same sign.

Relaxing the functional form is not straightforward. Even parametric extensions, by, for example, allowing for separate τ and β by stratum may be difficult to implement with commonly available data. For example, Chernozhukov et al. [2013] develop methods that allow the unobserved component α_c to enter the conditional expectation of Y_i in a nonlinear manner.

2.3 Propensity Score Methods

In contrast to the fixed effect literature, functional form assumptions are often avoided in the treatment effect literature (see Abadie and Cattaneo [2018] and Imbens and Wooldridge [2009] for recent surveys), by using more flexible estimators. Many of the recommended estimators in that literature go beyond estimating the conditional expectation of the outcomes given treatment and covariates. The concern with estimators that rely only on estimating the conditional expectation of the outcomes is that they are often sensitive to the specific estimation method employed, in particular in settings where the covariate distributions differ substantially between treatment groups. In the current setting, this would correspond to a concern that the probability of receiving the active treatment may differ substantially by both strata and covariates. The estimators that are recommended in that literature involve in some fashion or another estimating

the propensity score,

$$e(x,c) := \text{pr}(W_i = 1 | X_i = x, C_i = c).$$

Estimators in this literature include weighting on the inverse of the propensity score (Hirano et al. [2003]), matching on the propensity score (Abadie and Imbens [2016]), or blocking on the propensity score, often of them in combination with direct regression adjustment through doubly robust methods (Robins and Rotnitzky [1995]). One specific approach is to use the influence function (Bickel et al. [1998], Chernozhukov et al. [2016]). In the current case that would correspond to estimating $e(\cdot)$ and $\mu_w(\cdot)$, and then estimate τ as

$$\hat{\tau}^{\text{eif}} := \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\mu}_1(X_i,C_i) - \hat{\mu}_0(X_i,C_i) + W_i\frac{Y_i - \hat{\mu}_1(X_i,C_i)}{\hat{e}(X_i,C_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_0(X_i,C_i)}{1-\hat{e}(X_i,C_i)}\right\}.$$

In settings with a substantial number of units per cluster, we can directly implement these ideas (e.g., Yang [2016]).

The main issue with this approach in the current setting is that it relies on the number of sampled units per cluster being sufficiently large so that we can estimate the conditional potential outcome means and the propensity score with sampling error going to zero. This is not in the spirit of the fixed effects literature where, by differencing out the fixed effects, one can obtain consistent estimates of the parameters of interest in settings with a small number of units per cluster, sometimes as few as two. One approach would be to use some of the recent methods that allow for high-dimensional covariates (e.g., Farrell [2015], Athey et al. [2016], Chernozhukov et al. [2016]), and use the cluster indicators simply as additional control variables. The structure of the control variables, with the cluster indicator partitioning the population in many subpopulations, may prevent such methods from being effective.

In this paper we propose a new approach that allows us to exploit insights from the propensity score literature in settings with a finite number of sampled units per cluster. In fact, this number may be as small as two. We do so by imposing structure on the joint distribution of the treatment assignment and covariates in the clusters, so that we can characterize the propensity score as function of only a small number of individual characteristics.

2.4 An Alternative Representation of the Fixed Effect Estimator

To motivate our approach it is useful to observe is that we can characterize the fixed effect estimator in an alternative way. Using the notation for cluster averages of Y_i, W_i, and X_i respectively,

$$\overline{Y}_c := \frac{1}{N_c}\sum_{i:C_i=c} Y_i, \qquad \overline{W}_c := \frac{1}{N_c}\sum_{i:C_i=c} W_i, \qquad \overline{X}_c := \frac{1}{N_c}\sum_{i:C_i=c} X_i,$$

we can write the fixed effect estimator in a different way.

Lemma 1. (An Alternative Representation of the Fixed Effect Estimator - Mundlak [1978]) *Consider the regression*

$$Y_i = \alpha + W_i\tau + X_i^\top\beta + \overline{W}_{C_i}\delta + \overline{X}_{C_i}^\top\gamma + \varepsilon_i, \tag{2.6}$$

with the least squares estimates defined as

$$\left(\hat{\alpha}^{\text{ca}}, \hat{\delta}^{\text{ca}}, \hat{\gamma}^{\text{ca}}, \hat{\tau}^{\text{ca}}, \hat{\beta}^{\text{ca}}\right) := \arg\min_{\alpha,\delta,\gamma,\tau,\beta} \sum_{i=1}^{N}\left(Y_i - \alpha - W_i\tau - X_i^\top\beta - \overline{W}_{C_i}\delta - \overline{X}_{C_i}^\top\gamma\right)^2,$$

(where the superscript "ca" stands for cluster averages). Then:

$$\hat{\tau}^{\text{ca}} = \hat{\tau}^{\text{fe}}.$$

The proof can be found in Appendix B.1.

Comment 1: This alternative representation of the fixed effects estimator, mentioned in passing in Mundlak [1978], can be derived easily using omitted variable bias expressions. We state it as a formal result merely to facilitate the interpretation of the novel results below. The intuition for the equivalence of the two regressions is that bias from omitting cluster indicators comes from a non-zero effect of the stratum indicator, in combination with a correlation between the stratum indicator and the treatment indicator. If we have two clusters with the same distribution of conditional treatment probabilities, there is no bias from combining them into a single cluster. □

Comment 2: Altonji and Mansfield [2014] also use cluster averages as a method for controlling

for unobserved cluster differences. Their focus is on linear methods and the identifying power of such regressions in the context of structural models. □

COMMENT 3: These two representations show that the adjusting for differences between the clusters takes a very simple form, exploiting the additivity of the regression function. Instead of including indicators C_{ic} for the cluster, we can include the cluster average covariates, or subtract cluster averages from outcomes and covariates. Suppose we view (2.6) as a regression version of an attempt to compare treated and control units after adjusting for all differences between the cluster in $(X_i, \overline{W}_{C_i}, \overline{X}_{C_i})$. Then the program evaluation literature under unconfoundedness would suggest that if the distribution of these variables were substantially different for treated and control units, simple least squares regression might not be a reliable way of adjusting for these differences. Instead, more sophisticated methods of adjusting for such differences, involving the propensity score, and relaxing linearity and additivity of the regression function in $(X_i, \overline{W}_{C_i}, \overline{X}_{C_i})$, might be more effective. □

COMMENT 4: In addition (2.6) suggests that the differences between clusters are fully captured by differences in the values of the pair of averages $(\overline{W}_c, \overline{X}_c)$. In many cases, there may be concerns that the differences between clusters are more complex, and would require accounting also for differences in $\overline{XW}_c$, the cluster average of the product of W_i and X_i, or in the average value of higher order moments of X_i. Given the representation in (2.6) it would be natural to include averages of such functions in the regression function. □

Our proposed approach addresses these last two comments. It provides a formal justification for an unconfoundedness condition given sample cluster averages of functions of the covariates and treatment indicators, which motivates adjusting for those in a flexible, nonlinear way. It also suggests when it would be appropriate to include averages of additional functions of the covariates and treatment indicators beyond $(\overline{W}_c, \overline{X}_c)$.

3 An Alternative to the Fixed Effect Estimator

In this section we present our main results, We propose a new estimator for average treatment effects in the setting with clustered data. The estimator has features in common with the efficient influence function estimators from the program evaluation literature, as well as with the fixed effect estimators from the panel data literature. Unlike fixed effect estimators, it

can accommodate differences in potential outcome distributions between clusters that are not additive. There are two issues involved in our approach. First, we have to be careful in defining the estimand to account for the fact that there may be few units in a cluster. In general, we can not consistently estimate the overall average causal effect, because there are likely to be clusters with no treated or no control units. To take this into account, we define a subset of units for which we estimate the average effect. This subset will depend on fixed characteristics of units as well as on realizations of the sampling and assignment processes in a somewhat unusual manner. Of course, this is not completely new to our approach even in the panel data setting: implicitly standard fixed effect estimators do not estimate the average effect of the treatment if there is systematic variation in treatment effects by strata. However, by explicitly moving away from the focus on population quantities, we relax the conditions required for identification, compared to, say, those in Altonji and Matzkin [2005]. Second, we need to adjust for features of the clusters that cannot be estimated consistently under the asymptotic sequences we consider.

3.1 Some Preliminary Assumptions

The set up we consider has a large population of clusters. In the population, each cluster has a large number of units. We randomly sample a finite number of clusters and then sample a finite number of units from the subpopulation of sampled clusters. Large sample approximations to estimators are based on the number of sampled clusters increasing, with the average number of sampled units per cluster converging to a constant.

Assumption 3.1. (Balanced clustered sampling) *There is a super-population of clusters, we randomly sample n of them and for each cluster we randomly sample $N_c = |c|$ units, with $N = \sum_{c=1}^{n} N_c$ the total sample size. N_c is the same for all clusters*

This assumption describes the sampling process, it is not the only possible sampling scheme that we can allow for, but this is the simplest one. In particular, it is possible to generalize our results to the settings with variation in the number of sampled units for each cluster.

For each unit in the population the (unobserved) data tuple is given by $\{(Y_i(0), Y_i(1), W_i, X_i, U_i, C_i)\}_{i=1}^{N}$. The variable U_i is a cluster-level variable that varies only between clusters, so that it is equal to its cluster average for all units, $\overline{U}_{C_i} = U_i$ for all i.

Our second assumption imposes restrictions on the treatment assignment process:

Assumption 3.2. (Unconfoundedness within Clusters)

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, C_i. \tag{3.1}$$

This assumption implies that we can always compare individuals with the same characteristics within the cluster.

The second assumption imposes restrictions on the fixed effects.

Assumption 3.3. (Random effects)
For the unobserved cluster-level variable $\overline{U}_{C_i}$ we have the following:

$$\Big(Y_i(1), Y_i(0), X_i, W_i\Big) \perp\!\!\!\perp C_i \;\Big|\; \overline{U}_{C_i} \tag{3.2}$$

This assumption, what Altonji and Matzkin [2005] (Assumption 2.3 in their paper) call exchangeability, essentially turns the problem into a random effects set up: the labels of the clusters C_i are not important, only the cluster-level characteristics $\overline{U}_{C_i}$ are. This assumption allows us to conceptualize similarity of clusters.

Since $\overline{U}_{C_i}$ is measurable with respect to cluster indicator variable, an implication of the previous pair of assumptions is:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{U}_{C_i}. \tag{3.3}$$

Now we can also compare treated and control units in different clusters, as long as the clusters have the same value for $\overline{U}_{C_i}$.

3.2 Identification results

For the first identification result we need some additional notation. For each cluster c define define $\mathbb{P}_c$ to be the empirical distribution of (X_i, W_i) in cluster c. In the case with discrete X_i this amounts to the set of frequencies of observations in a cluster for each pair of values (W_i, X_i).[1]

[1] For the formal definition of this object including continuous X_i see Appendix A.

Proposition 1. (Unconfoundedness with empirical measure) *Suppose Assumptions 3.1-3.3 hold. Then:*

$$W_i \perp\!\!\!\perp \left(Y_i(0), Y_i(1)\right) \mid X_i, \mathbb{P}_{C_i} \tag{3.4}$$

For the proofs of the results in this section see Appendix A.

Comment 1: This result states that as long as units have the same characteristics, and they come from clusters identical in terms of $\mathbb{P}_{C_i}$, they are comparable. This is a propensity score type result in the sense that subpopulation with the same value for $(X_i, \mathbb{P}_{C_i})$ are balanced: the distribution of treatments is the same for all units within such subpopulations. □

Comment 2: One can view this result as a statement that $\mathbb{P}_{C_i}$ captures all the information about $\overline{U}_{C_i}$ from the data and thus it is enough to condition on it. While intuitive, this statement is not entirely correct. Clusters can be different in terms of empirical distribution of Y_i, and this can potentially help in predicting $\overline{U}_{C_i}$, but we are not utilizing this for the identification. Instead, we are using the fact that the conditional distribution of $\overline{U}_{C_i}$ given $(X_i, \mathbb{P}_{C_i})$ and W_i does not depend on W_i, and thus we are averaging over the same distribution for control and treated units. □

Define the set of observations that have at least one match in terms of empirical measure and covariates but have a different assignment:

$$\mathcal{I} := \left\{ i \in \{1, \ldots, N\} : \exists j \in \{1, \ldots, N\} \text{ s.t.} \mathbb{P}_{C_i} = \mathbb{P}_{C_j}, X_i = X_j, W_i \neq W_j \right\}.$$

For each observation $i \in \mathcal{I}$ define the set of matches (non-empty by construction):

$$\mathcal{J}(i) := \left\{ j \in \{1, \ldots, N\} \text{ s.t.} \mathbb{P}_{C_i} = \mathbb{P}_{C_j}, X_i = X_j, W_i \neq W_j \right\}.$$

Then for any $i \in \mathcal{I}$ we can construct the following "estimator" for $Y_i(1) - Y_i(0)$:

$$\hat{\tau}_i := W_i \left(Y_i - \frac{1}{|\mathcal{J}(i)|} \sum_{j \in \mathcal{J}(i)} Y_j \right) + (1 - W) \left(\frac{1}{|\mathcal{J}(i)|} \sum_{j \in \mathcal{J}(i)} Y_j - Y_i \right).$$

Define the average effect for units in $\mathcal{I}$:

$$\tau_{\mathcal{I}} := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\Big(Y_i(1) - Y_i(0)\Big).$$

A natural estimator for this object is the average of $\hat{\tau}_i$:

$$\hat{\tau}_{\mathcal{I}} := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\tau}_i.$$

We have the following simple results:

Proposition 2. *Suppose that the moments of the potential outcome distributions are finite, that X_i is a discrete random variable with finite support, and that the probability that $W_i = 1$ conditional on X_i and C_i is strictly between zero and one. Suppose also that Assumptions 3.1–3.3 hold. Then:*

$$\hat{\tau}_{\mathcal{I}} - \tau_{\mathcal{I}} = o_p(1).$$

Comment 3: Note that the identification result in Proposition 2 is not for a population average causal effect. The results in Altonji and Matzkin [2005] show that the random effects are not sufficient for nonparametric identification of population means, and that includes population average causal effects. That does not, mean, however, that we cannot identify any average causal effect, and the result here shows that we can identify average effects for subpopulations defined in terms of $\boldsymbol{W}$, $\boldsymbol{X}$ and $\boldsymbol{C}$. □

Comment 4: Proposition 2 can be directly used only in cases where the distribution of X_i is discrete and supported on the small number of points. Otherwise, we would never observe clusters with the same value of $\mathbb{P}_{C_i}$.Thus this result is more important from the conceptual point of view, rather than a basis for practical estimation. Nevertheless, it shows how adjusting for features of the cluster that are not estimated consistently (the empirical distribution of (W_i, X_i), not the population distribution) can still lead to consistent estimates of causal effects. □

In order to get an operational identification result, that is, one that works even with multi-valued X_i, we impose additional structure:

Assumption 3.4. (Exponential family) *Conditional on U_i distribution of (X_i, W_i) belongs*

to an exponential family with a known sufficient statistic:

$$f_{X_i,W_i|U_i}(x,w|u) \propto h(x,w)\exp\Big\{\eta^T(u)S(x,w)\Big\}, \tag{3.5}$$

with potentially unknown carrier h.

Define $S_i := S(X_i, W_i)$, and let $\overline{S}_c$ be the cluster average of S_i for cluster c.

COMMENT 5: This assumption restricts the statistical model for $(X_i, W_i)|\overline{U}_{C_i}$ but not for $Y_i|X_i, W_i, \overline{U}_{C_i}$. We also do not put any restrictions on the carrier $h(\cdot)$, which can be a general function of its arguments. This makes this model quite flexible. □

Theorem 1. (UNCONFOUNDEDNESS WITH SUFFICIENT STATISTIC) *Suppose Assumptions 3.1–3.4 hold. Then:*

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{S}_{C_i}. \tag{3.6}$$

Theorem 1 can be viewed as essentially a direct consequence of Proposition 1, but it is substantially more operational. It reduces the potentially high-dimensional object $\mathbb{P}_{C_i}$ to a lower dimensional average $\overline{S}_{C_i}$. It is also unusual in that one of the conditioning variables, $\overline{S}_{C_i}$, is not a fixed unit-level characteristic. Instead, it is a characteristic of the cluster and the sampling process. If we change the sampling process, say to sampling twice as many units per cluster, the distribution of $\overline{S}_{C_i}$ changes. Nevertheless, this conceptual difference in the nature of $\overline{S}_{C_i}$ relative to the unit-level characteristic X_i does not affect how it is used in the estimation procedures.

There is another key difference between the unconfoundedness condition in Theorem 1 and in Proposition 1. With continuous covariates, the latter essentially makes it impossible to have overlap. Indeed, unless we have individuals with the same value of covariates within the cluster, the distribution of W_i given X_i and $\mathbb{P}_{C_i}$ is degenerate. It is well known that overlap is crucial in the semiparametric estimation of treatment effects and without it, the identification is possible only under functional form assumptions.

The result in Theorem 1 is more useful because it allows us to control the degree of overlap as well. With $|c|$ being fixed the higher is the dimension of $S(\cdot)$ the closer we are to controlling for $\mathbb{P}_{C_i}$, and thus the smaller is the region for which we have overlap. It would be interesting to

balance this effect with other statistical effects that arise from having higher-dimensional $S(\cdot)$, but we leave this for future work. In this paper, we will assume that $S(\cdot)$ is known, fixed and there is a known region of the covariate space where we have overlap.

In particular, define propensity score:

$$e(x,s) := \mathbb{E}[W_i | X_i = x, \overline{S}_{C_i} = s] \tag{3.7}$$

We are making the following assumption:

Assumption 3.5. (KNOWN OVERLAP) *We assume that there exists $\eta > 0$ and a nonempty known set $\mathbb{A}$, such that for any $(x,s) \in \mathbb{A}$ we have $\eta < e(x,s) < 1-\eta$.*

COMMENT 6: This assumption has two parts: the first part restrict $e(x,s)$ to be non-degenerate on a certain set. This is necessary if we want to identify treatment effects without relying on functional form assumptions. The second part is different: we assume that the set is known to a researcher. This is a generalization of the standard overlap assumption, where we assume that the set $\mathbb{A}$ is equal to the support of the covariate space. See Crump et al. [2009]. □

3.3 An Example

As a natural example of the density that satisfies Assumption 3.4 consider the following family:

$$\begin{cases} \mathbb{E}[W_i|U_i] =: \pi(U_i) \\ X_i|U_i, W_i \sim \mathcal{N}\left(\mu(W_i,U_i), \sigma^2(W_i,U_i)\right) \end{cases} \tag{3.8}$$

It is easy to see that the conditional distribution of (X_i, W_i) given U_i has exponential family representation:

$$\begin{cases} f(X_i, W_i|U_i) \propto \exp\{\eta_1(U_i)X_i^2 + \eta_2(U_i)W_iX_i^2 + \eta_3(U_i)X_i + \eta_4(U_i)W_iX_i + \eta_5(U_i)W_i\}h(X_i, W_i) \\ \eta_1(U_i) := \frac{1}{\sigma^2(0,U_i)} \\ \eta_2(U_i) := \frac{1}{\sigma^2(1,U_i)} - \frac{1}{\sigma^2(0,U_i)} \\ \eta_3(U_i) := -2\frac{\mu(0,U_i)}{\sigma^2(0,U_i)} \\ \eta_4(U_i) := 2\left(\frac{\mu(0,U_i)}{\sigma^2(0,U_i)} - \frac{\mu(1,U_i)}{\sigma^2(1,U_i)}\right) \\ \eta_5(U_i) := \log\left(\frac{\pi(U_i)}{1-\pi(U_i)}\right) + \mu^2(1,U_i) - \mu^2(0,U_i) \end{cases} \tag{3.9}$$

Thus in this case the sufficient statistics is 5-dimensional: $\overline{S}_c = \left(\overline{X^2}_c, \overline{X}_c, \overline{X^2W}_c, \overline{XW}_c, \overline{W}_c\right)$.

4 Estimating Average Treatment Effects

In this section, we discuss two ways to exploit the results in Theorem 1. In the first, we use a linear model approach, where we depart from the fixed effect specification in (2.1) by using additional sufficient statistics, and by adjusting for those in a more general way, though maintaining much of the linear model structure. This is a straightforward approach that may be reasonable when the covariate and sufficient statistic distributions do not differ much by treatment status. In the second we use a general doubly robust approach that is more likely to be appropriate when the distributions differ substantially by treatment status.

4.1 A Linear Model Approach

First we select a set of sufficient statistics. Whereas implicitly the fixed effect approach uses $\overline{X}_c$ and $\overline{W}_c$, we may wish to include in addition the average of the product of X_i and W_i, $\overline{XW}_c = \sum_{i:C_i=c} X_iW_i/N_c$. Given $\overline{S}_c = (\overline{X}_c, \overline{W}_c, \overline{XW}_c)$, we can estimate a linear model with

$$Y_i = \alpha + \tau W_i + X_i^\top\beta + \overline{S}_{C_i}^\top\delta + W_i(X_i - \overline{X})^\top\theta + W_i(\overline{S}_{C_i} - \overline{S})^\top\gamma + \varepsilon_i,$$

where $\overline{X}$ and $\overline{S}$ are the sample averages of X_i and $\overline{S}_{C_i}$ respectively. This allows for more general associations between the potential outcomes and the covariates and sufficient statistics, as well as for interactions with the treatment.

We estimate the parameters by least squares:

$$(\hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\delta}, \hat{\theta}, \hat{\gamma}) :=$$

$$\arg \min_{\alpha,\tau,\beta,\delta,\theta,\gamma} \sum_{i=1}^{N} \left(Y_i - \alpha - \tau W_i - X_i^\top \beta - \overline{S}_{C_i}^\top \delta - W_i (X_i - \overline{X})^\top \theta - W_i (\overline{S}_{C_i} - \overline{S})^\top \gamma \right)^2.$$

Additionally define the following parameters:

$$(\hat{\alpha}_p, \hat{\beta}_p, \hat{\delta}_p, \hat{\theta}_p, \hat{\gamma}_p) :=$$

$$\arg \min_{\alpha_p,\beta_p,\delta_p,\theta_p,\gamma_p} \sum_{i=1}^{N} \left(W_i - \alpha_p - X_i^\top \beta_p - \delta_p^\top \overline{S}_{C_i} - W_i (X_i - \overline{X})^\top \theta_p - W_i (\overline{S}_{C_i} - \overline{S})^\top \gamma_p \right)^2.$$

and let $(\alpha_p, \beta_p, \delta_p, \theta_p, \gamma_p)$ be the corresponding population parameters.

Define $D_i := \left(1, W_i, X_i, \overline{S}_{C_i}, W_i(X_i - \overline{X}), W_i(\overline{S}_{C_i} - \overline{S})\right)$. We make the following standard assumptions about D_i and ε_i:

Assumption 4.1. (Projection assumptions) *The following restrictions are satisfied for D_i an ε_i:*

$$\begin{cases} \mathbb{E}[D_i D_i^T] \text{ is invertible} \\ \mathbb{E}[\varepsilon_i^4] < \infty \\ \mathbb{E}[\|D_i\|_2^4] < \infty \end{cases} \tag{4.1}$$

Assumption 4.2. (Linearity) *The conditional expectation satisfies*

$$\mathbb{E}[Y_i(w) | X_i, \overline{S}_{C_i}] = \alpha + \tau w + X_i^\top \beta + \delta^\top \overline{S}_{C_i} + w(X_i - \overline{X})^\top \theta + w(\overline{S}_{C_i} - \overline{S})^\top \gamma.$$

The proofs of the next lemmas can be found in Appendix B.1.

Lemma 2. *Suppose Assumptions 3.1–3.4, and 4.1-4.2 hold with sufficient statistic $\overline{S}_{C_i}$. Then*

the least squares estimator $\hat{\tau}_{\text{ls}}$ *is consistent for the average treatment effect* $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.

Define the asymptotic variance and its empirical analog:

$$\begin{cases} \mathbb{V} := \frac{\mathbb{E}\left[\left(\sum_{i:C_i=c} u_i \varepsilon_i\right)^2\right]/N_c}{\mathbb{E}[u_i^2]} \\ \hat{\mathbb{V}} := \frac{\sum_{c=1}^{n}\left(\sum_{i:C_i=c} \hat{u}_i \hat{\varepsilon}_i / N_c\right)^2 / n}{\sum_{c=1}^{n}\left(\sum_{i:C_i=c} \hat{u}_i^2 / N_c\right)/n} \end{cases} \tag{4.2}$$

where

$$\begin{cases} u_i := W_i - \alpha_p - X_i^\top \beta_p - \overline{S}_{C_i}^\top \delta_p - W_i(X_i - \overline{X})^\top \theta_p - W_i(\overline{S}_{C_i} - \overline{S})^\top \gamma_p \\ \hat{u}_i := W_i - \hat{\alpha}_p - X_i^\top \hat{\beta}_p - \overline{S}_{C_i}^\top \hat{\delta}_p - W_i(X_i - \overline{X})^\top \hat{\theta}_p - W_i(\overline{S}_{C_i} - \overline{S})^\top \hat{\gamma}_p \\ \hat{\varepsilon}_i := Y_i - \hat{\alpha} - X_i^\top \hat{\beta} - \overline{S}_{C_i}^\top \hat{\delta} - W_i(X_i - \overline{X})^\top \hat{\theta} - W_i(\overline{S}_{C_i} - \overline{S})^\top \hat{\gamma} \end{cases} \tag{4.3}$$

We have the following result:

Lemma 3. *Suppose Assumptions 3.1–3.4,4.1 hold with sufficient statistic* $\overline{S}_{C_i}$. *Suppose in addition that*

$$\mathbb{E}[Y_i | W_i, X_i, \overline{S}_{C_i}] = \alpha + \tau W_i + X_i^\top \beta + \delta^\top \overline{S}_{C_i} + W_i(X_i - \overline{X})^\top \theta + W_i(\overline{S}_{C_i} - \overline{S})^\top \gamma.$$

Then

$$\sqrt{n}(\hat{\tau}_{\text{ls}} - \tau) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}),$$

and

$$\hat{\mathbb{V}} = \mathbb{V} + o_p(1).$$

4.2 The General Case

In this subsection we collect several inference results for the general semiparametric estimator. All proofs can be found in Appendix B.2.

For the further use we use following notation for the conditional mean, propensity score and

residuals:

$$\begin{cases} \mu(W_i, X_i, \overline{S}_{C_i}) := \mathbb{E}[Y_i | W_i, X_i, \overline{S}_{C_i}] \\ e(X_i, \overline{S}_{C_i}) := \mathbb{E}[W_i | X_i, \overline{S}_{C_i}] \\ \varepsilon_i(w) := Y_i(w) - \mu(w, X_i, \overline{S}_{C_i}) \end{cases} \tag{4.4}$$

Note that these expectations are defined conditional on Assumption 3.1, which determines the distribution of $\overline{S}_{C_i}$.

We will use $\hat{\mu}_i(\cdot)$ and $\hat{e}_i(\cdot)$ for generic estimators of $\mu(\cdot)$ and $e(\cdot)$. Subscript i is used to allow for cross-fitting (Chernozhukov et al. [2016]). Define $\{A_i\} := \{(X_i, \overline{S}_{C_i}) \in \mathbb{A}\}$, where $\mathbb{A}$ is the (known) set with overlap in the distribution of $(X_i, \overline{S}_{C_i})$. Define true and estimated share of observations with overlap:

$$\begin{cases} \pi(\mathbb{A}) := \mathbb{E}[\{A_i\}] \\ \hat{\pi}(\mathbb{A}) := \frac{1}{n} \sum_c \frac{1}{N_c} \sum_{i:C_i=c} \{A_i\} \end{cases} \tag{4.5}$$

We assume the generic estimators $\hat{e}_i$ and $\hat{\mu}_i$ satisfy several high-level consistency properties. These restrictions are standard in the program evaluation literature.

Assumption 4.3. (High-level conditions) *The following conditions are satisfied for $\hat{e}_i$ and $\hat{\mu}_i$:*

$$\begin{cases} \eta < \hat{e}_i(X_i, \overline{S}_{C_i}) < 1 - \eta \ \ a.s. \\ \frac{1}{n} \sum_{c=1}^{n} \left[\frac{1}{N_c} \sum_{i:C_i=c} \{A_i\} (e(X_i, \overline{S}_{C_i}) - \hat{e}(X_i, \overline{S}_{C_i}))^2 \right] = o_p(1) \\ \frac{1}{n} \sum_{c=1}^{n} \left[\frac{1}{N_c} \sum_{i:C_i=c} \{A_i\} (\mu(W_i, X_i, \overline{S}_{C_i}) - \hat{\mu}_i(W_i, X_i, \overline{S}_{C_i}))^2 \right] = o_p(1) \\ \frac{1}{n} \sum_{c=1}^{n} \left[\frac{1}{N_c} \sum_{i:C_i=c} \{A_i\} (e(X_i, \overline{S}_{C_i}) - \hat{e}_i(X_i, \overline{S}_{C_i}))^2 \right] \\ \qquad \times \frac{1}{n} \sum_{c=1}^{n} \left[\frac{1}{N_c} \sum_{i:C_i=c} \{A_i\} (\mu(W_i, X_i, \overline{S}_{C_i}) - \hat{\mu}_i(W_i, X_i, \overline{S}_{C_i}))^2 \right] = o_p\left(\frac{1}{n}\right) \end{cases} \tag{4.6}$$

We also restrict moments of the residuals:

Assumption 4.4. (Moment conditions)

$$\begin{cases} \mathbb{E}[\varepsilon_i^2(k)|X_i, \overline{S}_{C_i}] < K \ \ a.s. \\ \mathbb{E}[\varepsilon_i^4(k)] < \infty \end{cases} \tag{4.7}$$

For arbitrary (subject to appropriate integrability conditions) functions $(\mu(\cdot), e(\cdot))$ define the following functional:

$$\psi(y, w, x, s, \mu(\cdot), e(\cdot)) := \mu(1, x, s) - \mu(0, x, s) + \left(\frac{w}{e(x, s)} - \frac{1-w}{1-e(x, s)}\right)(y - \mu(w, x, s)). \tag{4.8}$$

We focus on the following causal estimand:

$$\tilde{\tau}_{\mathbb{A}} = \frac{1}{\hat{\pi}(\mathbb{A})}\frac{1}{n}\sum_{c=1}^{n}\left(\frac{1}{N_c}\sum_{i:C_i=c}\{A_i\}\left(\mu(1, X_i, \overline{S}_{C_i}) - \mu(0, X_i, \overline{S}_{C_i})\right)\right) \tag{4.9}$$

This is a quantity that is random because of its dependence on $\{A_i\}$.[2]

Theorem 2. (Consistency) *Suppose Assumptions 3.1–3.4 and Assumption 4.3 hold. Then:*

$$\hat{\tau}_{\text{dr}} := \frac{1}{\hat{\pi}(\mathbb{A})}\frac{1}{n}\sum_{c=1}^{n}\left(\frac{1}{N_c}\sum_{i:C_i=c}\{A_i\}\psi(Y_i, W_i, X_i, \overline{S}_{C_i}, \hat{\mu}(W_i, X_i, \overline{S}_{C_i}), \hat{e}(X_i, \overline{S}_{C_i}))\right), \tag{4.10}$$

satisfies $\hat{\tau}_{\text{dr}} - \tilde{\tau}_{\mathbb{A}} = o_p(1)$.

For inference results we need to use $\hat{\mu}_i$ with cross-fitting. We also need to take account of the clustering. Define

$$\rho(c, \mu(\cdot), e(\cdot)) := \frac{1}{N_c}\sum_{i:C_i=c}\{A_i\}\psi(Y_i, W_i, X_i, \overline{S}_{C_i}, \mu(W_i, X_i, \overline{S}_{C_i}), e(X_i, \overline{S}_{C_i})),$$

so that

$$\hat{\tau}_{\text{dr}} = \frac{1}{n}\sum_{c=1}^{n}\rho(c, \hat{\mu}(\cdot), \hat{e}(\cdot))/\hat{\pi}(\mathbb{A}).$$

[2]It is straightforward to extend our inference results to a more standard target $\tau_{\mathbb{A}}$, in which case we will have a different (larger) variance.

Theorem 3. (Inference for semiparametric case) *Suppose Assumptions 3.1–3.4 and Assumption 4.3 hold. Assume that $\hat{\mu}_i$ is estimated using cross-fitting with L folders. Then:*

$$\sqrt{n}(\hat{\tau}_{\text{dr}} - \tilde{\tau}_{\mathbb{A}}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}), \qquad \text{where} \quad \mathbb{V} = \frac{\mathbb{E}\left[\xi_c^2\right]}{\pi^2(\mathbb{A})},$$

where ξ_c is defined in the following way:

$$\xi_c := \sum_{i \in c} \frac{1}{N_c}\{A_i\} \left(\frac{W_i}{e(X_i, \overline{S}_{C_i})} - \frac{1 - W_i}{1 - e(X_i, \overline{S}_{C_i})} \right) (Y_i - \mu(W_i, X_i, \overline{S}_{C_i}))$$

Finally, we address the estimation of variance. For this define the following empirical version of ξ_c:

$$\hat{\xi}_c := \sum_{i \in c} \frac{1}{N_c}\{A_i\} \left(\left(\frac{W_i}{\hat{e}(X_i, \overline{S}_{C_i})} - \frac{1 - W_i}{1 - \hat{e}(X_i, \overline{S}_{C_i})} \right) (Y_i - \hat{\mu}(W_i, X_i, \overline{S}_{C_i})) \right) \tag{4.11}$$

The proposed variance estimator is just the variance of $\hat{\xi}_c$:

$$\hat{\mathbb{V}} := \frac{1}{\hat{\pi}^2(\mathbb{A})} \frac{1}{n} \sum_{c=1}^{n} \left(\hat{\xi}_c - \frac{1}{n} \sum_{c'=1}^{n} \hat{\xi}_{c'} \right)^2. \tag{4.12}$$

The following proposition says that asymptotically variance of the estimated influence function is equal to the variance of the true influence function:

Proposition 3. (Variance consistency) *Suppose the assumptions of Theorem 3 hold. Then the variance estimator is consistent:*

$$\hat{\mathbb{V}} = \mathbb{V} + o_p(1). \tag{4.13}$$

5 Applications

5.1 Empirical Illustration

We consider data from Das et al. [2016], in this paper authors want to estimate the differences in quality between public and private healthcare providers in rural India. To achieve this,

they sent 15 standardized (fake) patients, each with three different cases to both private and public providers in 100 villages in 5 districts. For each provider the authors also observe several covariates, in our analysis, we will use gender and age.

In this setup, W is the indicator for a public/private healthcare provider, X is (gender, age). We define clusters by interacting patient, the case, and the district. This leads to potentially $15 \times 3 \times 5 = 225$ clusters, but in the data, only a fraction of this is observed. The distribution of the cluster sizes is summarized in Table 1.

We use a 5-dimensional sufficient statistic in this setup: $\left(\overline{W}_{C_i}, \overline{X}_{C_i}, \overline{WX}_{C_i}\right)$, where $X_i = (\text{age}_i, \text{gender}_i)$. We use $\mathbb{A} = \{i : 0 < \overline{W}_{C_i} < 1\}$. We are left with 520 observations out of 635. In order to construct the doubly-robust estimator we estimate a simple logit model. The results for the logit model are presented in Table 2. While none of the sufficient statistic is significant on its own, the LR test strongly rejects the model without them ($LR = 44.09$ at 5 degrees of freedom).

Estimation results are presented in Table 3. We report the standard fixed effect estimator along with a simple OLS estimator based on (X_i, S_i). We report these estimators for the full and restricted sample. For the restricted sample we also compute the doubly-robust estimator. We use cluster-bootstrap to compute the standard errors.

Results for all five estimators are qualitatively similar, which is expected because the data come from an experiment. We view this exercise as a proof of concept.

5.2 Simulations

For the simulation we consider the following DGP:

$$\begin{cases} U_i \sim \text{Bern}(0.5) \\ W_i | U_i \sim \text{Bern}(\pi(U_i)) \\ X_i | W_i, U_i \sim \mathcal{N}(\mu(W_i, U_i), 1) \\ Y_i | X_i, W_i, U_i \sim \mathcal{N}\left(\alpha(U_i) + \beta(U_i) X_i + \tau(U_i) W_i, 1\right) \end{cases} \tag{5.1}$$

Parameters $(\pi(U_i), \mu(W_i, U_i), \alpha(U_i), \beta(U_i), \tau(U_i))$ are provided in Table 4. In this case, the sufficient statistic is given by $(\overline{X}_{C_i}, \overline{W}_{C_i}, \overline{WX}_{C_i}, \overline{X^2}_{C_i})$.

For the simulation, we consider a situation with small clusters (4 units per cluster) and a large number of them (200 clusters). For each simulation, we consider two sets of estimators. The first set includes the standard fixed effect estimator, OLS estimator with $(X_i, S_i, (XS)_i)$ as covariates and the residual-adjusted OLS estimator, with weights based on logit propensity score. For the second set of estimators, we restrict the sample and consider only clusters with $0 < \overline{W}_{C_i} < 1$. For this sample, we again run the standard fixed effects model, OLS model with $(X_i, S_i, (XS)_i)$ as covariates and three residual-adjusted estimators, with propensity score model estimated by a simple logit model, flexible tree model, and random forest.

Results averaged over 500 simulations are presented in Table 5 and Table 6. In both full and restricted samples, the fixed effect estimator performs very poorly: the average result is equal to 7.51 which is outside of the convex hull of possible treatment effects (given by the interval $[4, 6]$). In both cases, the OLS estimator performs significantly better, and the doubly-robust estimator performs better in terms of RMSE. The best RMSE is achieved using a combination of OLS and a random forest estimator for the propensity score. The resulting estimator is nearly unbiased and has only a slightly higher standard deviation. Note that the restricted sample is 80% of the original one.

6 Extensions

6.1 The Setting with Panel Data

In this section we briefly consider the case with panel data. The ideas developed in the current paper for the cluster setting extend readily to the panel case, although the assumptions that would justify them might be more controversial in this setting. In the panel case the current paper fits into a recent literature that connects more explicitly the panel data literature with the causal treatment effect literature by allowing for general heterogeneity beyond additive effects. For example, in an influential paper Abadie et al. [2010] develops a synthetic control approach that focuses directly on estimating counterfactual outcomes for units exposed to the treatment of interest as a weighted average of outcomes for units exposed to the control treatment. Xu [2017] and Athey et al. [2017] build on Bai and Ng [2002, 2017] to develop matrix completion methods for this setting. Bonhomme and Manresa [2015] and Bonhomme et al. [2017] consider a set up where the units form clusters such that within clusters the units are homogenous, but

between clusters there may be heterogeneity beyond additive components. Which cluster a unit belongs to is is estimated using methods similar to k-means. de Chaisemartin and D'Haultfœuille [2018] focus on the staggered adoption case where, once a unit is exposed to the treatment, it remains exposed in all periods after the initial one. They investigate the interpretation of the two-way fixed effects estimator under general heterogeneity, but maintaining the assumption that the adoption date has no effects beyond the effect of being exposed in the current period. Athey and Imbens [2018] also focus on the staggered adoption case and consider randomization inference, allowing for general treatment effects of the adoption date. Imai and Kim [2016] discusses models that take the dynamic aspects of the panel setting more seriously than the simple two-way fixed effects set up, using graphical models of the type advocated by Pearl [2000]. de Chaisemartin and D'Haultfœuille [2017] study a two-group and two-period difference-in-differences setting where only some units in the second period treatment group receive the treatment, again allowing for general treatment effect heterogeneity. Borusyak and Jaravel [2016] study the staggered adoption case within a model with additive time and unit effects, but allowing for general dynamic treatment effects. In a novel approach Chernozhukov et al. [2017] use double ML approach to estimate causal effects in the panel setting where both the number of units and the dimension of the covariates create high-dimensional problems that require regularization. In order to deal with the unobserved heterogeneity, they exploit (approximate) sparsity assumptions on the errors in the regression of the fixed effects on the average individual characteristics and use the appropriate lasso procedure to control for them. In a linear panel data model with additive fixed effects Kock and Tang [2018] use sparsity assumptions directly on the fixed effects. At an abstract level, in our approach extended to the panel case, our exponential family assumption substitutes for the sparsity assumption in Chernozhukov et al. [2017] and Kock and Tang [2018].

Now let us consider the generalization of the set up in the current paper to the panel case. Suppose we have N observations on C individuals, and T time periods, so that $N = C \times T$. We observe Y_i for all units and a binary treatment W_i. Let $T_i \in \{1, \ldots, T\}$ denote the time period observation i is from, and let $C_i \in \{1, \ldots, C\}$ denote the individual it goes with.

For any variable Z_i, define the time and individual averages:

$$\overline{Z}_{\cdot t} := \frac{1}{C} \sum_{i:T_i=t} Y_i, \qquad \overline{Z}_{c\cdot} := \frac{1}{T} \sum_{i:C_i=c} Y_i,$$

and the overall average

$$\overline{Z} := \frac{1}{N}\sum_{i=1}^{N} Z_i,$$

and the residual

$$\dot{Z}_i = Z_i - \overline{Z}_{\cdot t} - \overline{Z}_{c\cdot} + \overline{Z}$$

Let $\hat{\tau}_{\text{fe}}$ be the least squares estimator for the regression

$$Y_i = \alpha_{T_i} + \beta_{C_i} + \tau W_i + X_i^\top \gamma + \varepsilon_i \tag{6.1}$$

Compare this to the least squares regression

$$Y_i = \tau W_i + X_i^\top \gamma + \delta \overline{W}_{\cdot T_i} + \mu \overline{W}_{C_i\cdot} + \psi \overline{X}_{\cdot T_i} + \varphi \overline{X}_{C_i\cdot} + \varepsilon_i$$

The two least squares estimators for τ are numerically identical. This suggests that we can view the standard fixed-time effects approach in (6.1) as controlling for time and individual level sufficient statistics. This view opens a road to generalizing the standard estimators.

At the same time, this type of generalization is not completely satisfactory. For one, controlling for future values of X_{it} and W_{it} seems controversial. Also, it seems that the outcome information should be used to control for individual-level heterogeneity. Finally, in the panel case, the definition of treatment effects is inherently more complex, because of the dynamic structure of the problem. For these reasons, we think that the approach of this paper while insightful should be refined to make it appropriate for the panel data settings. We leave this for future research.

6.2 Quantile Treatment Effects

Theorem 1 states that conditional on the covariates and the sufficient statistics we have the unconfoundedness condition:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \overline{S}_{C_i}.$$

This implies that we can study estimation of effects other than average treatment effects. In particular, for any bounded function $f : \mathbb{R} \to \mathbb{R}$ we can estimate $\mathbb{E}[f(Y_i(k))]$ using the following representation:

$$\mathbb{E}[f(Y_i(k))] = \mathbb{E}\left[\frac{\{W_i = k\} f(Y_i)}{e(X_i, \overline{S}_{C_i})}\right]$$

For example, we can estimate quantile treatment effects of the type introduced by Lehmann and D'Abrera [2006]. If we are interested in q-th quantile of the distribution of $Y_i(k)$ then (under appropriate continuity) we can identify it as a solution of the following problem:

$$c : \mathbb{E}\left[\frac{\{W_i = k\}\{Y_i \leq c\}}{e(X_i, \overline{S}_{C_i})}\right] = q$$

For the standard case under unconfoundedness Firpo [2007] has developed effective estimation methods that can be adapted to this case. We leave this for future research.

7 Conclusion

In this work, we proposed a new approach to identification and estimation in the observational studies with unobserved cluster-level heterogeneity. The identification argument is based on the combination of random effects and exponential family assumptions. We show that given this structure we can identify a specific average treatment effect even in cases where the observed number of units per cluster is small. From the operational point of view, our approach allows researchers to utilize all the recently developed machinery from the standard observational studies. In particular, we generalize the doubly-robust estimator and prove its consistency and asymptotic normality under common high-level assumptions. We also show that the standard

fixed effects estimation is a particular case of our procedure.

As a direction for future research, it will be interesting to see whether it is possible to utilize machine learning methods to learn sufficient statistics from the data. Additionally, it is essential to understand the statistical trade-off between the dimension of the sufficient statistic, cluster size and estimation rate for the propensity score. Finally, we view this work as a first step towards understanding a more challenging and arguably more practically important data design, where we observe panel data.

References

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

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

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. Econometrica, 84(2):781–807, 2016.

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.

Joseph G Altonji and Richard K Mansfield. Group-average observables as controls for sorting on unobservables when estimating group treatment effects: The case of school and neighborhood effects. Technical report, National Bureau of Economic Research, 2014.

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.

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

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

Manuel Arellano and Jinyong Hahn. A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. Technical report, CEMFI, 2006.

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

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 and Serena Ng. Determining the number of factors in approximate factor models. Econometrica, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Principal components and regularized estimation of factor models. arXiv preprint arXiv:1708.08137, 2017.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1):249–275, 2004.

Peter Bickel, Chris Klaassen, Yakov Ritov, and Jon Wellner. Efficient and adaptive estimation for semiparametric models. 1998.

Stéphane Bonhomme. Functional differencing. Econometrica, 80(4):1337–1385, 2012.

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. Technical report, IFS Working Papers, 2017.

Kirill Borusyak and Xavier Jaravel. Revisiting event study designs. 2016.

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

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, Matt Goldman, Vira Semenova, and Matt Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. arXiv preprint arXiv:1712.09988, 2017.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. Biometrika, pages 187–199, 2009.

Jishnu Das, Alaka Holla, and Aakash Mohpal. Quality and accountability in health care delivery: Audit-study evidence from primary care in india. THE AMERICAN ECONOMIC REVIEW, 106(12):3765–3799, 2016.

Clément de Chaisemartin and Xavier D'Haultfœuille. Fuzzy differences-in-differences. The Review of Economic Studies, 85(2):999–1028, 2017.

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

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1):1–23, 2015.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. Econometrica, 75(1):259–276, 2007.

Jinyong Hahn and Whitney Newey. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica, 72(4):1295–1319, 2004.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

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

Cheng Hsiao, H Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. Journal of Applied Econometrics, 27(5):705–740, 2012.

Kosuke Imai and In Song Kim. When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data? PhD thesis, Working paper, Princeton University, Princeton, NJ, 2016.

Guido Imbens and Jeffrey Wooldridge. Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47(1):5–86, 2009.

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

Anders Bredahl Kock and Haihan Tang. Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. Econometric Theory, pages 1–65, 2018.

Erich Leo Lehmann and Howard JM D'Abrera. Nonparametrics: statistical methods based on ranks. Springer New York, 2006.

Stephen L Morgan and Christopher Winship. Counterfactuals and causal inference. Cambridge University Press, 2014.

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

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. Econometrica: Journal of the Econometric Society, pages 1–32, 1948.

Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

M Hashem Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica, 74(4):967–1012, 2006.

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.

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

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.

Shu Yang. Propensity score weighting for causal inference with multi-stage clustered data. arXiv preprint arXiv:1607.07521, 2016.

A Identification results

First, we need to formally define $\mathbb{P}_c$. For this fix an arbitrary linear order $\succsim$ on $\mathcal{X} \times \{0,1\}$ (e.g., a lexicographic order). For any cluster c consider a tuple $A_c = \{(X_i, W_i)\}_{i \in c}$, order elements of A_c with respect to $\succsim$ and define $\mathbb{P}_c = \big((X_{(1)}, W_{(1)}), \ldots, (X_{(c)}, W_{(c)})\big) \in (\mathcal{X} \times \{0,1\})^c$. Under Assumption 3.1 this construction ensures that $\mathbb{P}_c$ is a well-defined random vector. It is clear that there is a one-to-one relationship between this vector and the empirical distribution of (X_i, W_i) within the cluster which makes the notation appropriate.

Below we will use the following definition of conditional independence. Let X, Y, Z be three random elements and A, B be the elements of the $\sigma(X)$- and $\sigma(Y)$-algebras, respectively. The $X \perp\!\!\!\perp Y | Z$ if the following holds:

$$\mathbb{E}[\{X \in A\}\{Y \in B\}|Z] = \mathbb{E}[\{X \in A\}|Z]\mathbb{E}[\{Y \in B\}|Z] \tag{1.1}$$

In the proofs below we are using A and B as generic elements of the appropriate σ-algebras, without explicitly specifying them.

We start stating several lemmas that are important for the first identification result (Proposition 1). The first lemma says that given the (X_i, W_i, U_i) other covariates cannot help in predicting $(Y_i(0), Y_i(1))$.

Lemma A1. (Statistical exclusion) *Under Assumptions 3.1, 3.3 the following is true:*

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp \{(\mathbb{P}_{C_j}, X_j, W_j)\}_{i=1}^N | X_i, W_i, U_i \tag{1.2}$$

Proof. From the repeated application of the iterated expectations and Assumptions 3.1, 3.3 we have the following:

$$\begin{aligned}
&\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|X_i, W_i, U_i] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|\{X_i, W_i, U_i, C_i\}_{i=1}^n]|X_i, W_i, U_i] = \\
&\mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|\{X_i, W_i, U_i, C_i\}_{i=1}^n]|X_i, W_i, U_i] = \\
&\mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, W_i, U_i]|X_i, W_i, U_i] = \\
&\qquad \mathbb{E}[\{\{(\mathbb{P}_{C_j}, X_j, W_j)\}_{j=1}^N \in B\}|X_i, W_i, U_i]\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, W_i, U_i]
\end{aligned} \tag{1.3}$$

Equality between the first and the last expression implies the independence result. □

The second lemma states that only $\mathbb{P}_{C_i}$ are useful in predicting U_i.

Lemma A2. (Statistical sufficiency) *Under Assumption 3.1 the following holds:*

$$U_i \perp\!\!\!\perp \{W_j, X_j\}_{j=1}^N | \mathbb{P}_{C_i} \tag{1.4}$$

Proof. The proof follows from the following equalities:

$$\begin{aligned}
&\mathbb{E}[\{U_i \in A\}\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}] = \\
&\mathbb{E}\left[\mathbb{E}[\{U_i \in A\}\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j=1}^N, \{C_j = C_i\}_{j=1}^N]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}, \{W_j, X_j\}_{j:C_j=C_i}]|\mathbb{P}_{C_i}\right] = \\
&\mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}\mathbb{E}[\{U_i \in A\}|\mathbb{P}_{C_i}]|\mathbb{P}_{C_i}\right] = \\
&\qquad\qquad \mathbb{E}\left[\{\{W_j, X_j\}_{j=1}^N \in B\}|\mathbb{P}_{C_i}\right]\mathbb{E}\left[\{U_i \in A\}|\mathbb{P}_{C_i}\right]
\end{aligned} \tag{1.5}$$

The third equality holds by random sampling (observations in different clusters are independent), the fourth equality holds by exchangeability of data within the cluster. □

Proof of Proposition 1: We start with the following equalities:

$$\begin{aligned}
&\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\qquad\qquad \mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}]
\end{aligned} \tag{1.6}$$

The last equality follows from Lemma A1. As a next step we have the following result:

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, U_i]|W_i, X_i, \mathbb{P}_{C_i}] = \\
&\mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, U_i]|X_i, \mathbb{P}_{C_i}] = \\
&\quad \mathbb{E}[\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}, U_i]|X_i, \mathbb{P}_{C_i}] = \mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}] \quad (1.7)
\end{aligned}$$

The first equality follows directly from Assumption 3.2, the second equality follows from Lemma A2. Combining the two chains of equalities we get the following:

$$\mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|W_i, X_i, \mathbb{P}_{C_i}] = \mathbb{E}[\{(Y_i(1), Y_i(0)) \in A\}|X_i, \mathbb{P}_{C_i}] \quad (1.8)$$

which proves the conditional independence. □

Corollary A1. (Exclusion in exponential families) *Under the assumptions of Lemma A1 the following is true:*

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp \{(\overline{S}_{C_j}, X_j, W_j)\}_{j=1}^N | X_i, W_i, U_i \quad (1.9)$$

Proof. Because S_{C_i} is a function of $\mathbb{P}_{C_i}$ the result follows from Lemma A1. □

Lemma A3. (Sufficiency in exponential families) *Under Assumptions 3.1 and 3.4 the following holds:*

$$U_i \perp\!\!\!\perp \{W_j, X_j\}_{j=1}^N | \overline{S}_{C_i} \quad (1.10)$$

Proof. The proof is exactly the same as in Lemma A2 with S_{C_i} used instead of $\mathbb{P}_{C_i}$. The fourth equality now holds directly by the exponential family assumption. □

Proof of Theorem 1: The same as for Proposition 1, use Corollary A1 and Lemma A3 instead of Lemmas A1 and A2. □

Corollary A2. *For any function f such that $\mathbb{E}[|f(Y(k))|] < \infty$ the following is true:*

$$\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\ \{W_i = 0\}\mathbb{E}[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\}\mathbb{E}[f(Y_i(1))|X_i, \overline{S}_{C_i}] \quad (1.11)$$

Proof. The proof follows from the following equalities:

$$\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \mathbb{E}[\mathbb{E}[f(Y_i)|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \\ \mathbb{E}[\mathbb{E}[f(Y_i)|W_i, X_i, U_i]|\{W_j, X_j, \overline{S}_{C_j}\}_{j=1}^N] = \mathbb{E}[f(Y_i)|W_i, X_i, \overline{S}_{C_i}] = \\ \{W_i = 0\}\mathbb{E}[f(Y_i(0))|X_i, \overline{S}_{C_i}] + \{W_i = 1\}\mathbb{E}[f(Y_i(1))|X_i, \overline{S}_{C_i}] \quad (1.12)$$

where the third equality follows from Corollary A1, the fourth from Lemma A3 and the final one from Proposition 1. □

B Inference results

Notation: We are using standard notation from the empirical processes literature adapted to our setting. For any **cluster-level** random vector X_c: $\mathbb{P}_n(X_c) := \frac{1}{n}\sum_{c=1}^{n} X_c$ and $\mathbb{G}_n(X_c) := \sqrt{n}\left(\mathbb{P}_n(X_c) - \mathbb{E}[X_c]\right)$. Define $B_i = (X_i, \overline{S}_{C_i})$ and $D_i := (W_i, B_i)$.

B.1 Linear case

Proof of Lemmas 2 and 3:

Denote by $\hat{\beta}_{ls}$ and $\hat{\gamma}_{ls}$ the OLS coefficients in the main and auxiliary regression, respectively. We have the standard linear projection expansion:

$$\begin{aligned}
\hat{\beta}_{ls} &= \left(\frac{1}{n\times c}\sum_{c=1}^{n}\sum_{i:C_i=c} D_i D_i^T\right)^{-1} \frac{1}{n\times |c|}\sum_{c=1}^{n}\sum_{i:C_i=c} Y_i D_i = \\
\beta &+ \left(\frac{1}{n\times |c|}\sum_{c=1}^{n}\sum_{i:C_i=c} D_i D_i^T\right)^{-1} \frac{1}{n\times |c|}\sum_{c=1}^{n}\sum_{i:C_i=c} D_i \varepsilon_i = \\
\beta &+ \left(\mathbb{E}[D_i D_i^T]\right)^{-1} \frac{1}{n}\sum_{c=1}^{n}\frac{1}{|c|}\sum_{i:C_i=c} D_i\varepsilon_i + o_p\left(\frac{1}{\sqrt{n}}\right) = \\
&\beta + \left(\mathbb{E}[D_i D_i^T]\right)^{-1} \frac{1}{n}\sum_{c=1}^{n}\phi_c + o_p\left(\frac{1}{\sqrt{n}}\right) \qquad (2.1)
\end{aligned}$$

where $\phi_c := \frac{1}{|c|}\sum_{i:C_i=c} D_i\varepsilon_i$. This implies consistency and asymptotic normality and also implies that $\|\hat{\beta}_{ls} - \beta\|^2 = O_p\left(\frac{1}{n}\right)$. The same argument implies that $\|\hat{\gamma}_{ls} - \gamma\|^2 = O_p\left(\frac{1}{n}\right)$.

For the variance of $\hat{\tau}_{ls}$ we have the following expression:

$$\mathbb{V}_{ls} = \frac{\mathbb{E}\left[\left(\frac{1}{|c|}\sum_{i:C_i=c} u_i\varepsilon_i\right)^2\right]}{\mathbb{E}[u_i^2]} \qquad (2.2)$$

We are using the plug-in estimator:

$$\hat{\mathbb{V}}_{ls} = \frac{\mathbb{P}_n\left(\frac{1}{|c|}\sum_{i:C_i=c} \hat{u}_i\hat{\varepsilon}\right)^2}{\mathbb{P}_n\left(\frac{1}{c}\sum_{i:C_i=c}\hat{u}_i^2\right)} \qquad (2.3)$$

By construction we have:

$$\begin{cases} \hat{u}_i = u_i + (\gamma - \hat{\gamma})^T B_i \\ \hat{\varepsilon}_i = \varepsilon_i + \left(\beta - \hat{\beta}\right)^T D_i \end{cases} \tag{2.4}$$

For the numerator we have the following

$$\begin{aligned} \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i \hat{\varepsilon}_i \right)^2 = \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i u_i \right)^2 + \\ 2\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i u_i \right) \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i R_{1i} + \sum_{i:C_i=c} \varepsilon_i R_{2i} + \sum_{i:C_i=c} R_{1i} R_{2i} \right) + \\ \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i R_{1i} + \frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i R_{2i} + \frac{1}{|c|} \sum_{i:C_i=c} R_{1i} R_{2i} \right)^2 \end{aligned} \tag{2.5}$$

which implies the following bound:

$$\begin{aligned} \left| \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i \hat{\varepsilon}_i \right)^2 - \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i u_i \right)^2 \right| \leq \\ 2 \sqrt{ \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i u_i \right)^2 \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i R_{1i} + \sum_{i:C_i=c} \frac{1}{|c|} \varepsilon_i R_{2i} + \frac{1}{|c|} \sum_{i:C_i=c} R_{1i} R_{2i} \right)^2 } + \\ \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i R_{1i} + \frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i R_{2i} + \frac{1}{|c|} \sum_{i:C_i=c} R_{1i} R_{2i} \right)^2 \end{aligned} \tag{2.6}$$

Jensen's inequality implies the following bounds:

$$\begin{cases} \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i R_{1i} \right)^2 \leq \mathbb{P}_n \sum_{i:C_i=c} (u_i R_{1i})^2 \\ \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i R_{2i} \right)^2 \leq \mathbb{P}_n \sum_{i:C_i=c} (\varepsilon_i R_{2i})^2 \\ \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} R_{1i} R_{2i} \right)^2 \leq \mathbb{P}_n \sum_{i:C_i=c} (R_{1i} R_{2i})^2 \end{cases} \tag{2.7}$$

CS inequality implies:

$$\begin{cases} R_{1i}^2 \leq \|\hat{\beta} - \beta\|_2^2 \|D_i\|_2^2 \\ R_{2i}^2 \leq \|\hat{\gamma} - \gamma\|_2^2 \|B_i\|_2^2 \end{cases} \tag{2.8}$$

This implies the following bound for the averages:

$$\begin{cases} \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} R_{1i}^2 u_i^2 \leq \|\hat{\beta} - \beta\|_2^2 \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} \|D_i\|_2^2 u_i^2 = o_p(1) \\ \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} R_{2i}^2 \varepsilon_i^2 \leq \|\hat{\gamma} - \gamma\|_2^2 \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} \|B_i\|_2^2 \varepsilon_i^2 = o_p(1) \\ \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} R_{1i}^2 R_{2i}^2 \leq \|\hat{\beta} - \beta\|_2^2 \|\hat{\gamma} - \gamma\|_2^2 \mathbb{P}_n \frac{1}{|c|} \sum_{i:C_i=c} \|B_i\|_2^2 \|D_i\|_2^2 = o_p(1) \end{cases} \tag{2.9}$$

Combining all these together we have the following:

$$\left| \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i \hat{\varepsilon}_i \right)^2 - \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \varepsilon_i u_i \right)^2 \right| \leq o_p(1) \tag{2.10}$$

The same argument implies the following bound:

$$\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i^2 \right) = \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \left(u_i^2 + 2u_i R_{2i} + R_{2i}^2 \right) \right) = \mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i^2 \right) + o_p(1) \tag{2.11}$$

Finally, combining all bounds together we have the following:

$$\frac{\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i \hat{\varepsilon} \right)^2}{\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} \hat{u}_i^2 \right)} = \frac{\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i \varepsilon \right)^2}{\mathbb{P}_n \left(\frac{1}{|c|} \sum_{i:C_i=c} u_i^2 \right)} + o_p(1) = \mathbb{V} + o_p(1) \tag{2.12}$$

Proof of Lemma 1: There are two well known preliminary results we use. First, consider the two regression functions

$$Y_i = X_i\beta_X + Z_i\beta_Z + \varepsilon_i,$$

$$Y_i = X_i\gamma_X + \eta_i,$$

$$Z_i = X_i\Delta_X + \nu_i,$$

with the corresponding least squares estimators $\hat{\beta}_X$, $\hat{\beta}_Z$, $\hat{\gamma}_X$, and Δ_X. Then, the omitted variable bias formula states that

$$\hat{\gamma}_X = \hat{\beta}_X + \hat{\Delta}_X\hat{\beta}_Z.$$

Second, consider the two regression functions

$$Y_i = X_i\beta_X + Z_i\beta_Z + \varepsilon_i,$$

and,

$$Y_i = (X_i - Z_i\Gamma)\beta_X + Z_i\beta_Z + \varepsilon_i.$$

The least squares estimators for β_X based on the two regression functions are identical.

Next, consider the regression function

$$Y_i = \sum_{s=1}^{S} \alpha_s S_{is} + W_i\tau + X_i\beta + \varepsilon, \tag{2.13}$$

with the least squares estimators $\hat{\alpha}_s$, $\hat{\beta}$ and $\hat{\tau}$. Because $\overline{W}_{S_i}$ and $\overline{X}_{S_i}$ are linear combinations of the indicators S_{is}, it follows by the second preliminary result that the least squares coefficients $(\hat{\tau}, \hat{\beta})$ are the same as those based on the regression function

$$Y_i = \sum_{s=1}^{S} \alpha_s S_{is} + (W_i - \overline{W}_{S_i})\tau + (X_i - \overline{X}_{S_i})\beta + \varepsilon, \tag{2.14}$$

Because S_{is} is uncorrelated with $W_i - \overline{W}_{S_i}$ and $X_i - \overline{X}_{S_i}$ it follows by the first preliminary result that the least squares coefficients $(\hat{\tau}, \hat{\beta})$ are also identical to those based on the regression function

$$Y_i = (W_i - \overline{W}_{S_i})\tau + (X_i - \overline{X}_{S_i})\beta + \varepsilon, \tag{2.15}$$

Because the means of $W_i - \overline{W}_{S_i}$ and $X_i - \overline{X}_{S_i}$ are zero, it follows by the first preliminary result we get the same estimates for τ and β if we estimate by least squares the regression function

$$Y_i = \alpha + (W_i - \overline{W}_{S_i})\tau + (X_i - \overline{X}_{S_i})\beta + \varepsilon. \tag{2.16}$$

Because $\overline{W}_{S_i}$ and $\overline{X}_{S_i}$ are both uncorrelated with both $W_i - \overline{W}_{S_i}$ and $X_i - \overline{X}_{S_i}$ it follows by the first preliminary result that the least squares coefficients $(\hat{\tau}, \hat{\beta})$ are also identical to those based on the regression function

$$Y_i = \alpha + \overline{W}_{S_i}\delta + \overline{X}_{S_i}\gamma + (W_i - \overline{W}_{S_i})\tau + (X_i - \overline{X}_{S_i})\beta + \varepsilon. \tag{2.17}$$

Finally, using the second preliminary result again it follows that the least squares estimators for τ and β are the same if based on the regression function

$$Y_i = \alpha + \overline{W}_{S_i}\delta + \overline{X}_{S_i}\gamma + W_i\tau + X_i\beta + \varepsilon_i. \tag{2.18}$$

B.2 Semiparametric case

We start with a reminder on notation:

$$\begin{cases} \mu(D_i) := \mathbb{E}[Y_i|D_i] \\ e(B_i) := \mathbb{E}[W_i|B_i] \\ \varepsilon(k) := Y_i(k) - \mu(k, B_i) \\ \psi(y, w, x, s, \mu(\cdot), e(\cdot)) := \mu(1, x, s) - \mu(0, x, s) + \left(\frac{w}{e(x,s)} - \frac{1-w}{1-e(x,s)}\right)(y - \mu(w, x, s)) \\ \rho(c, \mu(\cdot), e(\cdot)) := \frac{1}{|c|}\sum_{i:C_i=c}\{A_i\}\psi(Y_i, W_i, X_i, \overline{S}_{C_i}, \mu(W_i, X_i, \overline{S}_{C_i}), e(X_i, \overline{S}_{C_i})) \\ \xi_c := \sum_{i\in c}\frac{1}{N_c}\{A_i\}\left(\frac{W_i}{e(X_i,\overline{S}_{C_i})} - \frac{1-W_i}{1-e(X_i,\overline{S}_{C_i})}\right)(Y_i - \mu(W_i, X_i, \overline{S}_{C_i})) \end{cases} \tag{2.19}$$

In order to prove Theorem 2 we consider a more general case that allows for misspecification. First we prove Lemma B4 which states that we get identification if either the propensity score or the conditional mean is potentially misspecified. Then we prove Proposition B1 which is a general consistency result under possible misspecification. Theorem 2 follows as a special case. After that we prove Theorem 3 and Proposition 3.

Lemma B4. *Assume that at least one of the following statements is true:*

$$\begin{cases} \tilde{\mu}(W_i, X_i, \overline{S}_{C_i}) = \mu(W_i, X_i, \overline{S}_{C_i}) \\ \tilde{e}(X_i, \overline{S}_{C_i}) = e(X_i, \overline{S}_{C_i}) \end{cases} \tag{2.20}$$

If the assumptions of Theorem 1 hold then we have the following result:

$$\mathbb{E}[\rho(c, \tilde{m}, \tilde{e})] = \mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}\tau(B_i)\right] \tag{2.21}$$

where $\tau(B_i) := \mu(1, B_i) - \mu(0, B_i)$.

Proof. By construction we have the following:

$$\begin{aligned} &\mathbb{E}[\rho(c, \tilde{\mu}, \tilde{e})] = \mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1, B_i) - \tilde{\mu}(0, B_i) + \left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - \tilde{e}(B_i)}\right)(Y_i - \tilde{\mu}(D_i))\right)\right] = \\ &\mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}(\tilde{\mu}(1, B_i) - \tilde{\mu}(0, B_i))\right] + \sum_{i \in c} \frac{1}{|c|}\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - \tilde{e}(B_i)}\right)(Y_i - \tilde{\mu}(D_i))\right] \end{aligned} \tag{2.22}$$

For the second part we have the following (using unconfoundedness):

$$\begin{aligned} &\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - \tilde{e}(B_i)}\right)(Y_i - \tilde{\mu}(D_i))\right] = \\ &\mathbb{E}\left[\mathbb{E}\left[\{A_i\}\left(\frac{W_i}{\tilde{e}(B_i)} - \frac{1 - W_i}{1 - e(\tilde{B}_i)}\right)(Y_i - \tilde{\mu}(D_i))|B_i\right]\right] = \\ &\qquad \mathbb{E}\left[\{A_i\}\left(\frac{e(B_i)\,(\mu(1, B_i) - \tilde{\mu}(1, B_i))}{\tilde{e}(B_i)} - \frac{(1 - e(B_i)\,(\mu(0, B_i) - \tilde{\mu}(0, B_i))}{1 - \tilde{e}(B_i)}\right)\right] \end{aligned} \tag{2.23}$$

This implies that if either $\tilde{\mu}(D_i) = \mu(D_i)$ or $\tilde{e}(B_i) = e(B_i)$ then $\mathbb{E}[\rho(c, \tilde{m}, \tilde{e})] = \mathbb{E}\left[\sum_{i \in c} \frac{1}{|c|}\{A_i\}\tau(B_i)\right]$. □

Proposition B1. (Consistency with wrong specifications) *Assume that the following conditions hold for* $(\hat{\tilde{e}}, \hat{\tilde{\mu}})$*:*

$$\begin{cases} \mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)^2\right)=o_p(1) \\ \mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{e}}(B_i)-\tilde{e}(B_i)\right)^2\right)=o_p(1) \\ \eta<\tilde{e}(B_i)<1-\eta \ \ a.s. \\ \eta<\hat{\tilde{e}}(B_i)<1-\eta \ \ a.s. \\ \mathbb{E}[\tilde{\varepsilon}_i^2(k)]<\infty \end{cases} \tag{2.24}$$

where $\tilde{\varepsilon}_i(k): Y_i(k)-\tilde{\mu}(k,B_i)$*. Additionally assume that the conditions of Lemma B4 hold. Then we have the following:*

$$\mathbb{P}_n\rho(c,\tilde{\mu},\tilde{e})=\mathbb{P}_n\rho(c,\tilde{\mu},\tilde{e})+o_p(1)=\mathbb{E}[\rho(c,\tilde{\mu},\tilde{e})]+o_p(1) \tag{2.25}$$

Proof. To prove the consistency result we need to separate the functional into two parts:

$$\begin{aligned}\rho(c,\tilde{\mu},\tilde{e})=&\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1,B_i)+\frac{W_i}{\tilde{e}(B_i)}(Y_i-\tilde{\mu}(1,B_i))\right)- \\ &\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(0,B_i)+\frac{1-W_i}{1-\tilde{e}(B_i)}(Y_i-\tilde{\mu}(0,B_i))\right)=\rho_1(c,\tilde{\mu},\tilde{e})-\rho_0(c,\tilde{\mu},\tilde{e})\end{aligned} \tag{2.26}$$

In what follows we are working only with the first part of the functional, the second can be analyzed in the exactly the same way. Define the empirical version:

$$\rho_1(c,\hat{\tilde{\mu}},\hat{\tilde{e}}):=\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\hat{\tilde{\mu}}(1,B_i)+\frac{W_i}{\hat{\tilde{e}}(B_i)}(Y_i-\hat{\tilde{\mu}}(1,B_i))\right) \tag{2.27}$$

We can decompose this expression into three parts:

$$\rho_1(c,\hat{\tilde{\mu}},\hat{\tilde{e}}) = \sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{\mu}(1,B_i)+\frac{W_i}{\tilde{e}(B_i)}(Y_i-\tilde{\mu}(1,B_i))\right)+$$
$$+\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)\left(1-\frac{W_i}{\hat{\tilde{e}}(B_i)}\right)\right)+$$
$$\sum_{i\in c}\frac{1}{|c|}\{A_i\}(Y_i-\tilde{\mu}(1,B_i))W_i\left(\frac{1}{\hat{\tilde{e}}(B_i)}-\frac{1}{\tilde{e}(B_i)}\right)=\rho_1(c,\tilde{\mu},\tilde{e})+R_{1c}+R_{2c} \quad (2.28)$$

The result will follow once we prove two approximations:

$$\begin{cases}\mathbb{P}_n R_{1c}=o_p(1)\\ \mathbb{P}_n R_{2c}=o_p(1)\end{cases} \quad (2.29)$$

We start with the second one. Observe that we have the following:

$$|\mathbb{P}_n R_{2c}|\leq\mathbb{P}_n|R_{2c}|\leq\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}|\tilde{\varepsilon}_i(1)|\left(\frac{\{A_i\}W_i}{\tilde{e}(B_i)\hat{\tilde{e}}(B_i)}\right)\{A_i\}\left|\tilde{e}(B_i)-\hat{\tilde{e}}(B_i)\right|\leq$$
$$\max_i\left(\frac{\{A_i\}W_i}{\tilde{e}(B_i)\hat{\tilde{e}}(B_i)}\right)\sqrt{\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}\tilde{\varepsilon}_i^2(1)}\sqrt{\mathbb{P}_n\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\tilde{e}(B_i)-\hat{\tilde{e}}(B_i)\right)^2}=$$
$$O_p(1)\sqrt{O_p(1)}\sqrt{o_p(1)}=o_p(1) \quad (2.30)$$

For the first term we have the following:

$$R_{1c}=\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)\left(1-\frac{W_i}{\hat{\tilde{e}}(B_i)}\right)\right)=$$
$$\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)\left(1-\frac{W_i}{\tilde{e}(B_i)}\right)\right)+$$
$$\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(\left(\hat{\tilde{\mu}}(1,B_i)-\tilde{\mu}(1,B_i)\right)W_i\left(\frac{\hat{\tilde{e}}(B_i)-\tilde{e}(B_i)}{\tilde{e}(B_i)\hat{\tilde{e}}(B_i)}\right)\right)=R_{11c}+R_{12c} \quad (2.31)$$

The first part can be bounded in the following way:

$$|\mathbb{P}_n R_{11c}| \leq \mathbb{P}_n |R_{11c}| \leq$$
$$\max_i \left| \frac{\{A_i\}(W_i - \tilde{e}(B_i))}{\tilde{e}(B_i)} \right| \times \sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{m}}(1, B_i) - \tilde{\mu}(1, B_i) \right)^2 \right)} =$$
$$O_p(1) \times o_p(1) = o_p(1) \quad (2.32)$$

The second part can be bounded in the following way:

$$|\mathbb{P}_n R_{12c}| \leq \mathbb{P}_n |R_{12c}| \leq \max_i \left(\frac{\{A_i\} W_i}{\tilde{e}(B_i) \hat{\tilde{e}}(B_i)} \right) \times$$
$$\sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{\mu}}(1, B_i) - \tilde{\mu}(1, B_i) \right)^2 \right)} \sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\tilde{e}}(B_i) - \tilde{e}(B_i) \right)^2 \right)} =$$
$$O_p(1) \times o_p(1) \, o_p(1) = o_p(1) \quad (2.33)$$

Combining all the results together we have the proof. □

Proof of Theorem 2: Observe that $\hat{e}$ and $\hat{\mu}$ satisfy the assumptions of Proposition B1 with $\tilde{\mu}$ and $\tilde{e}$ equal to m and e. As a result, combining Proposition B1 and Lemma B4 we get the following:

$$\frac{1}{\hat{\pi}(A)}\mathbb{P}_n\rho(c,\hat{\mu},\hat{e}) = \frac{1}{\hat{\pi}(A)}\left(\mathbb{E}[\rho(c,\mu,e)] + o_p(1)\right) =$$
$$\left(\frac{1}{\pi(A)} + o_p(1)\right)\left(\mathbb{E}[\rho(c,\mu,e)] + o_p(1)\right) = \frac{1}{\pi(A)}\mathbb{E}[\rho(c,\mu,e)] + o_p(1) =$$
$$\frac{1}{\pi(A)}\mathbb{E}[\{A_i\}\tau(X_i,\overline{S}_{C_i})] + o_p(1) \quad (2.34)$$

Proof of Theorem 3: The start of the argument is the same as in proof for the consistency result. We decompose the empirical version of $\rho_1(c, \hat{m}, \hat{e})$:

$$\rho_1(c, \hat{m}, \tilde{e}) - \sum_{i \in c} \frac{1}{|c|} \{A_i\} \mu(1, B_i) = \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\frac{W_i}{e(B_i)} (Y_i - \mu(1, B_i)) \right) +$$
$$+ \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left((\hat{\mu}(1, B_i) - \mu(1, B_i)) \left(1 - \frac{W_i}{\hat{e}(B_i)} \right) \right) +$$
$$\sum_{i \in c} \frac{1}{|c|} \{A_i\} (Y_i - \mu(1, B_i)) W_i \left(\frac{1}{\hat{e}(B_i)} - \frac{1}{e(B_i)} \right) = \xi_{1c} + R_{1c} + R_{2c} \quad (2.35)$$

The result will follow once we prove the following:

$$\begin{cases} \mathbb{P}_n R_{1c} = o_p \left(\frac{1}{\sqrt{n}} \right) \\ \mathbb{P}_n R_{2c} = o_p \left(\frac{1}{\sqrt{n}} \right) \end{cases} \quad (2.36)$$

In exactly the same way as before we can decompose R_{1c} into R_{11c} and R_{12c}. For R_{12c} we have the following:

$$|\mathbb{P}_n R_{12c}| \leq \mathbb{P}_n |R_{12c}| \leq \max_i \left(\frac{\{A_i\} W_i}{e(B_i) \hat{e}(B_i)} \right) \times$$
$$\sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{\mu}(1, B_i) - \mu(1, B_i) \right)^2 \right)} \sqrt{\mathbb{P}_n \left(\sum_{i \in c} \frac{1}{|c|} \{A_i\} \left(\hat{e}(B_i) - e(B_i) \right)^2 \right)} =$$
$$O_p(1) \times o_p \left(\frac{1}{\sqrt{n}} \right) = o_p \left(\frac{1}{\sqrt{n}} \right) \quad (2.37)$$

For R_{11c} we use the following argument:

$$\mathbb{E}\left[(\mathbb{P}_n R_{11c})^2\right] = \mathbb{E}\left[\left(\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] \leq$$

$$|L|\sum_{l\in L}\mathbb{E}\left[\left(\sum_{c:l(c)=l}\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] =$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] \leq$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)\left(1-\frac{W_i}{e(B_i)}\right)\right)^2\right] =$$

$$|L|\sum_{l\in L}\sum_{c:l(c)=l}\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i\in c}\frac{1}{|c|}\left(\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\left(\frac{e(B_i)(1-e(B_i)}{e^2(B_i)}\right)\right)\right] \leq$$

$$K\frac{1}{n}\mathbb{E}\left[\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\right)\right] \quad (2.38)$$

Using this we get the following:

$$\mathbb{E}\left[|\mathbb{P}_n R_{11c}|\right] \leq \sqrt{\mathbb{E}\left[(\mathbb{P}_n R_{11c})^2\right]} \leq$$

$$\frac{K}{\sqrt{n}}\mathbb{E}\left[\mathbb{P}_n\left(\sum_{i\in c}\frac{1}{|c|}\left(\hat{\mu}_{-l(c)}(1,B_i)-\mu(1,B_i)\right)^2\right)\right] = o\left(\frac{1}{\sqrt{n}}\right) \quad (2.39)$$

This implies (by Markov's inequality) that $\mathbb{P}_n R_{11c} = o_p\left(\frac{1}{\sqrt{n}}\right)$

$$\mathbb{E}[R_{2c}^2|\{D_i\}_{i=1}^N] \leq \sum_{i\in c}\frac{1}{|c|}\mathbb{E}[\varepsilon_i^2|D_i]\left(\frac{\{A_i\}W_i}{e^2(B_i)\hat{e}^2(B_i)}\right)\{A_i\}\left(e(B_i)-\hat{e}(B_i)\right)^2 \leq$$

$$\max_i\left(\frac{\{A_i\}\mathbb{E}[\varepsilon_i^2|D_i]W_i}{e^2(B_i)\hat{e}(B_i)}\right)\sum_{i\in c}\frac{1}{|c|}\{A_i\}\left(e(B_i)-\hat{e}(B_i)\right)^2 \quad (2.40)$$

We also have the following:

$$\mathbb{E}[R_{2c}|\{D_i\}_{i=1}^N] = 0 \quad (2.41)$$

Using these two things we get the following:

$$\mathbb{E}[(\mathbb{P}_n R_{2c})^2 \,|\{D_i\}_{i=1}^N] \leq \max_i \left(\frac{\{A_i\}\mathbb{E}[\varepsilon_i^2|D_i]W_i}{e^2(B_i)\hat{e}(B_i)}\right) \times \\ \frac{1}{n}\mathbb{P}_n \sum_{i\in c} \frac{1}{|c|}\{A_i\}\left(e(B_i) - \hat{e}(B_i)\right)^2 \leq K \times o_p\left(\frac{1}{n}\right) = o_p\left(\frac{1}{n}\right) \quad (2.42)$$

This implies that $\mathbb{E}[(\mathbb{P}_n R_{2c})^2] = o\left(\frac{1}{n}\right)$ (because $(\hat{e}-e)^2$ is bounded by 1) and thus $R_{2c} = o_p\left(\frac{1}{\sqrt{n}}\right)$.

Proof of Proposition 3: Similarly to all other proofs we can divide ξ_c into two parts ξ_{1c} and ξ_{0c}. We will analyze ξ_{1c}, analysis for ξ_{0c} is the same. We have the following decomposition:

$$\hat{\xi}_{1c} - \xi_{1c} = \sum_{i \in c} \frac{1}{|c|} \{A_i\} \left((\mu(1, B_i) - \hat{\mu}(1, B_i)) \frac{W_i}{\hat{e}(B_i)} \right) + \sum_{i \in c} \frac{1}{|c|} \{A_i\} (Y_i - \mu(1, B_i)) W_i \left(\frac{1}{\hat{e}(B_i)} - \frac{1}{e(B_i)} \right) = R_{11c} + R_{12c} \quad (2.43)$$

For the first term we have the following bound:

$$\mathbb{P}_n R_{11c}^2 \leq \mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} \left((\mu(1, B_i) - \hat{\mu}(1, B_i))^2 \frac{W_i}{\hat{e}^2(B_i)} \right) \leq \left(\max_i \frac{\{A_i\} W_i}{\hat{e}^2(B_i)} \right) \times \mathbb{P}_n \left(\frac{1}{c} \sum_{i \in c} \{A_i\} (\mu(1, B_i) - \hat{\mu}(1, B_i))^2 \right) = O_p(1) o_p(1) = o_p(1) \quad (2.44)$$

For the second term we have the following bound:

$$\mathbb{P}_n R_{12c}^2 \leq \mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} \{W_i\} \varepsilon_i^2(1) \frac{(\hat{e}(B_i) - e(B_i))^2}{\hat{e}^2(B_i) e^2(B_i)} \leq \sqrt{\left(\mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} \frac{\{A_i\} (\hat{e}(B_i) - e(B_i))^4}{\hat{e}^4(B_i) e^4(B_i)} \right) \left(\mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} \{W_i\} \varepsilon_i^4(1) \right)} \leq K \sqrt{\left(\mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} (\hat{e}(B_i) - e(B_i))^2 \right) \left(\mathbb{P}_n \frac{1}{c} \sum_{i \in c} \{A_i\} \{W_i\} \varepsilon_i^4(1) \right)} = o_p(1) O_p(1) = o_p(1) \quad (2.45)$$

Putting these results together we have the following:

$$\mathbb{P}_n(\hat{\xi}_{1c} + \hat{\xi}_{2c})^2 - \mathbb{P}_n(\xi_{1c} + \xi_{2c})^2 = \mathbb{P}_n(\xi_{1c} + \xi_{2c} + R_{11c} + R_{12c} + R_{01c} + R_{02c})^2 - \mathbb{P}_n(\xi_{1c} + \xi_{2c})^2 = \mathbb{P}_n(\xi_{1c} + \xi_{2c})(R_{11c} + R_{12c} + R_{01c} + R_{02c}) + \mathbb{P}_n(R_{11c} + R_{12c} + R_{01c} + R_{02c})^2 \leq \sqrt{\mathbb{P}_n(\xi_{1c} + \xi_{2c})^2 4 \mathbb{P}_n(R_{11c}^2 + R_{12c}^2 + R_{01c}^2 + R_{02c}^2)} + \mathbb{P}_n(R_{11c}^2 + R_{12c}^2 + R_{01c}^2 + R_{02c}^2) = \sqrt{O_p(1) o_p(1)} + o_p(1) = o_p(1) \quad (2.46)$$

This argument also implies that $\mathbb{P}_n(\hat{\xi}_{1c}) = \mathbb{P}_n(\xi_{1c}) = o_p(1)$ and thus we have the final result:

$$\frac{1}{\hat{\pi}^2(A)}\left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})^2 - \left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})\right)^2\right) - \frac{1}{\pi^2(A)}\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2 =$$
$$\frac{1}{\hat{\pi}^2(A)}\left(\mathbb{P}_n(\hat{\xi}_{1c}+\hat{\xi}_{2c})^2 - \mathbb{P}_n(\xi_{1c}+\xi_{2c})^2\right) + \left(\frac{1}{\hat{\pi}^2(A)} - \frac{1}{\pi^2(A)}\right)\mathbb{P}_n(\xi_{1c}+\xi_{2c})^2 + O_p(1)o_p(1) =$$
$$O_p(1)o_p(1) + o_p(1)O_p(1) + O_p(1)o_p(1) = o_p(1) \quad (2.47)$$

Table 1: Distribution of cluster sizes

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	6	9	9.6	13	24

Table 2: Logit model

	Estimate	s.e.
Intercept	-2.60	2.64
age	0.03	0.01
gender	-1.12	0.46
$\overline{W}$	5.17	5.01
$\overline{age}$	-0.02	0.03
$\overline{(age \times W)}$	0.00	0.09
$\overline{gender}$	0.50	2.28
$\overline{W \times gender}$	0.71	3.93

Table 3: Estimates of causal effect

	$\hat{\tau}_{fe}$	$\hat{\tau}_{ols}$	$\hat{\tau}_{fe}(A)$	$\hat{\tau}_{ols}(A)$	$\hat{\tau}_{dr}(A)$
Est.	-1.62	-1.60	-1.60	-1.55	-1.44
s.e.	0.23	0.25	0.24	0.25	0.30

Table 4: Parameters for the simulation

	$\pi(U_i)$	$\mu(0, U_i)$	$\mu(1, U_i)$	$\alpha(U_i)$	$\beta(U_i)$	$\tau(U_i)$
0	0.4	-1	2	2.59	1	4
1	0.6	1	-2	2.71	-1	6

Table 5: Simulation results: full sample

	τ	$\hat{\tau}^{fe}$	$\hat{\tau}^{ols}$	$\hat{\tau}^{dr}_{logit}$
Mean	5.00	7.53	5.23	5.16
SD	0.08	0.15	0.18	0.18
RMSE		2.53	0.28	0.24

Table 6: Simulation results: restricted sample

	τ	$\hat{\tau}^{fe}$	$\hat{\tau}^{ols}$	$\hat{\tau}^{ols-ad}_{logit}$	τ^{ols-ad}_{tree}	τ^{ols-ad}_{rf}
Mean	5.00	7.51	5.21	5.19	4.87	5.03
SD	0.08	0.15	0.18	0.19	0.31	0.19
RMSE		2.51	0.26	0.26	0.33	0.17